Scientific Writing for Computer Science Students

Wilhelmiina Hämäläinen

Course material
September 20, 2006 (revised May 2021)
Department of Computer Science
University of Joensuu

Preface

This material was originally prepared for the Scientific writing course (2006) of the International Master's Programme in Information Techology (IMPIT) in the Department of Computer Science, University of Joensuu, to help the students to write their masters theses in English. The style advices are based on existing literature on scientific writing (e.g., [1, 2, 4, 3]), but the instructions have been applied to the conventions of the computer science field.

The contents have been proofread and slightly revised May 2021 and revision is to be continued. I wish that the result is useful for the reader!

In Kuopio, 24th May 2021

Wilhelmiina Hämäläinen

Contents

Chapter 1

Introduction

Three learning goals:

1. How to write scientific texts in computer science?
2. How to write in English?
3. How to write a master's thesis?

1.1 Goal 1: How to write scientific text is cs?

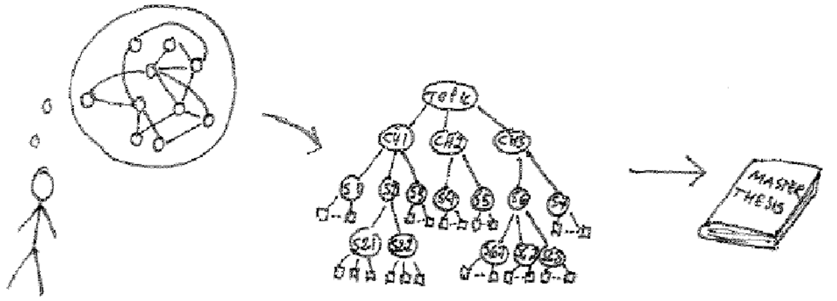
- general style
- how to use references
- equations, pictures, tables, algorithms

- useful tools (latex, bibtex, picture editors)

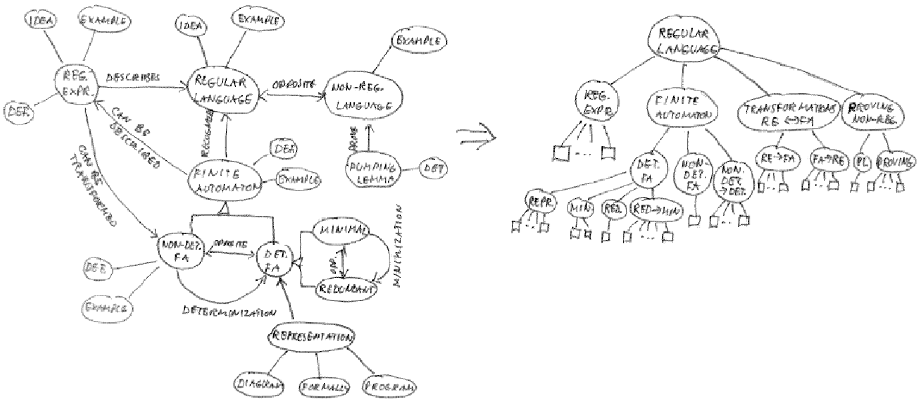
1.1.1 Problem

Writing w is a mapping from a set of ideas I to a set of scientific texts S, $w : I \rightarrow S$.

Problem: Given a set of ideas $i \in I$, produce $f(i) \in S$



1.1.2 Example



1.1.3 Instructions

1. Organize your ideas in a hierarchical manner, as a tree of ideas t ("minimal spanning tree" of the idea graph).

2. Write the tree t as text such that

- the root node of t corresponds to your topic (title)
- its children correspond to chapters
- their children and grand-children correspond to sections and subsections
- leaf nodes correspond to paragraphs (actual text)

1.1.4 Writing tree t

Each node n of tree t contains three fields:

- $title(n)$: the heading (or subheading) of the chapter, section or subsection. In leaf nodes (paragraphs) *NULL*
- $children(n)$: n's children (chapters, sections or subsections). In leaf nodes *NULL*.
- $content(n)$: description of the idea in n. In non-leaf nodes very brief, in leaf nodes longer.

The following algorithm describes how to walk through t in preorder and write it as a sequence $s \in S$ (scientific text):

1.1.5 Properties of a good tree t

- Tree t is balanced: The main chapters have approximately the same depth, i.e., the paths from the root to a leaf have approximately equal length, usually ≤ 4 or at most ≤ 5 [1]. In Introduction the typical path length is 3 and in Conclusions 2 (no sections).
- Each node in t has a reasonable number of children k: $k \geq 2$ (no orphans!) and typically $k \leq 7$ (maximum $k \leq 10$).
- For all leaf nodes n, the sizes of $content(n)$ are balanced: each paragraph contains at least two sentences, but is not too long (e.g. ≤ 7 or ≤ 10 sentences).

