How to write a good PhD thesis and survive the viva

Stefan Rüger
Knowledge Media Institute
The Open University, UK

V 0.87[*]— 12th July 2013

Abstract

The paper gives advice on how to write a good PhD thesis in a Computing subject in the UK — assuming that one has done interesting, novel academic work and "just" needs to write up. The context of this document relates to my experience, opinion and practice in various roles as supervisor, examiner or examination panel chair in more than two dozens of PhD examinations. What I have expressed here is likely to be relevant to my own PhD students or those whose PhD thesis I examine within the British system. Many aspects of this paper are universally true for a number of subjects and countries while some are more specific to Computing and the UK.

1 What is a PhD?

Most universities in the UK award a PhD degree for demonstrating the ability to carry out independent research to academic standards. Normally, a successful PhD candidate shows this ability by

- having studied a particular area within a subject for three to four years
- having made at least one new discovery and/or at least one contribution to the knowledge of a sub-area within the chosen area of the subject
- having written a thesis about that area, placing the own independent, novel contribution in context and comparing it critically with other approaches and
- having defended the thesis in the so-called viva, which is a discussion with examiners, who are experts in the area

The thesis is a monograph, ie, a self-contained piece of work, written solely by the PhD candidate and no-one else. It sets out a certain problem that the candidate has worked on, possibly within a larger team, under guidance of one or more academic advisors. It motivates and defines the problem, reviews existing approaches to the problem, identifies through critical analysis a clear gap for a possible novel academic contribution, and spells out a so-called hypothesis, which is a proposed explanation for the problem or a proposed solution to a problem. The thesis also explains in sufficient detail, and justifies, the work undertaken to decide on the hypothesis (or hypotheses as the case may be). This work typically involves a combination of further literature

[*] This document is in development. Please refer to it as http://people.kmi.open.ac.uk/stefan/thesis-writing.pdf rather than passing on a copy. You can leave feedback at http://people.kmi.open.ac.uk/stefan/thesis-writing-feedback.php.

studies, theoretical analysis, experimental design, data collection, carrying out the experiments, data analysis, and drawing conclusions. A good thesis also delineates the limitation of the work done or the conclusions drawn and outlines possible future research directions.

Novelty. Most universities publish a document that lays open the expectations towards the thesis, including the central requirement of contributing something novel. For example the University of London publishes a 18 page document on their website entitled "Regulations for the degrees of MPhil and PhD"[1], which states with respect to the requirement of novelty *[The thesis shall] form a distinct contribution to the knowledge of the subject and afford evidence of originality by the discovery of new facts and/or by the exercise of independent critical power.* Now you know. Even if a university does not publish an equivalent document and expects their PhD students to pick up an understanding of novelty during mandatory PhD workshops or by osmosis, it will be something along above lines. Please note that the University of London — like most universities — allows to demonstrate originality through the exercise of independent critical power (an excellent literature analysis that contributes to the understanding of a field) in absence of a discovery of new facts. Although this route is rarely taken, above definitions affords PhD students with an alternative plan in case their studies have not led to the discovery of new facts.

Depth, not breadth! Although one of the expectations of a PhD thesis is that it contains something novel as its centre piece, this is often confined to a narrow sub-area of a field. There is no need to spark off a new research field. The award of a PhD only documents the ability to carry out independent research to academic standards. A narrow area is normally sufficient for this. Nowadays, having a PhD is an indispensable prerequisite for becoming an academic (lecturer in the UK, professor in most other countries) and it certainly indicates the ability to study a particular problem to a considerable depth. However, the PhD does not necessarily require breadth, which most universities also demand from their academics. Indeed, many European universities require a second formal qualification, often called habilitation, that demonstrates breadth within the subject for obtaining the right to teach at university level. Universities in most countries require at least corresponding postdoctoral experience as proof of breadth.

Planned or serendipitous? Sometimes the novel aspect of a PhD thesis can be well planned, for example, when the effectiveness of a newly admitted medication is studied in combination with certain other therapies, and it is obvious from the outset that this setup alone will create new insights. However, more often than not it is unclear at the beginning of a PhD study whether, and if so which, novel aspects are likely to be discovered. Sometimes a PhD student starts with a particular idea for solving a problem and ends up discovering something unrelated. Hence, much of the writing in a PhD thesis is back-fitted to what has been achieved during the PhD studies. As a consequence, text that was written for intermediate examinations or progress reports needs to be rescinded or radically rewritten for the actual PhD thesis. Although the requirement of a *distinct contribution* may sound frightening for many PhD students in their first year, once they look back on what they have done over the years of their study, they invariably find that they have genuinely discovered, researched and illuminated new aspects of a given problem area.

The viva. In the UK the viva is a discussion between the PhD student and two independent experts in the field behind closed doors, optionally in presence of one supervisor as observer (who regulations normally forbid to take part in the discussion) and possibly in presence of a panel chair who oversees the formal aspects of the viva but is unlikely have read the thesis. The two experts act as examiners, whose task is to check whether the thesis represents the student's

[1] `http://www.london.ac.uk/fileadmin/documents/students/postgraduate/research_facilities/MPhil_PhD_regs_from_Sept_2009.pdf` last modified Sept 2009, accessed on 3 March 2011

own work, whether or not the student can competently talk about it and explain or defend the choices made, whether the presented work is sufficient in novelty, quality and scale, and that the data, the experiments and the analysis support the drawn conclusions.

Voila! That's it.

This paper cannot help with the first years of PhD study and the underlying work that goes into a PhD thesis — that is another story. It is all about writing a successful thesis and enjoying the viva.

Let's get started!

2 Writing a thesis

At the point of writing up, PhD students are unlikely to have written many theses: typically one to document an undergraduate final year project and possibly another one at Master level. The PhD thesis spans a much longer time scale and broader subject area, though, and requires correspondingly more care than any previously written thesis.

2.1 Structuring the thesis

A thesis normally has the same first-level structure as any research paper:

1. Introduction
2. Motivation
3. Related work
4. Experiments
5. Conclusions

Of course, there are many variations to this core structure. Most frequently, the part of the thesis with own contributions is expanded to two to three chapters. There is much freedom: a PhD thesis can have different parts, for example for theoretical and experimental work, or different parts for different methods.

Consistent and coherent narrative. Ideally, PhD work leads to publications before the thesis is written. It is immensely satisfying to enter a viva with four strong publications under the belt: how can examiners dispute novelty, interestingness, contribution to a field, critical thinking etc when the peers in a field have already certified these qualities by accepting papers at competitive conferences or journals? The temptation then is to take these papers, write an overall introduction and conclusion, and staple these together. While this is acceptable or even desirable in some places, most notably The Netherlands, the British system generally rules this simple option out. A British thesis needs to be a smooth monograph with a consistent and coherent narrative. Ideas, critical analysis, data, experiments and evaluations from own previous publications can normally be reused, provided that these were created in the context of the PhD studies and declared, but these elements have to support a bigger, new narrative and need to be freshly interwoven to generate a consistent holistic work. This often means that a single literature review (the related-work section above) is generally favoured over a number of chapter-specific literature reviews that had their origin in the related-work sections of published papers. Different circumstances are likely to create different scenarios, but the general advice here is to plan for a monolithic consistent and coherent narrative rather than gluing together one's own previous work.

Select publishable material. Most universities expect "publishable" quality of the thesis as a whole or of its constituent parts. This is the litmus test for the selection of material to go into the thesis. Are the reported experiments rigorous? Are the data they use accepted by the research community? Would the literature review, the methodology, the setup and the conclusions withstand peer review? Note that publishable does not necessarily mean published. Although I generally advocate that the thesis should contain all that was done during the studies, one really should select material that survives critical scientific scrutiny. The thesis is not a place for material that could not possibly be published independently elsewhere.

Make headlines informative. Another thing to note is that above structure headlines (introduction, motivation, related work, experiments, conclusions) tell us nothing about the content. I generally prefer headlines such as

1. Night-time traffic lights and colour-blindness
2. Accident statistics in Europe
3. Red-green glasses, audio warnings and licence restrictions
4. New special glasses: a comparative study in Denmark
5. Strong evidence for effectiveness of special glasses

Which set of headlines is more informative? I made this example up, of course, but it illustrates my argument for "content-type" headlines, which I generally prefer over generic, nondescript headings. Besides keeping tradition I can only see some weak arguments for using generic headings: with these, at least one knows where to locate the main contributions of the thesis. Good descriptive chapter headlines, however, also make it clear whether a chapter is repetition of known work or part of the central contribution. Ultimately, everyone has a choice of how much to convey in the title of a section.

Avoid isolated subsections. Theses normally have a three-level structure as two-level structures are often too coarse. Occasionally, some sections warrant a fourth level of subdivision. It is a matter of taste whether or not to number these, but I find 4-level-numbers look rather awkward. Irrespective of how deep one subdivides the thesis, it is considered bad practice to produce a subsection without an sibling: Rather than having a single subsubsection called "3.2.1 Ratings" within a subsection "3.2 Recommender systems" consider renaming the subsection to "3.2 Recommender systems and ratings" while not using a third level here.

2.2 Signposting

One of the common sources of misunderstanding when reading a thesis is generated by wrong expectations and, hence, the danger of going off on a tangent. Clear signposting within a thesis helps keeping everyone on track.

2.2.1 The title of the thesis

The thesis title is the most obvious signpost. There are many different schools about what should be in a title. At one end of the spectrum there are grand titles claiming to cover a whole field ("On Music Information Retrieval"), while at the other end of the spectrum the title gives a specific characterisation of what exactly has been done ("The analysis of the effect of modified Melfrequency Cepstral Coefficients for the improvement of retrieval tasks within late-medieval lute collections of up to 50 pieces"). With a title like this one may wonder whether the experiments were carried out during full-moon or not.

There is good justification, though, for each type of title: for example, the more generic titles are justifiable if the main contribution of the thesis is the critical reflection on the whole field. A number of distinct contributions also justify a generic title if that is the least common denominator of the contributions. I think that the best title informs about the problem studied and the type of methods used. Something like "Semantics and Statistics for Automated Image Annotation" is concise, punchy and informative.

Do not to overclaim or undersell. For example, do not use "multimedia retrieval" in the title when every single collection actually used solely consists of images; do not exclusively put "Markov random fields" in the title when the thesis systematically examines a whole arsenal of statistical techniques. Good titles only ever crystallise towards the end of the PhD study. Watch out, though: some universities have regulations that make it very hard to change the thesis title once a certain point in the process has passed.

2.2.2 High-level summary

The abstract is a high-level overall summary of the thesis. Think of it as the elevator pitch[2] for experts. It should not dwell on the motivation of the problem, but instead give a realistic and sober picture of methods, achievements and limitations. One of its main formal characteristic is that it needs to be self-contained without reference to papers, sections of the thesis, figures or equations, so it can be stored in abstract databases and distributed independently from the thesis.

Although it is best to write the abstract as the very last thing, an early draft is typically needed at the time when the supervisor invites potential examiners. This draft should be revisited and checked once again before submitting the thesis. As examiner I have learned to read the abstract another time after having finished reading the thesis: Far too often I find that some of the good intentions were not carried out, but are still promised in the abstract. This part of the thesis warrants particular attention to detail and style, as it will be circulated much wider, and read more often, than the whole thesis. The best abstracts I have seen summarise the thesis in the first sentence and expand on this in the remainder of the page. Also note that the university's degree regulation is likely to specify formal restrictions, eg, that the abstract must not exceed 300 words.

2.2.3 The introduction chapter

The introduction chapter spells out the problem area, motivates its study, makes the research hypothesis (or hypotheses as the case may be) explicit, details the contributions of the thesis, and contains an explicit walk through the thesis structure. If the reader (examiner) later on is met with a surprise then something must be wrong in the introduction: this should not happen. Putting a list of one's contributions into the introduction is a good way of managing the reader's expectations. If the PhD work generated published papers, the very end of the introduction (or the contribution section of the introduction) is a good place to list them. Only list those papers that relate to the PhD thesis and that are of reasonable quality. If necessary, make it clear how they relate to the thesis. The introduction should be at a beginner level and ought to be accessible to first-year PhD students.

[2] Meeting an important customer in the elevator by chance one has 20 seconds to sell something.

2.2.4 Signposting as you go along

Many theses repeat a small signposting exercise at the beginning of each chapter (apart from the introduction, of course) to keep in with the good practice of signposting. Some add an explicit introduction section to each content chapter as well as an intermediate conclusion section to reflect on the chapter's value. The need for chapter-level signposting depends on the complexity and length of the thesis and their chapters; it can vary from chapter to chapter.

I would advise, though, to stop signposting below the second level of subdivision: it can become irritating to be told three times what one is going to read before being given the opportunity to read it, particularly, if the depth of contents does not meet the expectations that have been raised through signposting. I personally prefer to discreetly weave chapter level signposting into the text before the first chapter subsection without having an explicit chapter subsection called "introduction". There is something to be said about explicit conclusion-type subsections for each content chapter: they can tremendously aid the flow of arguments in an unobtrusive way.

2.3 Code of practice for research

A PhD thesis must follow the accepted code of practice for research. Much of this is part of the PhD education within workshops or is learned through the practice of being an active lab member and taking part in research projects. I will pick up a few aspects below that are important for writing the thesis, mainly how to cite sources, how to paraphrase, what plagiarism is, how important reproducibility of academic work is and ethical issues that may crop up.

Good reasons to cite. I cannot stress enough that the entire thesis must be *your own account* of the presented academic work and its context. This is actually the most important rule of writing a PhD thesis: no cutting and pasting, and no copied artwork, figures or tables from others either! In every case where an author presents other people's ideas, methods, criticisms, conclusions etc, it must be made unmistakably obvious who the originator is and what the original source is. There are a number of good reasons for this: First of all, references give credit to the originator. Secondly, they distance the author from the ideas, methods etc — after all, they may not have been that great or accurate! Thirdly, they give the reader an opportunity to follow up details in the original source. Fourthly, their use demonstrates the author's knowledge in an area. Fifthly, they open up the possibility of critical own analysis and dialogue with the presented ideas of others. A good PhD thesis contains many references to original work and demonstrates a thorough, critical understanding of the context of the own work. Any one of above reasons should trigger a clear reference in the text that points to a suitable original source. Every non-obvious claim needs a reference that backs it up, perhaps with the exception of common knowledge that undergraduates in the discipline can reasonable be expected to have. If in doubt, it is advisable to put in a suitable reference, for example to a particular section or page of a textbook (not the whole book).

Literal citations. Word-for-word citations from a source should normally not be necessary in science or engineering subjects. If, on rare occasions, a literal citation becomes desirable then it needs to be clearly indicated by using double quotes, by changing the font and with an immediate reference to the source before or after the quote. If the word-for-word citation is longer than a line then it is usually formatted as an indented block. Copied artwork, figures or tables must reference the source straight away in the caption, ideally naming and acknowledging the permission from the publisher. The only exception to this is artwork, figures or tables that *you* (not co-authors) prepared for previously published papers. Still, even in this case, one may have to obtain permission as it is usual practice to transfer the copyright to the publisher. If the work was jointly authored with others and the artwork in question was produced by one of the co-authors this must be treated in the same way as artwork from an unrelated paper.

Paraphrasing. Sometimes it is necessary to summarise other people's work. This should be done in one's own words to adapt to the style and needs of the thesis, and to demonstrate an understanding of it. A reference to the original work is indispensable to give credit to the originators and to differentiate the paraphrased work from own work. The reference for paraphrased work normally only ever stretches to the enclosing paragraph. If it is necessary to use more than one paragraph for describing other work, the reference needs to be repeated. It is always best to clearly separate own work and thoughts from paraphrased work of others, for fear of creating ambiguities. If own artwork replaces artwork that others have produced earlier, then there is normally no need to ask the publisher for permission, but it is likely that the inherent intellectual ownership must be acknowledged with a reference to the original artwork.

Plagiarism is usually defined as copying or paraphrasing from others without reference to the source. Plagiarism in a thesis is completely unacceptable. It does not matter if this affects only a particular phrase, a sentence, a paragraph, a page or artwork. It is also irrelevant if the plagiarism was carried out with the intention to deceive or is just owing to sloppiness or carelessness. Collecting material and descriptions from other papers in one's electronic notebook during a literature research in year one may seem innocent enough. Reusing these when writing up the thesis having forgotten they come from elsewhere still constitutes plagiarism: no intention is required for the fact of plagiarism.

Assuming Uthor and Xpert conclude in their landmark 2008 paper "Dead ducks cannot fly when their body is floppy" then a sentence such as "The elasticity of poultry after life causes crashes during flight" in someone else's later work without reference to the work of Uthor and Xpert (2008) must also be considered plagiarised, even if there is little overlap in vocabulary.

I once came across a particularly perfidious case where a submitted research paper cited a reference in the following manner 'Assuming for the moment that X is known, we can show that Y can be obtained through the so-and-so algorithm (RefAuthor et al 2000), and ..." followed by a only very slightly paraphrased copy of a whole section from RefAuthor et al (2000). Clearly, there is a reference to the source, but the wording suggests that the reference only extends to the so-and-so algorithm, while the copied passages all appear to be the original work within the research paper, not that of the reference. What the authors should have done instead is writing something to the effect of "We deploy the model of RefAuthor et al (2000) faithfully, which we introduce here from the original source in full for the sake of being self-contained." In addition, each paraphrased paragraph should have had one reference back to the original paper by RefAuthor et al. I think this case is perfidious, as technically this does not meet the definition of plagiarism, because there is a reference to the source. Nevertheless, it has all the hallmarks of academic misconduct, as the reader is tricked into thinking there is much original work here when in fact there is not.

The problem with plagiarised parts of a thesis is that they appear to have been created by the thesis author, who receives the recognition for the work by others and at the same time deprives them from deserved recognition. This is unethical, unfair and stands in conflict with an ability to carry out independent research. It is customary that PhD candidates sign a declaration of originality when submitting their thesis that all submitted material is their own except where referenced (or something to that effect). Plagiarism clearly contradicts this declaration rendering it false. Knowingly signing a false declaration with intent to gain an unfair advantage is commonly considered to be a defining element of fraud, which can then become a matter for the courts.

Academic misconduct. Plagiarism in itself, even without proving deception on part of the author, is considered to be academic misconduct for which an awarded PhD can be revoked years or even decades later: The awarding institution basically confirms that the bearer can carry out independent research to academic standard; hence, the institution has a right, even duty, to revoke the PhD if new evidence shows that the criteria for awarding it had not been met at the time. A PhD can be also revoked for all other sorts of academic misconduct, for example for making up data, experimental results vel cetera. If there is evidence for deception, for example through its scale or its manner, academic misconduct including plagiarism can be considered fraudulent in a legal sense. Authors, who were not cited, can sue the fraudulent thesis author for damages. I expect that the offender's employer would also be in a position to sue for damages, while professional bodies might withdraw their recognition as well. In theory, I imagine, government bodies who awarded funding for research based on fraudulent papers could sue the perpetrator with a view to recover the awarded grant sums. Any single one of these consequences is far worse than not obtaining a PhD in the first place, as high-profile cases regularly demonstrate. Academic misconduct including plagiarism damages academia in general, and the discipline and university where it happens in particular.

Overlapping publications. Most journals and conferences require their submissions to be original and unpublished work that is not considered for publication elsewhere. The reselling of the same results in different venues is generally thought of as unprofessional and frowned upon. This practice is sometimes called self-plagiarism, though technically this is a self-contradictory term, as plagiarism is defined as using *other* people's work without reference.

There are cases, though, in which overlapping publications are justified, for example, to reach a different audience that would otherwise not know about the particular research. In any case, it is vital to be completely transparent about the fact that a publication contains material by the same authors published elsewhere and equally vital to cite the original paper. For example, a submission to a Neural Computing conference could stress "We present experiments and results previously published in a paper *Can dead ducks fly?* at the International Conference for Poultry Motion (Tudent et al 2008), which we have recast as Neural Computing question and which specifically elaborates on the function of the brain during flying activity, something which we have not covered previously."

It is also not atypical that an idea was first published as a poster in a conference with limited supporting evidence, then later gets re-examined in a conference paper with more thorough experiments, and is later expanded on in a journal article that adds some theoretical underpinning and further comparisons against rival approaches. This practice is sometimes even part of a publication strategy that has the added benefit of offering PhD candidates an opportunity to present before an international audience or of being part of a conference where trends are set.

One problem with duplicate or overlapping publications is that it is common practice for authors to transfer the full copyright to the publisher, so that a duplicate paper is bound to create a legal issue. It is dishonesty at this level (assigning copyright one no longer has and implying originality when in fact the same was already published) that makes strongly overlapping publications a problem of academic conduct. Another problem with overlapping publications is that they project a false image of productivity in a CV.

Once published, it is normally no longer the authors who own the material. It can happen, though, that publishers decide to reprint a paper in a different context, because they own the legal right to do so. A book chapter written for a particular book may be used by the publisher to bolster up another book even without telling the authors. There is the moral question of the customers, who have bought both books expecting completely different contents only to find out that they have paid for the overlapping chapter twice. However, this is for the publisher to worry about, not the authors. Coincidently, authors can always try to negotiate which rights they confer to the publisher. I once was asked to write a book chapter, but because I meant to write a book about the same subject, I agreed with one publisher that I grant only non-exclusive publishing rights and keep the copyright of the book chapter, and with the other publisher that I could use the material of the book chapter in the full book. As always, transparency is key.

In the UK it is common practice that PhD candidates keep the copyright of the thesis with the proviso that a certain number of copies must be kept by the institution and with the encouragement that the thesis should be put into the electronic institutional repository. Hence, there is normally no legal problem to publish parts of the PhD thesis afterwards elsewhere. It is also commonly accepted in the UK that a PhD thesis may contain the results of previously published papers, in fact some supervisors define the scope of a PhD thesis as everything one did during the studies.

Reproducibility is a key aspect of science. To this effect, a lab-book must be kept, which details the scientific journey throughout the PhD studies. This lab-book together with the collected data, experimental design, the programmes for data processing and demonstrators, if any should be preserved. If, for example, the underlying data were published together with the experimental design and the programmes, then other labs could reproduce the experiments from the thesis. It is good practice to have the lab-book, data, programmes and demonstrators available on the day of the viva, should one of the examiners want to inspect any of the underlying work. After all, if the PhD candidate lacks the information to faithfully reproduce own work, how can one expect this work to be written up so that others can reproduce it? The thesis may need to describe some of the data or experimental procedure in great detail, perhaps more so than appears right for an interesting read. Particularly long and dry lists or repetitive procedures, which are vital for the work to be reproducible, can be moved to an appendix of the thesis, if this level of detail is not necessary to follow the main arguments of the thesis.

Ethical issues and privacy. Most ethical issues are specialised and have their very own rules of conduct, for example how to treat human tissue, or how to conduct medical research trials. These rules are outside the scope of this paper, but if applicable must be strictly adhered to within the thesis. One ethical issue that appears relatively often, and throughout disciplines, is the use of subjects in surveys or case studies. These can expect that their contribution is appropriately anonymised before any publication, and the thesis counts as publication for that matter. The presentation of case studies at companies may need to be discussed with the company, so that the thesis does not inadvertently give away sensitive business information. Any data sets that are prepared for inclusion in the thesis as appendix or that are published alongside on the internet must be carefully scrutinised with respect to privacy, business sensitivity or other ethical issues that may exist.

2.4 Content chapters

This is the core of the thesis and contains the fruits of your hard work. Here you explain, why the problem that you introduced is relevant, who has done what to tackle this, where opportunities lie and how you went about to contribute in this area. Here you present your thoughts, your criticism, your understanding, your experimental design, your summary of the data, how you analysed them, your visualisations, your insights and your conclusions. In theory all of this should have been planned from the outset, but in reality nothing might have gone this way, and you may need to create a story around the work that you have done.

2.4.1 Motivation

This chapter explains why the problem is worthwhile looking at. What is involved? Why is it difficult? Who would benefit from a solution of this problem?

2.4.2 Related work

Spot the gap. The requirement of a coherent thesis as a whole means that it is best to disentangle the related work sections of one's own papers and write an appropriately named chapter about related work in the area from scratch. This serves two purposes: to give an overview of the state of the art and to convincingly identify a gap of knowledge, techniques, solutions for the problem that the thesis tackles. This gap, of course, is a great justification for the work done to fill it.

Update the literature review. This is also a very good opportunity to repeat the literature research that was carried out at the start of the PhD studies. Occasionally, I see a fresh PhD thesis where the newest citation is three years old: for me this is a clear indication that the necessary homework has not been carried out before writing up and that the thesis is already out-of-date at the time of examination!

2.4.3 Experiments

Most science and engineering PhD theses have an experimental section.

Justify choices. Here, it is customary to explain the methodology and the experimental design, including the data sets and how they were obtained. Every choice ought to be explained and convincingly justified.

Negative results. It is a common view that most of the experimental work carried out should be reported with the notable exception of experiments that yielded "negative results". These are results such as "Method X does not improve Y" and are particularly hard to sell: Seemingly negative results might have come about by errors in experimental procedure or plain incompetence, and one is rarely in a position to exclude the possibility of these, in particular if the methods or experimental procedures are complex and involved.

I hasten to add that some disciplines have a tradition, even duty, to report on negative results, for example when a medical trial with a certain drug does not demonstrate improvements. It is easy to see why such results must be published as they are vital for the understanding of the effects of the drug. To give an explicit example: If a certain drug is known to lower cholesterol levels in the blood, but does not demonstrate an overall drop in the mortality rate of patients, then something else must be going on, and this is important to know. Let us assume that 19 independent and large long-term studies show that this drug does not statistically significantly lower the mortality rate, and a 20th study shows that the same drug *does* lower the mortality rate. It is vital to consider *all* studies to arrive at conclusions. Everyone who knows the $p = 5\%$ rule of statistical significance, will immediately understand that in twenty independent comparisons of same-performance quantities, 5%, ie, one in twenty, is allowed to imply significance where none is there. It would be immensely unethical, and scientifically unwarranted, of pharmaceutical companies to try to sell the drug citing only this one specific study.

So, my advice about not reporting negative results only refers to one-off complicated experimental designs, where a number of things could have gone wrong, and where the negative result cannot convincingly be pinned down to a particular reason. Clearly, this situation must be distinguished from one where a large number of identical independent experiments were carried out, and only the single one of them with the desired result was published. The latter would be unethical, and indicate gross academic misconduct, indeed.

One does not necessarily have to beat the state of the art. Some PhD work tries to beat the current state of the art in, say, automated image annotation with new methods. In the best of all worlds, one of the new methods beats the performance of the best previously known algorithms. If the experiments do not bear such results out, will it mean that the PhD work is void through a lack of novelty? Fortunately not. One can trade this situation off by a failure analysis, a very thorough literature study and a correspondingly more critical analysis of the state of the art.

Statistical significance. Always carry out appropriate statistical significance tests of results where appropriate. Think about and justify the choice of statistical test. Also note that if A cannot be shown to be statistically significantly better than B, one should not talk about one being slightly better than the other based on the performance numbers (this is a common mistake).

Importance of experimental design. Some value in a PhD thesis is drawn from careful experimental design. It is best practice to only change one parameter at a time; to use datasets that are publicly available or at least make datasets available; to describe experiments in a way so that they are reproducible; and, particularly in Computing, to set up experiments in an automated batch fashion. If different data sets have been used for different parts of the thesis owing to historic reasons, ask yourself the question how much the thesis gains in value by re-running the experiments on the same data sets. In some cases a whole new set of interesting conclusions can be drawn by comparing approaches on the same data set. Having a batch programme that runs the experiments is not only desirable to re-run the same algorithms on different datasets later on, it also is a very good way of recording and preserving the experimental conditions of experiments. These programmes should be kept together with the experimental data in the electronic lab book.

Selection bias. In its simplest form selection bias refers to statistical samples from sets being made in a non-representative way. Many a user study at Computing departments is biased towards the technology savvy young male students. A more sinister form of selection bias appears when the experimenter deliberately looks for and selects subsets of datasets for which an algorithm works well without reporting those subsets for which the same algorithm does not work well.

There is also the less often recognised pitfall of selection bias when selecting the best algorithms for a certain problem or dataset: Very often a set of algorithms is evaluated over the dataset, each of which with its own performance number, and the best-performing algorithm is then selected to be the most appropriate. This comparison alone can introduce a selection bias in the sense that it may have been random peculiar effects that make this particular algorithm look good on this particular data set. Rather than claiming the winning algorithm has the observed performance p, one ought to verify this on a separate dataset that has not been involved in the algorithm selection step. To give a concrete example, assume that we are in the business of predicting gold-price movements over quarter years from historic data. We use the data from the year 2012 and compare a number of algorithms (each predicting 4 bits, namely whether the gold price goes up or down in each of the four quarter years). In fact, the best algorithm is likely to have had a perfect prediction over the four quarters. If not, one could easily imagine to add to the comparison set all 16 possible algorithms that predict fixed up-down patterns no matter what the underlying input is. One of these fixed-pattern algorithms will definitely predict exactly the observed pattern of gold-price movement and, hence, be declared the winner. Does it mean that this simple fixed-pattern predictor is likely to have the same (perfect) performance on all other data sets? Of course not! For the same reason a fund manager, who has brought about above-average profits year-on-year for ten years in a row, need not be one with above-average profits in the next year. If only there are enough fund managers (say, at least $1024 = 2^{10}$) and all generate random profits there is bound to be one who has been consistently good over the last ten years.

It is easy to see in this example that the selection of algorithms based on performance can overfit the data, but this bias is easily forgotten in more complex situations. In machine learning it is customary to split the data set into training and test sets in order to avoid overfitting of the algorithm's parameter on the training set. In the same way, when comparing algorithms on a test set, one needs to reassess the winning algorithm on an independent validation set to avoid this type of selection bias.

2.4.4 Tutorial value

In some sense, the two most important readers of the thesis are the examiners. However, a thesis is not only an instrument to show one's own understanding of the subject matter; it is also a good way to introduce fellow researchers or future PhD students to the area. Hence, any thesis should have a good tutorial value and be self-contained to a certain extent.

Write for a 1st year PhD student. In general, the reader of a thesis can be expected to have a fairly general knowledge, perhaps at undergraduate level, of the discipline. Anything above this level or anything that is critical for the understanding of the thesis ought to be explained, at least with an overview of the most important points and one or more references to further reading. For example, if the thesis frequently utilises Fourier transforms and Discrete cosine transforms, it makes sense to go to the trouble of explaining both and repeat known facts as far as they are relevant for the presented work. It is also desirable to be consistent in the expectations of the reader: It would look odd, for example, to explain the Fourier transforms but not Discrete cosine transform.

Avoid being too detailed if it is common mathematical knowledge, can easily be gathered form introductory text books, or is otherwise general pre-university knowledge. For example, a section that painstakingly explains trigonometric functions only serves to illustrate that the author is uncomfortable with mathematics.

If in doubt, it is better to provide the necessary background in the thesis and make the thesis self-contained.

2.5 The exit strategy

Self criticism. The last chapter in a thesis is normally about overall conclusions and discussions of the presented work. One of the goals of a thesis is to demonstrate critical thinking, and this includes being critical towards one's own work. I consider it good practice to have a subsection in the last chapter that deals with the limitations of the presented work. This allows one to focus on the boundary of what has been achieved and creates another natural opportunity to demonstrate one's own understanding of the own contribution. It is important to exercise critical power for a realistic assessment of one's contributions; one should not be bashful nor exaggerate.

Future work. Some of this analysis will lead to thoughts about future work. Good research always opens up new questions, and a subsection about future work helps voicing thoughts on this. Almost inevitably one will have had more plans for research than one could manage. Giving an informed view of how best to continue the particular line of enquiry demonstrates academic research credentials. Sometimes the quality of research carried out shines more by illuminating the path ahead than by what has been solved.

End on a positive note. Sometimes the work has led to deeper insights or gave rise to recommendations, and these should be put into the last chapter. It is best to end a thesis on a positive note; the limitations should come first, then future work before reflecting on what has been achieved or could be achieved by following the strand of research that has been opened in the thesis.

2.6 Front matter, back matter and appendices

2.6.1 The front matter

This is the part of the thesis that appears before the body of text, ie, that comes before the introduction. It starts with a title page, the format of which will normally be prescribed by degree regulations or can be gleaned from other theses of the institution in the library. Then there is an optional page with a dedication followed by an abstract (see Subsection 2.2.2). The degree regulations are likely to require a particular declaration, eg, that this is all one's very own work and that no part of the thesis has been submitted as part of any other degree or qualification. Normally, any acknowledgements come after this. The table of contents is a very useful instrument for signposting and should be put in the front matter as well.

Acknowledgements. The thesis is a formal piece of writing, and the dedication and acknowledgements are one of the few occasions where you can let your character, personality and life philosophy shine through. Once written, try to read the acknowledgement with the eyes of yourself ten years down the line, with the eyes of an employer five years later and with the eyes of the examiners and colleagues now. By the time one write the acknowledgements, some PhD students may be in the most exhausted or most elated mood ever, but if one expressed this in an untempered way, how would it read later on? If someone told the world that doing a PhD was the most stressful and frightening experience ever, how will prospective employers interpret this when the same person applies for a research post? An acknowledgement gone wrong is like a facebook entry that one comes to regret later. This is not the right place to give negative feedback to the environment or dwell on insider jokes. In particular, one should refrain from what can be seen as hidden messages to advisors, colleagues in the lab, the university or the host country. Always check the acknowledgements for unintended messages: for example, a four-page acknowledgement section thanking everyone in the address book including the cats and dogs of the neighbour's nephew, while barely spending half a line acknowledging the role of one's advisors, may give an unintended message of a broken supervision structure.

Consistency checks. Many word processing templates make it easy to provide a list of figures, tables and algorithms, but I have never gained any benefits from these. I consider them to be superfluous, and my advice is simply to leave them away. The single one role that I can see for these lists is in a near-final draft stage, where they give a great opportunity to check whether one has been consistent in the style of all captions. Are the captions sufficiently informative? Do they share the same style of capitalisation? It is also wise to check the table of contents, which is a vital part of the front matter, for consistency of spelling, capitalisation, and most importantly whether it tells a good story in itself.

Numbering. The pages that come before the introduction are normally counted in Roman numbers (i, ii, iii, iv, . . .), while the page number is reset to Arabic 1 with the introduction. The front matter must have an even number of pages in case the thesis is printed double-sided later on; if necessary, insert a blank page before the introduction. The reason for this is that all books must have odd page numbers on the right side and even numbers on the left side when opened in the middle. When I read a stack of unbound sheets of paper I know that at the end of an even numbered page the sheet is spent and I need to take another one from the stack. Some prefer all chapters to start on the right hand side in a book and, hence, force the first page of every chapter to be on an odd page number.

2.6.2 Bibliography and citations

What to cite? When referring to specific work by others it is mandatory to cite, preferably articles in highly authoritative primary sources, eg, high-impact journals or conferences. Sentences that begin with "It is commonly accepted that . . ." normally demand at least one, better more, citations in support of this claim, which may also come from well-respected secondary sources such as text books. Interesting work sometimes appears first in workshop papers, often in unpolished form, then in conferences or journals with more substantive arguments. Before citing a workshop paper, it may be worthwhile checking whether an extension to this work has appeared in better venues in the meantime.

Which sources to avoid? If possible at all, reduce the number of references to blogs, web sites, unrefereed magazines and tertiary sources such as newspapers, encyclopediae, personal communication, wikipediae etc. The main problem with these references is that they tend to be either not archived in their original form or not peer reviewed or neither. When citing URLs it is good practice to state when you as author have last accessed this URL. Content to web-pages changes over time, and knowing an "accessed on" date gives the reader a fighting chance to figure out what might have been there on that day, for example using the wayback machine[3].

Read all sources! It is paramount that one has read every single paper that one cites in the thesis and also to verify that it is of good enough quality to warrant the citation. Even if the rest of the world cites a particular paper in support of a particular claim, how can one be sure this is really so without having read it? If the thesis author finds it difficult to gain access to a source, why should a reader fare better? I noted once that a thesis had cited a spoof paper, which was only half a page long and had appeared in the MIREX 2010 AMS task submissions. Its sole purpose was to document a trivial random baseline, something for which no citation is actually necessary or desirable. The fictitious author of the paper had the name Rainer Zufall, meaning "Pure chance" in German. The paper claimed him (Rainer is a male first name) to be visiting researcher of a respectable and well-known place. I wonder how many unwarranted citations this Zufall guy manages to accumulate owing to uncritical citations.

Only indirectly cite unread sources. There are only rare circumstances under which it is acceptable to cite a resource without having read it. These may occur when an argument is supported by a paper that ultimately relies on a resource that is no longer available or was published in an inaccessible language. It is only correct to then indicate this indirect citation as it is, for example, "Author (1988) traces the name of this method back to the original Latin edition of Monk and Other (1234)".

Chicago style. There is a multitude of different citation styles, and each discipline has its own tradition and preference. If I have a choice I generally prefer the Chicago style (Author

[3] `http://www.archive.org/` last accessed on 12 Apr 2011

year) and I think this style is particularly germane for PhD theses: reviewers are likely to know the literature and most players in the field. Authors and year are in many cases sufficient for the experts to recall the actual paper. If the reference style just uses nondescript numbers then reviewers need to look up every single reference, which puts an unnecessary burden on them.

Activate authors. It is acceptable, even desired, to make authors of references the subjects in a sentence, eg, "Uthor and Xpert (2008) showed that dead ducks cannot fly". In fact, this method can be used with any citation style, even the pure numeric list style, and is likely to improve the readability of the thesis. It is commonplace to only use surnames with this method. If the author list contains three or more people the remainder of the author list can be contracted to "et al", an abbreviation of the Latin "et alii" meaning "and others". For example, author references look like "Einstein (1905)", "Einstein and Bargmann (1944)" or "Einstein et al (1941)". Einstein did not publish many papers with multiple co-authors, but if he had written a second paper in the year 1941 with more than one co-author, and if both papers had been cited in the same source, these would appear as 1941a and 1941b, respectively, to break the ambiguity. It is important to be meticulous with the spelling of names. Einstein's 1941 collaborative paper was with Valentine Bargmann and Peter Gabriel Bergmann, and although there is only one letter different in Bargmann and Bergmann these are different people. It is also important to be precise in whether a paper was written solely by one author or two authors (in which case both surnames are spelled out) or by a longer list of authors. Only then can *et al* be used to shorten the list, and this is normally consistently done for three or more authors. Some style handbooks require three authors to be spelled out still. For all these reasons it is best to utilise a programme that manages the bibliographic list and references to it.

The other typical use of citation is to support an argument where it is made; here is an example (Author et al 2003). I consider it bad style, though, to use references as objects. One might think that a sentence such as "The dead ducks in (Uthor and Xpert 2008) could not fly" looks acceptable, but this changes when the publisher style prescribes to set references as footnotes. Then the same sentence becomes "The dead ducks in^{117} could not fly", which looks decidedly odd.

Create consistent and meaningful bibliographic entries. Each entry in the bibliography section should contain the full author list (no et al here!), title, publisher, year of publication (with a letter, if necessary, to break otherwise ambiguous citations) and page numbers at the very least. After all, having read each of the papers that is cited, this should be easy. Try to be consistent how papers appear in the bibliography.

My personal preference is to be concise and leave away much unnecessary information such as the number of times this conference has taken place, but still provide cues as to place and month of a conference. This might help some to remember a paper by context. My preference goes as follows: Supply the unabridged journal name for journal papers. For conference papers give the full conference name with the usual acronym of the conference and the place in brackets. Drop the year after the acronym unless the year of publication of the proceedings differs from the year the conference took place. Do not mention the location of well-known publishers (ACM, IEEE, Springer, Addison Wesley etc). Do not mention the editors of papers in conference proceedings. Capitalise the title of the paper or book as one would a sentence, ie, capitalise the first word, and the other words as they would appear in a sentence. Capitalise all words of journal names or conference names except for articles, prepositions and conjunctions. The Springer proceedings series "Lecture Notes in Computer Science" is sufficiently well-known in the Computing field to be abbreviated as Springer LNCS followed by the volume number. If you plan on publishing your thesis as electronic book, consider putting in a DOI link. Here are two examples for my preference:

S Tudent, A Uthor and E Xpert: *Can dead ducks fly?* In International Conference on Poultry Motion (ICPM, Urbana-Champaign, IL), Springer LNCS 4321, pp 456–466, Jul 2008. DOI: 10.1007/3-54045199-7_99

A Uthor and E Xpert: Why dead ducks cannot fly. *International Journal of Poultry*, 22(3), pp 265–287, 2008b. DOI: 10.1080/13658810701626277

In any case be aware of conferences that have changed their name. For example, ISMIR's full name changed from "International Symposium on Music Information Retrieval" via "International Conference on Music Information Retrieval" to "International Society for Music Information Retrieval Conference" within the span of a decade while keeping the acronym.

2.7 Notation, glossary and index

It is always useful to create a notation list for own use while writing the thesis explaining the meaning of each mathematical variable and non-standard symbol. This list can be useful for the reader too, in particular, if it refers to the page, where the symbol is introduced. The same is true for concepts and acronyms, which go into something called a glossary together with a small explanation and where they were defined first. Clearly, both a notation section and a glossary are optional; everything that could go in there should be explained in the text at the time of first use. If they are created, the best place for them is rather at the end of a thesis than at the beginning, where they might intimidate or irritate the reader who feels the need to go through them before tackling the thesis. I have seen a few vivas, where the examiners required the PhD candidate to add a notation and/or glossary to the thesis because s/he was using symbols, concepts and acronyms gratuitously, and the examiners felt they had a hard time remembering everything that was defined somewhere in the text.

Links. Electronic versions of the PhD might benefit from links of mathematical symbols, concepts and acronyms to appropriate places. I find it extremely useful to interlink these document internally, but also not to put extraordinary attraction to the fact that there are links. Rather than underlining these internal links or colour them brightly, I prefer to use a slightly different colour for the links (say a dark blue or a dark grey that stands in harmony with the black colour of the text) for fear of distracting the reader otherwise.

2.7.1 Appendices

The role of appendices is to provide space for supportive material that would get in the way of a good narrative, is not really necessary to understand the thesis, but may, for example, be necessary to reproduce the experiments in the thesis. The thesis might, for example, present machine learning models for automatically recognising 500 different concepts. Rather than listing all 500 concepts in a table, one could chose 20 representative concepts for a table to give a flavour of them and list the complete list of concepts in an appendix. Many regulations have a word limit for a PhD thesis, but typically the material in the appendix does not count towards this word limit. In the same way, examiners are regularly not expected to read or examine on the material in the appendix, but they can pick up observations they have made.

2.8 Look and feel, tone, grammar and style

A PhD thesis in Computing is an examination work that delivers the main evidence for the ability to execute scientific research independently, ie, that you have got what it takes to be a good scientist. Hence, there are some restrictions as to the style of the thesis. Some universities also have format restrictions as to minimum margin and the font size. Often regulations require the thesis be printed one sided and in double line spacing. This gives examiners space to place comments and questions on their copy.

Neutral tone. The thesis should deploy neutral vocabulary based on observation, facts and analysis. There must also be a clear separation in place between observation and interpretation, in general, and value judgements in particular.

2.8.1 Active voice

It is best to write as much in active voice as possible, something which Rupert Sheldrake advocated in a 2001 New Scientist article "Personally speaking".[4] Scientific publishers encourage the use of active voice with similar arguments. *We measured the lift of dead ducks in a wind canal* just reads so much better than *The lift of dead ducks was measured in a wind canal.*

The only problem in a thesis is that using "we" in active voice evokes the natural question by the examiner "Who do you mean by 'we'? You and ...?". I have heard this question a couple of times and have had to ask this in vivas myself.

For all ideas, work and the experiments that are genuinely yours (and a fair proportion should normally be yours in the thesis!) I suggest using "I/my" in the thesis. This is particularly important when you want to emphasise your very own idea or contribution. I realise that much scientific work in Computing is based on collaborations, which typically is also reflected and documented in joint authorship of scientific papers. As a consequence, there will be a wide span of ownership: Some of the discussed ideas will be genuinely your brainchild, some will have been developed in discussions with your advisors and collaborators, and yet others will have been outrightly suggested by others. It is good practice to agree with your supervisor as to which contributions warrant using the first person singular. In any case, use common sense when attributing ideas, methods, work, papers: Never call a paper that was jointly published with others *my paper*, even if you did most of the writing. Never claim an idea, process or methodology your own through the use of language when it is not, be it that it is commonly known or that previous publications show it is really someone else's.

The thesis is a monograph written by yourself as a single author, and the proper use of "I/my" allows you to express which elements are genuinely yours.

Refrain from using the royal "we" (ie, when you actually mean yourself). This sounds grand, remote and stilted. There are still many cases, though, where "we" is appropriate, even necessary:

- When including the reader in the process, eg, *In the following chapter we* (meaning the reader and the author) *will work through several examples of code-breaking with frequency analysis.*
- When reporting directly on joint work: *Tudent et al (2008) showed that dead ducks cannot fly. We conducted these experiments in a wind tunnel with frozen ducks from Waitrose. Since then we changed the source of the ducks to Aldi and Lidl, but the conclusion stayed the same (2009b).*

Some scientists still advise against the use of active voice, because they think that science should be objective implying that the identity of the experimenter is irrelevant. Although I disagree

[4] http://www.sheldrake.org/Articles&Papers/pdf/personally_speaking.pdf, last accessed 17 Mar 2011

with this position, I recommend not to overdo the use of "I/we". When using first person singular, make it count for the very own contributions.

There are many ways of using active language without "I/we" pronouns: *Fig 1 clearly shows that the lift of dead-duck wings is significantly smaller than* ... as opposed to *It has been shown that ... (Fig 1).* Another good example would be *Algorithm 1b achieves a much better performance than the baseline* to avoid the focus on the experimenter. *I achieve a much better* ... is a bad example of unnecessary focus on the experimenter, and phrases like these should be avoided. If one wants to emphasise that this is the own algorithm, one could for example write *My Algorithm 1b* ...

In summary

- Use active voice wherever possible
- Use figures, algorithms, tables, data, authors of cited papers as subject in active voice
- Use "we/our" when reporting on joint work or when including the reader
- Use "I/my" when indicating your own genuine contribution, but sparingly elsewhere

2.8.2 Be consistent and precise

Adopt commonly accepted terms and if necessary define them at their first use. Always use the same term for the same object or concept. Novels may benefit from using a varying degree of synonyms to lighten up reading, but in a thesis about image retrieval, always call images images, and not pictures, photographs, pictorial matter, icons, pics etc. Be acutely aware that many terms have a different meaning in different communities, in particular when one of the examiners comes from a slightly different area. In Information Retrieval, for example, the performance of algorithms is often taken to be its *effectiveness*, ie, how well relevant documents can be retrieved. In the database world, performance typically refers to *efficiency* of retrieval, ie, the time taken to retrieve documents addressed by the query. So, it might be a good idea to write about effectiveness or efficiency, as the case may be, rather than about performance.

Know the meaning of jargon. When you cannot define a term, think twice before using it. This does not mean that one should define every term that is used, but one should be able to. The danger by assimilating jargon is that one might use it in a wrong way or in a way that has a different meaning to different people. In my book, the worst offenders are "more scalable", "noisy data", "non-linear effect" and variously "the probability", "the likelihood" and "the chance". If I were to receive one pound Sterling every time I see these when it is clear from the context that the author could not define what they mean, I would be a rich man. I will go through some examples in more detail.

The technical definition of a *scalable* algorithm is one where the asymptotic requirements of resources (CPU time, memory, disk space, number of servers etc) can be bounded by a linear function of the problem size. Simplified, if one doubles the problem size, no more than double the resources are needed to solve the problem. So, algorithms are either scalable or not scalable, and almost every use of "more scalable" is ill-conceived. I once read a PhD thesis that claimed a particular algorithm was "quite scalable", because, as the author helpfully explained, it had a runtime behaviour of $O(n^2)$. With the common definition of scalable, this algorithm is not scalable at all!

Surprisingly often *non-linear* is used as a synonym for magically, somehow woefully complicated. To start with, there is a simple definition of *linear* that a surprising number of Computing graduates get wrong: A function f is linear if it is additive, ie, $f(x+y) = f(x) + f(y)$ for all x and y, and homogeneous, ie, $f(\alpha x) = \alpha f(x)$ for all suitable α and x. This notion requires that the variables must allow addition and scaling, and indeed linear functions are those that preserve the structures of vector spaces. Non-linear only means this: that a certain function is not linear in one or more of its arguments. First of all, one should be clear which quantity is non-linear in respect of which input variable, and second of all one should be specific about which quality of the underlying structure it is that makes life difficult for experimentation, observation or mathematical treatment.

Noisy data are far too often cited as culprits for experiments gone wrong, in a similar way as non-linear effects are. Almost every occasion, where noisy data are mentioned, this is accompanied by a lack of understanding or lack of explanation what exactly is meant by this term. In its proper sense noise is signified by unwanted random changes in measurements brought about by the environment. By having a good understanding of the source of the noise one is often able to separate a useful signal from the noise. In data-driven Computer Science sometimes datasets are made use of that were created and collected for a completely different purpose. For example, when tagged flickr images are used as ground truth for automated image annotation it may well happen that an image tagged "rose" shows a person, or colour, and not the plant. This has to do with insufficient data-cleaning or the inappropriate use of a data set that is otherwise perfectly in good order. To say that the images or the tags are noisy is misleading and distracts from deficiencies of the experimental setup. What should be done is to explain where the data set came from, what the original purpose was, how it was envisaged to be used, and to estimate the quality of the data for the particular purpose by, say, looking closely at a sufficiently big random subset. The presence of tags that do not correspond to objects within the image, and vice versa, can thus be quantified and, even better, its effect on the machine learning methods for automated image annotation can be studied by gradually adding similarly wrong tag-image pairs to the data set and see how the predictive power of the algorithms suffers. This is much better than shrugging ones shoulders and pointing to "noisy data" in the way of explaining experimental outcome.

As a last example of jargon, I would like to mention the improper use of the term *the probability*. The main problem is that probabilities come from random processes, and these need explaining beforehand.

For example, assume that an image collection of $n > 0$ images is tagged each with an arbitrary subset of $m > 0$ unique tags. In this context alone, the phrase "the probability that image i is tagged with label t" has no intrinsic meaning. Technically, we are looking at a set of image-tag pairs. If it was a particular random process that was responsible for creating a joint distribution of image-tag pairs then this process defines a probability p_{it} for each (i, t) pair, indeed. Vice versa, any set of non-negative numbers p_{it} that sum to one defines a random generation process. If, however, we are just given a fixed tagged image collection, then a particular image i would either be tagged with a particular tag t or not. There is no randomness involved at this level, and we can look up the answer by inspecting the data set. In fact, the whole data collection can be coded as binary matrix b, whose element b_{it} is one or zero, depending on whether or not image i is tagged with tag t. We can answer questions such as, given a particular tag t and a uniformly randomly chosen image i, what is the probability that the image is tagged with t? The answer is $\sum_i b_{it}/n$. We can also answer the question, given a particular image i and a uniformly randomly chosen tag t, what is the probability that image i is tagged with this tag? The different answer is now $\sum_t b_{it}/m$. The summation is over tags t here, not images as previously. This example shows that we can have easily different answers for the interpretation of "the probability that image i is tagged with label t", and we have not yet looked at complicated random processes for selecting images or tags at random!

The important lesson is that each use of "the probability" needs a painstaking explanation of what is random (the data or the access to the data or something else), and also which distributions give rise to the random process. Bertrand's paradox is a text book example for why one cannot be explicit enough about this point: Given a circle and its inscribed equilateral triangle, what is the probability p that a randomly chosen chord of the circle is longer than a side of the triangle? Depending on which random process selects the chord there are different answer to this question. If both end-points of the chord are selected uniformly randomly from the circumference of the circle then $p = 1/3$. On the other hand, each uniformly randomly chosen point from the inner part of the circle uniquely defines a chord by taking it to be its midpoint; randomly choosing a chord in this way results in $p = 1/4$. There is yet another "natural" method of choosing a chord at random which yields $p = 1/2$. This is why there is no such thing as *the* probability.

2.8.3 Grammar

English grammar is descriptive, that is, it is acceptable to write in the same way as others commonly use the language. Of course, English is the predominant language of scientific publications in almost every corner of the world. Correspondingly wild and varied is the usage of grammar and, in particular, punctuation. Given that publishers have to be cost efficient, there are fewer editorial checks than is desirable. Hence, papers that have managed to get published in academic outlets are rarely a good role model for the use of grammar and language. If you are not a native speaker, and even if you are, you might want to consider having your final version professionally proof read by a copy editor.

A certain lack of attention to grammar is acceptable, provided the clarity of what one wants to express does not suffer, which has to be the overriding principle. Other than that it is important to maintain a consistent house style.

The most important punctuation rule. Lynne Truss wrote a moderately funny book about punctuation "Eats, shoots and leaves" (Profile Books, 2003). It provides a good overview of how punctuation can help to clarify the meaning. If there is only one punctuation rule that you can manage to learn and follow then it should be the one that indicates the difference between non-restrictive clauses and restrictive clauses. I best give an example: "The experimental batch contained 100 test tubes. I analysed all tubes, which had frosted caps." This is vitally different to "... I analysed all tubes that had frosted caps." In the first case I would have analysed all 100 test tubes of the experimental batch. I would also disclose that every one of the test tubes had a frosted cap. In the second case I would only have analysed those test tubes that happened to have frosted caps, which for all we know might have been any number between 0 and 100.

Always make up your mind whether a clause is non-restrictive (in which case it ought to be a "which" clause that is set apart through commas; the clause gives additional information that could be left away without changing the meaning of the main sentence) or restrictive (in which case it ought to be a "that" clause without a comma). Although it is technically acceptable to use "which" in a restrictive clause (no comma) I would discourage this use. Either use "which" (comma) or "that" (no comma) to be absolutely clear whether or not the clause restricts the scope of a previous noun.

2.8.4 Graphs and illustrations

Always provide your own original illustrations. If it seems easier to copy illustrations from elsewhere, reconsider. If it is really, really important to copy, gain permission, cite the source in the caption of the figure, and acknowledge the rights of the creator and copyright owner in a credits section at the end. The caption of re-created artwork must also cite the source from which the idea came, otherwise the illustration is in danger of being seen as plagiarism.

Script all graphs. It is best to use a single powerful tool, for example, Gnuplot, to create all the graphs in a thesis. This a great opportunity to have a unified appearance of all labels on axes, bar graphs, pie charts, line graphs, scatter plots etc. Most screenshots of most tools such as spreadsheets show disastrously bad quality for inclusion in a thesis. The trouble is that computer monitors operate at 72 dpi resolution, while printed material is normally 600 dpi. Enlarged screenshots often look pixelated or have far too small labels, sometimes both. Learning to use a good graph creation tool is useful for all publications, not only the thesis. The best aspect of Gnuplot is that it is a script language, ie, one can separate the data from their appearance. The script itself serves as a self-documenting way of how the graph was created, and one can automate the creation of many graphs without the risk of copy/paste errors. Keeping all the underlying data with the programmes that created and manipulated them is good practice and supports the scientific principle of reproducibility. Using a mechanism such as Gnuplot scripts helps this tremendously.

Assume graphs are printed black and white. Although the use of colour printers is on the rise, I would still at this point in time advise to use colour sparingly in graphs, or if at all, only as an orthogonal redundant scheme: For example, one might use red dots and blue crosses, so that a black and white copy still contains all the necessary information.

Use the right type of graph. Only use connecting line graphs if there is some sort of relation and natural order on the x-axis, eg, demonstrating a development of a quantity over the x-axis (eg, time). If it makes no sense to interpolate the quantities, for example, performance versus algorithm name, then use bar graphs, not interconnecting line graphs. If one had interconnected the performance of algorithm 2 with the performance of algorithm 3, this would imply there was something that relates these two algorithms, and that there was such a thing as algorithm 2.5. When putting graphs side to side to facilitate comparisons, it is a good idea to fix the scale of the axes of neighbouring graphs to be the same.

Consistently label across graphs and tables: Use the same type of lines and symbols for the same algorithms or experimental conditions; list the algorithms in the same order in all legends; list results in tables in the same order etc. Also, try to make sure that the labels in illustrations and graphs have a consistent font and size throughout the thesis. The font of the labels should either be the same as the text font or slightly smaller. I have seen a thesis once where the smallest label was rendered in a 4 pt font size, thus illegible, while the biggest label sported a 44 pt font — double the size of the chapter headlines — and quite possibly could have burned a pattern into my retina.

Avoid screenshots. I advise against the use of screenshots, but if they are an important part of the story, there are tricks to improve their appearance: enlarge the on-screen fonts (eg, through ctrl-+ in browser windows); make windows particularly large before taking a screenshot, sometimes it is possible to use a virtual desktop with a size larger than the screen; post-process screenshots to transform offending background colours to lighter versions or invert/change the colour scheme altogether. It is good practice to look at the printed illustrations and ask the following questions: Can I see the necessary detail? What do I expect the reader to take away from this figure? Are my claims and conclusions born out in the way the figure appears? Have I chosen the most appropriate representation? Does my representation unfairly distort the data?

Create meaningful examples. When illustrating structures, say, a set of documents, create meaningful examples. In particular, avoid lazy copies of the same nondescript object. In order to illustrate the data format of, say, documents, deploy real examples. Avoid graphs or descriptions such as "title$_1$: bla, content$_1$: bla bla bla; title$_2$: bla, content$_2$: bla bla bla; title$_3$: bla, content$_3$: bla bla bla; ...". Always create illustrations imaginatively and strive to be as informative, useful and realistic as possible.

2.8.5 Mathematical notation and conventions

The number one rule here is to be consistent and use the same symbol for the recurring quantities. For example, if image size is an important consideration throughout the thesis, make it a particular variable, say number n of pixels, and keep it that way. Try not to use n for anything else. In the same way, once you decided to use the number of pixels as indicator for size, avoid later using another quality of images, say, width, as indicator for size.

In the following I will explain a few mathematical conventions. My running example will be of images that are labelled with concepts from a fixed set.

Variable names. Mathematicians always prefer short, typically one-letter symbols for items because these utilise space well and avoid formulae being broken up over lines too often. This is in stark contrast to programming, where long and meaningful variable names are encouraged.

It is customary to explain the meaning of variable names when they are introduced. For example, mathematicians are likely to present the Euler characteristics χ as

$$\chi = v - e + f,$$

where v, e, and f are, respectively, the numbers of vertices, edges and faces in a given polyhedron. As a programmer one might want to simply write something like

$$\chi = n^{\mathrm{vertices}} - n^{\mathrm{edges}} + n^{\mathrm{faces}}.$$

This has two disadvantages: Not explaining what one means by n^{vertices}, n^{edges} or n^{faces} while relying on the mnemonic nature of the labels is considered hand-waving. Secondly, any prolonged reasoning about the quantities takes up a fair amount of space. Hence, I do not encourage index labels of this kind.

Lowercase single letters in italics stand for variable objects in their own right (numbers, integer index numbers, images, words) and can stand for function names (f, g etc). They also stand for collections of objects such as vectors and matrices (see below).

Uppercase single letters in italics are often used for sets and for variables that are fixed most of the time such as the number of images or a parameter of a method (for example the radius of a Gaussian blur function).

Any sequence of letters in italics is taken as the product of the single letter variables, eg, $max = m \cdot a \cdot x$.

Normal (Roman) fonts almost always are for names of functions that are longer than one letter such as max, sin, cos, log, tf or idf. To all intents and purposes, these functions are constant in their meaning, ie, they are defined once, if not commonly known, and never redefined or overloaded with anything else. Roman font is also used for mnemonic labels (which I discourage).

Fonts and font variations. The standard mathematical font is italics, while normal (non-italics) font is used for function names and labels. Occasionally, Greek letters are used for angles or parameters of methods. All I would say is that, unless there are convincing arguments, you should not introduce other font variations than these three. In particular, I urge you to stay clear of boldfacing and fancy fonts to create a multitude of symbols. The single one exception to this is the convention to use uppercase blackboard letters, eg, $\mathbb{N}$ or $\mathbb{R}$, for certain sets.

Every field within mathematics, computing and engineering has its own conventions, and it is difficult to advise more generally on naming strategies. Given that there are not too many letters in the alphabet, it may be a bit tricky to not double-book variable names. The most important lesson is to introduce variables when you use them the first time in this particular meaning.

Sets and variable types. Always, always, always, always, always, always specify the type and range of a variable. This is done by explicitly defining or at least describing the set of possible values. Standard numerical sets are the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of natural numbers, the sets $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ of integers, rational, real and complex numbers, respectively. Subsets are defined with the "such that" operator, written as : or |; here are some examples:

$$\begin{aligned}
[a, b] &:= \{x \in \mathbb{R} \,|\, a \le x \le b\} \text{ with } a, b \in \mathbb{R} \\
(a, b) &:= \{x \in \mathbb{R} \,|\, a < x < b\} \text{ with } a, b \in \mathbb{R} \\
S &:= \{n \in \mathbb{N} \,|\, n = m^2 \text{ for a } m \in \mathbb{N}\}
\end{aligned}$$

The first two examples define closed and open real intervals as usual, while the third example defines the set, named S here, of squared natural numbers. Note that in these examples x and n are "local" variables that, like summation indices, have no meaning outside the definitions.

Of course, sets do not have to be numeric only. It is perfectly fine to write something like "Let D be a set of images that one might find in an image database or image collection." It is also usual to define sets *en passant* as in the following example: "Our dataset D consists of images each of which is associated to a particular subset of a set C of available concepts." Please note that C is introduced as "the set C of available concepts" and not as "the set of available concepts C"; the latter is wrong.

Tuples and vectors. Mathematics is all about examining structures, which are mostly encoded in terms of sets and functions; sometimes functions are also called operators. One of the most fundamental set operators is the Cartesian product $A \times B$ of two sets A and B; it contains all ordered pairs, where the first component is an element of A and the second element is in B. For example, $\{0, 1\} \times \{5, 6, 7\} = \{(0, 5), (0, 6), (0, 7), (1, 5), (1, 6), (1, 7)\}$.

$A \times A$ is also written as A^2, giving rise to the definition of $A^n = A \times A \times \ldots \times A$ (n times). A^n is the set of all n-tuples with elements from A.

In the following, I will introduce two possible notations of assigning concepts from a fixed set C to a particular image. It is always best to define the type and structure of a variable when you introduce it. For example, you could say "Let $t \in \{0, 1\}^n$ be a binary vector that indicates which of n concepts are expressed in a particular image: $t_i = 1$ if the ith concept is present and 0 otherwise." It is clear from this definition that t is composed of n binary numbers, which gives meaning to $t_1, t_2, \ldots, t_n$ as variables in their own right: these are the components of the vector t. The vector t can also be understood as a particular function

$$\begin{aligned}
t\colon \{1, 2, \ldots, n\} &\to \{0, 1\} \\
i &\mapsto t_i \qquad (1)
\end{aligned}$$

that maps an index $i \in \{1, 2, \ldots, n\}$ to a binary number. This view is particularly useful, when you wish to say which concept is expressed in an image. Alternatively, one might define a function u that maps concepts (not index numbers) directly to binary numbers:

$$\begin{aligned}
u\colon C &\to \{0, 1\} \\
c &\mapsto u_c \qquad (2)
\end{aligned}$$

You see, t and u have a very similar mathematical structure. The difference is that u maps concepts directly to binary numbers, while with t you need to have an enumeration of all concepts in mind: $C = \{s_1, s_2, \ldots, s_n\}$. Basically, t_i tells us, whether the concept s_i is present in the image. All I want you to note here is that it is useful to extend the concept of a vector to one where the index can take on all sorts of countable sets rather than just a subset of $\mathbb{N}$. In programming the notation of t can be likened to work with parallel arrays (t and s), while u would be known as associative arrays, where the array index can be an arbitrary object.

To enumerate or not to enumerate that is the question. The difference between those two notations is subtle but significant. I will give a few examples of the usage of t and u. How many concepts does the image have? Answer:

$$\sum_{c \in C} u_c = \sum_{i=1}^{n} t_i$$

You can say "concept c's presence is given by u_c" or "concept s_i's presence is given by t_i". "Let r_c be the importance of concept c" or "Let r_{s_i} be the importance of concept s_i".

You will have noted that the t notation requires at times to talk about the number n of concepts, that you had to introduce the enumerated list s of concepts and that you easily come across double-indexed items such as r_{s_i}. In practice, you will find that many authors get tired of talking about s_i all the time; they then talk incorrectly about concept i, where they should talk about the ith concept s_i. Some get tired of double-indices and contract r_{s_i} to r_i. Although the attentive and forgiving reader can correct all these mistakes in their head, this notation generates some scope of confusion. If you accept that i in definition (1) and c in definition (2) are local variables that can have any name without changing the definitions, then the t notation (1) needs four global variables in practice (t, s, C and n) while the u-notation only requires two global variables (u and C). In the rare circumstances when you need to talk about the number of concepts you can use the generic set cardinality operator $|\bullet|$, so that you say $|C|$ rather than n. The reader will be thankful for not having to track so many variables.

Summarising, the direct mapping of concept membership values to binary numbers via u is elegant, and I generally prefer this notation over the one with parallel indices.

Powersets. Although the use of u with subscripts suggests u is used like a vector, one should better think of u as a function. To make this clear, you can talk about u as $c \mapsto u_c$ or $u_\bullet$. All three mean the same, namely a function that gives each concept in C a binary number. In this context it is useful to introduce the *powerset* B^A of two sets A and B as the set of all possible functions from A to B. With this notation one can simply write $u \in \{0,1\}^C$ or, equivalently, $u_\bullet \in \{0,1\}^C$ if you want to express that the argument of u is written as subscript.

Should you have come across the notion of a powerset for the first time, this notion may initially be a little confusing, but it is really not deep. As soon as you enumerate the elements of C with the list $s_1, s_2, \ldots, s_n$ you are in a position to map $\{0,1\}^C$ to $\{0,1\}^n$ component by component and realise that both notations are equivalent and yield the same structure. Mathematicians use powersets all the time — it is a very basic notation.

Matrices. This is a generalisation of the vector concept above. Rather than having one index set, a matrix will have two index sets. For example, $g_{\bullet\bullet} \in X^{Y \times Z}$ defines a matrix with values in X that uses two indices, the first with values in Y and the second with values in Z. I give an example how this notation can be used in practice:

> Our dataset D consists of images each of which is associated to a particular subset of a set C of available concepts. The ground truth is a matrix $g \in \{0,1\}^{D \times C}$ with $g_{dc} = 1$ if the image d is associated with concept c, and 0 otherwise. In other words, the ground truth matrix g is a binary matrix, where the row $g_{d\bullet}$ is a binary membership vector indicating which concepts belong to the image d. We can also interpret this vector as a membership function $c \mapsto g_{dc}$. Its inverse (a relation) partitions C into two subsets: $g_{d\bullet}^{-1}(1)$, which is the set of concepts assigned to image d, and $g_{d\bullet}^{-1}(0)$, which is the set of concepts *not* assigned to d. We denote the number of *true* or *relevant* concepts with $|g_{d\bullet}|$; technically, we may think of the L_1 norm of the vector $g_{d\bullet}$ or the cardinality of the set $g_{d\bullet}^{-1}(1)$. Finally, the column $g_{\bullet c}$ of matrix g expresses which images belong to concept $c \in C$, and we extend the notations $|g_{\bullet c}|$ and $g_{\bullet c}^{-1}$ analogously.
>
> Most multi-label classification algorithms can be represented as a matrix $f{:}D \times C \to [0,1]$ that computes a likelihood $0 \leq f_{dc} \leq 1$ that image d belongs to class c. For each image d, all concepts can then be ranked according to descending values of the vector $f_{d\bullet}$. We denote with $\mathrm{rank}^f_{d\bullet}(c) \in \{1, \ldots, |C|\}$ the position of the c-entry of row vector $f_{d\bullet}$ after sorting.
>
> We can either assign the top n highest ranking concepts to image d or chose to assign only those concepts c whose likelihood f_{dc} passes a (possibly concept-dependent) threshold. Either way, we obtain the classification through concept ranking. We call the concepts assigned in this way *labels* and call the associated binary matrix $l \in \{0,1\}^{D \times C}$, while the *true* concepts are those encoded as ground truth g.

This notation is precise, elegant and concise in that it only introduces two sets, the set C of concepts and the set D of images, and three structures, the ground truth g, the likelihood values f and the assigned labels l. The other variables c and d are only local variables to name specific concepts or images and they could have any name. Everything else is derived through notation and does not necessarily require own variable names.

Sequences. Sometimes one would like to have a sequence of numbers, vectors or matrices. These are normally indexed with superscripts as in "let $v^1, v^2, \ldots, v^m \in \mathbb{R}^n$ be m concept vectors, one for each image, each of which describes which concepts are assigned to a particular image." The ith component of the jth vector is written as v^j_i. This is not much different from the matrix notation above, but sequences are normally expected to define a certain order, ie, they imply the notion that v^1 comes in some sense before v^2. The only disadvantage of the sequence notation is that it may be confused with the notation for power, ie, if v was also an entity for which it makes sense to carry out multiplications then v^2 could be misunderstood as $v \cdot v$ rather than the concept vector for the second image.

Tensors. One can easily extend the concepts of sequences and matrices to more general tensors that can have any number of subscripts and superscripts.

Graphs. [todo]

[issue warning to check whether the introduced structure reflects the problem adequately: eg, set vs multiset]

The danger of typographic notation. The main principle in mathematical structuring is to get one's sets and functions right. A function is defined on a set, called domain, and assigns exactly one thing (element of another set called codomain) to every element from the domain. The elements of the domain can be tuples, in which case the function has more than one argument. The application of a function (f, g, addition $+$, exponentiation, implicit product) to its arguments (x, y) can be written in a multitude of ways, eg, $f(x), fx, f_x, f^x, {}^x\!f, xf, g(x,y), x+$

$y, x^y, xy\ldots$ Some of these notations are only used in specific fields, and some are typographically confused, eg, fx means that function f is applied to argument x, while xy means that the product operator is applied to two arguments, x and y. The same confusion exists between f^x (application of f to x) and x^y (x to the power of y). It is important that the context makes it clear what is what.

Note that once the function has been applied, the information about the argument has been lost: while the vector $f \in \{0,1\}^{\mathbb{N}}$ contains the information about all its components f_x, as soon as you consider a particular f_x this is just a single binary number, either 0 or 1, and this one-bit number has lost the information about x and about f. An expression such as $g(f_x) := f_x + f_{2x}$ is ill-defined. The argument of g is one of two numbers, 0 or 1. What was meant was $g(f,x) := f_x + f_{2x}$, which is perfectly well-defined.

Sometimes, a definition makes use of changing fonts, or uppercase and lowercase letters, which is also problematic: "We denote sets with caligraphic letters $\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$ and denote their cardinality with the corresponding uppercase letter $A, B, C\ldots$". Although conventions like these can work they are wasteful, because they use up a whole range of symbols. Also, does the use of a variable M imply there is a set $\mathcal{M}$ somewhere?

Others introduce conventions such as "We denote matrices with uppercase letters $A, B, C, \ldots$ and denote their elements with the corresponding lowercase letters $a_{ij}, b_{ij}, c_{ij}\ldots$". This is a similarly unnecessary waste of symbols. Even worse, some define a matrix as $A := \{a_{ij}\}$, which not only uses up two variable names, A and a, but also unnecessarily overloads the meaning of curly brackets, which are already used to construct sets, and lets a_{ij} stand for the list of all matrix elements, while the true meaning of a_{ij} is that of a single matrix element. There is nothing wrong with a being the full matrix, and a_{ij} being one particular element addressed by the indices i and j: this is all one needs.

The last bad example of typographic notation is one where the name (not the value) of the argument of a function is an important part of a definition: "Let $g \in \{0,1\}^{n \cdot m}$ be a matrix, g_i be the ith row vector, and g_j be its jth column vector". What is g_3 then? The 3rd row vector or the third column vector or something else? The other problem is, of course, that the author wanted to define an $n \times m$ matrix, but has defined a vector of length nm. It clearly is much better to first define the index sets N and M and introduce "$g \in \{0,1\}^{N \times M}$, where $g_{i\bullet}$ is the row vector associated to $i \in N$ and $g_{\bullet j}$ is the column vector associated to $j \in M$.

Punctuation of displayed maths. Every formula on a standalone line must be treated for punctuation as if it were an inline phrase in normal text. This means that in a fair number of cases a displayed formula must end in a period or a comma. As in normal text, there should be no space before the punctuation mark. If a standalone formula ends a sentence and is introduced by a colon then there is no need for a period at the end of it. Take care not to introduce a unwanted paragraph breaks before or after any standalone formula, in particular when the formula is in the middle of a sentence.

Never mix algorithmic notation with mathematical notation. The mathematical notation in your thesis helps clarifying concepts, structures, ideas and processes. Its implementation can differ from how you defined mathematical objects. For example, a sparse binary term-document matrix (expressing which document contains which terms) may be mathematically convenient for exploring relations, but your implementation would actually store a rugged array akin to an index in a book that gives you immediate access to the list of documents that contain a particular term, which is what you need at run-time for a search engine.

It is very wise to clearly separate algorithmic implementation from the mathematical structure. This is particularly true for a line such as $i = i + 1$, which is hair-raising in a mathematical context but acceptable as part of an algorithm (though I still would write $i \leftarrow i{+}1$ to express this assignment in an algorithm). In particular refrain from programme language specific notations such as $t[1]$ or $u\{c\}$ for vector elements or $x * y$ and $x\hat{}y$ for multiplications and exponentiation. Use the mathematically conventional t_1, u_c, xy and x^y instead.

Summarising, mathematics is like a strongly typed programming language with single letter variable names. Sets and functions are the predominant means to describe the underlying mathematical structure of objects. If you use much mathematics consider a notation summary as an appendix of the thesis.

2.8.6 Typesetting tricks of the trade

I highly recommend LaTeX as the tool of choice for writing a PhD; it is arguably the best typesetting system for a project of this size and complexity. Latex's biggest plus is its expert typesetting of mathematical formulae, and it has beautiful mathematical fonts that blend in with the text fonts. Another advantage of Latex is its remarkable stability over time. Only recently I deployed Latex on a nearly two-decade-old paper and it generated an identically looking output. Try this with Wordstar 7.

The few theses I have seen typeset in something else than Latex inevitably suffer from weak references (sometimes missing them in the bibliography, sometimes listing papers that were never referenced), an overly complicated way to create bibliography, index, table of content etc, and most of all from shoddy mathematical typesetting, particularly for inline mathematical terms and equations such as this simple trigonometric formula: $\sin 2\alpha = 2 \sin \alpha \cos \alpha$

There are now front-ends to Latex, eg, Lyx, that do a reasonable job in giving near-what-you-see-is-what-you-get experience.

Crossreferences and navigational links. Latex provides an excellent system of crossreferences for chapters, sections, subsections, equations, tables, figures, theorems, algorithms, you name it. In particular, a companion programme called Bibtex knows about scientific references, and comes with a number of popular citation styles. Latex's hyperref package allows you to easily include live implicit navigational links into the final pdf document. You can put external links into your documents as well and embed images and videos into the final document, too. In fact, the package backref adds to bibliographic references on which page(s) in the thesis each reference was used. This can be useful to see if you overly rely on a certain paper and cite it over and over again. This also gives the examiner who inspects your bibliography section an easy way to check how you use certain references. When you deploy the backref packages with the hyperref package, then you can click on a reference in the electronic thesis and be transported to the paper, where you can click on the right page number after the reference to be brought back.

It is vital for a project of the size of a thesis to always define labels and make symbolic references within the text. Latex resolves these to the appropriate values when typesetting the document, and you can move your material around as much as you like and still have the correct printed crossreferences. This is particularly important for citations to ensure the correct format and spelling of author names. A good bibliography programme lets one use symbolic references to the full paper, the author names only, or the year only. For example, with Bibtex's Chicago style one could write `\citeauthor{bivector-fields}'s \citeyear{bivector-fields} landmark paper`, which is automatically changed to "Einstein and Bargmann's 1944 landmark paper" from the bibliography database without worrying that one remembered the spelling of the authors correctly or the year of publication. There is also `\citeN{bivector-fields}` that uses a citation as noun and produces "Einstein and Bargmann (1944)"; the `N` in `\citeN` stands for noun. The other standard way of putting in a reference into the text is through `\cite{bivector-fields}`, which produces the regular form, in Chicago style "(Einstein and Bargmann, 1944)".

Number 1 pitfall: Function names. Despite all the strengths of Latex to just do the right thing for typesetting, there are still a few pitfalls that the unwary regularly get caught out by. One is caused by naivety to type something like `$$ sin 2a = 2 sin a cos a $$` for the above trigonometric formula without realising that a sequence of letters is always typeset as the implicit product of single-letter variables. In fact, the space character in math mode is ignored, because Latex manages all spacing itself. As a consequence above looks horribly wrong:

$$sin 2a = 2 sin a cos a$$

You must tell Latex that sin and cos are function names. Function names for known, existing functions are set in normal Roman font, and not in italics (which is the default for single-letter variable names). Luckily sin and cos are predefined as `\sin` and `\cos`. Now `$$ \sin 2a = 2 \sin a \cos a $$` will create the desired look of

$$\sin 2a = 2 \sin a \cos a.$$

If a function name is not known by Latex, it is very easy to define it yourself: If you want functions named tf and idf for term frequency and inverse document frequency, say, these would be defined somewhere at the beginning of the thesis as `\def\tf{\mathop{\rm tf}\nolimits}` and similarly for idf. You can then use `\tf` and `\idf` in the same way as `\sin` and `\cos`.

If you prefer that the subscripts are typeset under the function name rather than next to it, you can drop the `\nolimits` command in the definition: `\def\argmin{\mathop{\rm argmin}}`. With this you can typeset correctly

$$j^*(a) = \mathop{\rm argmin}_{j\in\{0,\dots,n\}} |a_{\bullet j}|_2$$

as the column index for which the column vector of the $n \times n$ matrix a has the smallest Euclidean length $|\bullet|_2$.

Typesetting products. The rule about implicit products is strongly anchored in mathematical typesetting, so much so that using the asterisks $*$ looks very amateurish: $\sin(2*a) = 2*\sin(a)*\cos(a)$. Please avoid the use of $*$ for simple products unless you want to own up to the fact that you have never opened a maths book.

Simple products should always be expressed as the juxtaposition of the symbols without using an explicit operator. If you decide to use a multiplication operator for clarity, please use a central dot `\cdot` only. Sometimes you have to add brackets, because an implicit product has a higher precedence than an explicit product:

$$1/ab = 1/(a \cdot b) \neq 1/a \cdot b = b \cdot 1/a = b/a$$

The very observant will have noted that the formula $\sin 2a = 2 \sin a \cos a$ actually contains two different implicit products one without any space for the highest precedence ($2a$), higher even than applying the monadic sine function, and one implicit product created by the small space between $\sin a$ and $\cos a$, which is of less high precedence than the monadic function application. Thus is mathematical convention.

Typesetting labels. If you feel you must use descriptive labels for clarity of variable names as in $n^{\rm faces}$, for example, (but note I discouraged their use earlier), please remember that you typeset the index label in Roman font (eg, `$n^{\rm faces}$`, so the labels do not look like products of variables.

Blackboard symbols for mathematical number sets. Originally, the basic mathematical number sets were typeset as boldface capital letters. In mathematics lectures these would be written on the blackboard with double strike letters, such as $\mathbb{N}$, $\mathbb{R}$, or $\mathbb{C}$. The latter have found their way back into text books and are now considered the standard way to typeset basic mathematical number sets. I have defined simple Latex commands `\def\IN{\mathbbb N}`, and similarly for `\IR` etc to represent the set $\mathbb{N} := \{1, 2, 3, \ldots\}$ of positive integers, the set $\mathbb{R}$ of real numbers and so on. You only need to include `\usepackage{amsmath,amssymb,amsbsy,amsfonts}` in the preamble of your thesis to make these definitions work.

Typesetting function definitions. Here is an example for how a function is typeset correctly:

$$\begin{aligned} f{:}\,\mathbb{R} &\to \mathbb{R} \\ x &\mapsto f(x) = x^2 \end{aligned} \tag{3}$$

The Latex code is as follows:

```
\begin{align}
  f{:}\,\IR &\to \IR \notag\\
  x &\mapsto f(x) = x^2 \label{def-f}
\end{align}
```

The curly brackets around the colon are important to make Latex remove the space normally uses around it (the colon is otherwise a binary operator). The `\,` gives just the right amount of space after the colon. Please also note the difference of the two arrows used in the function definition. The simple arrow `\to` is used between domain and codomain of a function, while the `\mapsto` arrow expresses how elements of the domain are mapped to some element in the codomain.

Capitalisation in Bibtex titles. For some reason titles fields of certain types of publications including “inproceedings” conference papers are automatically changed to lowercase. To counter that names such as Markov are downcased, you ought to use braces ({M}arkov) in the title field.

Abbreviations. I personally do not put a period after abbreviations for a number of reasons: it looks nicer, and the period does not get confused with the end of a sentence while there is hardly any scope for ambiguity by dropping this period. If you prefer abbreviations with periods, as the majority of people do, it is important to indicate to LaTeX when the space after it is a normal space and not the end of a sentence, as in “`Mr.\ Alfonso`”. Alternatively, there is also the non-breaking space to avoid a line or page break at this position, as in “`Ms.~Alfonso`”. The main reason for either of these techniques is to avoid LaTeX deploying a larger space after a period that indicates the end of a sentence.

Units. There should always be a non-breakable space (~) between the value and the unit, eg, 10 m, typeset as `10~m`. The unit byte[5] is abbreviated as B. Some use lowercase b to abbreviate the unit bit. In contrast to this, I think that bit should never abbreviated to avoid confusion with byte and because bit is already an abbreviation of binary digit. There are two systems of prefixes, one decimal and one binary: While k (kilo), M (mega), G (giga), T (terra) and P

[5] As an aside, information is a physical quantity, which the true SI enthusiast measures in Joule per Kelvin: one bit is $k\log(2) \approx 9.569938 \cdot 10^{-24}$ J/K, where k is the Boltzmann constant. 1 pJ/K ≈ 12.2 GiB ≈ 13.1 GB.

(peta) are prefixes for 1000, 1000^2, 1000^3, 1000^4 and 1000^5, respectively, the IEC has declared Ki (kibi), Mi (mebi), Gi (gibi), Ti (tebi) and Pi (pebi) as the standard prefixes for 1024, 1024^2, 1024^3, 1024^4 and 1024^5, respectively.

3 The Viva

3.1 Selecting the examiners

[todo]

3.2 The roles of the participants

[todo]

External examiner

Internal examiner

Chair

Observer

Candidate

3.3 Preparing for the viva

One of the most common questions of the candidate is "Should I prepare a talk?" When I am examiner I normally do not want to see a talk or slides as I have read the thesis and know it quite well. Of course, every examiner is different, and in theory they can ask for anything (talk, slides, lab book, implementation, demos, raw data sets) but they ought to have said so beforehand. I consider a reasonable time frame for examiners to make these requests to be one week before the viva. I would advise to find out from the chair of the viva a week before the viva, whether there was such a request. If not, one can reasonably assume that nothing more is expected than being there and engage in a discussion with the examiners.

Everyone will want the candidate to be at ease, and a typical way to start off a viva is the question "tell me about the most significant aspects (your contributions or similar) of your work in 5 to 10 minutes". It is a good idea to prepare something along these lines in varying length from a 30 second elevator pitch to a longer version spoken freely without aid.

If as a candidate you happen to have a demo, a talk, some instructive slides, key diagrams etc, by all means have them ready in an unobtrusive way, so that you only need to switch on a monitor or projector to access these. There is a bit of chutzpah involved in offering a demo during the viva without being asked to give one, and I would suggest if one really wants to show a demonstration, to promise one for *after* the viva.

It is essential to bring a copy of the thesis to the viva. This copy must be exactly the same as the submitted one. This is to be able to answer questions such as "At page 17, line 3 you say... what do you mean by ...?". It may be a good idea to bring a spare thesis for the observer, so that he or she can follow the discussions and so that there is a spare one in the (very rare!) case of an examiner forgetting to bring their copy along.

The examiners are meant to give detailed feedback for suggested changes after the viva, but it is useful for the candidate, too, to make notes of some of their points during the viva (bring pen and paper). The candidate may find it hard to keep up with taking notes. It is helpful if the observer can persuaded beforehand to agree to take some notes, too. As observer I always take notes in a viva, particularly as in this role I am not allowed to say anything unless prompted by the examiners.

Some more things the candidate can do before the viva: making sure there is paper to write on, perhaps also a flip-chart with pens (test that they actually write), making sure the room is well aired, has a comfortable temperature, that the table and chair layout is conductive to an oral exam, that there are some glasses, a jar with tap water etc. Ideally, the chair acts as host of the meeting, ie, looks after these aspects and generally the wellbeing of everyone (scheduling a break, ensuring there is is coffee, tea, biscuits, etc). However, chairs may be busy or may have been parachuted into the viva from a different part of the university. Agreeing with the chair beforehand whose task it is to look after the hosting aspects of the viva is a useful piece of preparation. This becomes even more acute if the external examiner participates remotely with video or teleconference facilities. In such a case it must be clarified beforehand who sets up the conference call and who supplies technical support should the need arise.

That is normally all there is to prepare for the candidate.

Although any of above seems self-evident, things can go wrong: I once attended a viva as one of two external examiners. Because there was no internal examiner, the university required someone internal to chair. This person had misgivings about having been drafted into the chair role, and was intent to demonstrate to everyone what a waste of time this was. The other external was supposed to come in through a video conference. The supervisor visited specifically from a sabbatical, but had other meetings lined up rather than participating in the viva as observer. Initially, the viva room was locked and no one had a key; after gaining access, the room turned out to be overheated; the conference system did not work; there was no technical support; there was not even water. The poor candidate found himself to be debugging the video conference system, and had to start the viva with a disinterested chair, who read the newspaper, with a booming voice from the remote external examiner from loudspeakers (we never got the video to work) all without support from the absconding observer. A bit of communication and preparation beforehand could have alleviated some of the problems.

3.4 On the day

My basic advice is: Be yourself, and talk about your work. The examiners should not ask things outside the remit of your work. It really should be one of the most enjoyable days of your life, because never again will you be able to talk to specialists about your work for that length of time.

One key thing to remember is that there is some implicit negotiation going on wrt the changes that might be required of you. During the viva, always think about which points you are happy to concede and which ones you want to put up a fight for. One should not fight when it is about trivial things: for one, it will not cost one much work to concede small points, for another fighting these has the danger as coming across as unnecessary bellicose. Also one should not all too easily agree suggestions for bigger change either: this is an opportunity to show independent thinking, critical thought and for standing one's ground. Again, good judgement is key: examiners can get irritated by PhD candidates, who defend obvious deficiencies. Like in sports it can be useful to raise one's hand and admit "mea culpa". Even then one can often negotiate whether a deficiency can be seen as outside the scope of the thesis or has to be mended.

All of these types of discussions (in good doses) make for a great viva. It is not helpful to come across as pushover or as overly aggressive and defensive. Summarising, there is a bit of a balancing act involved and good judgement on which fights to pick.

3.5 The corrections, if any

[todo]