[1] i.e., the lowest units above paragraphs are subsections or subsubsections

Alg. 1 WriteTree(t)

Input: tree of ideas t
Output: scientific text s

```
begin
    Write $title(n)$
    if ($n$ is not a leaf node)
        begin
        // Writing an introductory paragraph:
            Write $content(n)$
            for all $u = child(n)$
                WriteTree($u$)
        end
    else
        // Writing a main text paragraph:
        Write $content(n)$
end
```

- For all non-leaf nodes n, the sizes of $content(n)$ are balanced. These introductory paragraphs (chapter or section previews) can be very brief. They just give an overview what will be covered in that chapter or section. Exceptionally, you can use more than one paragraph in the chapter previews. Notice that it is possible to skip the previews altogether, but be systematic!
- For all leaf nodes n_i in preorder, $content(n_i)$ can refer only to previously written contents $content(n_1), ..., content(n_{i-1})$. E.g., you cannot define a non-deterministic automaton as an opposite of a deterministic automaton, if you haven't given the definition of a deterministic automaton, yet. Exception: you can briefly advertise what will be described in the future. E.g., "We will return to this problem in Chapter X".

1.2 Goal 2: How to write English?

Every week we will spend some time with English grammar and expressions.

We will practice at least the following topics:

- dividing the text into paragraphs, sentences and clauses
- possessive case (expressing the owner)
- verb tense and number
- word order in sentences
- use of articles
- punctuation
- useful words and expressions

Other important topics??

Idea: personally selected exercises!

1.3 Goal 3: How to write a master's thesis?

Writing a master's thesis is not just writing, but you have to read a lot of material, make experiments, and analyze the results.

The process has the same phases as a software project or any problem solving activity:

1. **Defining the problem**: Discuss with your supervisor and define what is the problem. Try to understand it in a larger context: other related problems and subproblems. Read some introductory article about the topic or select the main books written about your topic. You can already generate several ideas how to solve it, but don't fix anything, yet.
2. **Specification**: Specify your topic carefully. Don't take too large topic! Invent a preliminary title for your thesis and define the content in a coarse level (main chapters). Ask your supervisor's approval! Decide with your supervisor what material you should read or what experiments to make.
3. **Design**: Define the content more carefully: all sections and brief descriptions what you will write in each of them. Define the main concepts you will need and fix the notations. Then you can write the chapters in any order you want. Make also a work plan: what you will do and when.

4. **Implementation**: You can write the thesis after you have read all material or made all experiments. However, it is a good idea to begin writing already when you are working. Often you have to make changes to your design plan, but it is just life! Ask feedback from your supervisor, when your work proceeds.

5. **Final work**: Check language and spelling, missing or incomplete references. Check that the structure is coherent. Write an abstract.

Note: In practice it is easier to write other chapters, if you have an introduction, which defines the problem and helps to stay focused. However, often you have to rewrite the introduction again in the end, when everything else is ready. Conclusions are also written in the end.

1.4 Scientific writing style

Main goal: exact, clear, and compact.

- Compact is usually clear!

- Other desirable properties: smooth and objective

1.4.1 Exact

- Word choice: make certain that every word means exactly what you want to express. Choose synonyms with care. Don't be afraid of repetition.

- Avoid vague expressions which are typical for the spoken language. E.g., the interpretation of words which approximate quantities ("quite large", "practically all", "very few") depends on the reader and the context. Avoid them especially if you describe empirical observations.

- Make clear what the pronouns refer to. The reader shouldn't have to search the previous text to determine their meaning. Simple pronouns like "this", "that", "these", "those" are often the most problematic, especially when they refer to the previous sentence. Hint: mention the noun, e.g. "this test".
  → See Section Pronouns.

- Avoid ambiguous and illogical comparisons. These are often due to missing words or nonparallel structures. Examples:
  "Female students draw concept maps more often than male students."

  "The students' points were lower than the average computer science students."

  → See Section Parallel constructions.
- Anthropomorphism: do not attribute human characteristics to machines or other inanimate things. E.g., a computer cannot understand data, an experiment cannot control variables or interpret findings, a table or a figure cannot compare results.
- Incorrect grammar and careless sentence structures can create ambiguities!

1.4.2 Clear

- Use illustrative headings which describe the essential content of a chapter or a section.
- Write a brief introductory paragraph or sentence in the beginning of each chapter or section with subsections.
- Divide the text logically into sentences and paragraphs.
  - Direct, declarative sentences with simple, common words are usually best.
  - Paragraphs should be logically uniform and continuous.

  → See Section Sentences
- Place the adjective or the adverb as close as possible to the word it modifies.
  → See Sections Adverbs and Word order.
- Avoid **scientific jargon** = continuous use of technical vocabulary when it is not relevant.
- Write numbers as digits when they refer to sizes or exact measurements. Otherwise the general rule is to write numbers < 10 as words. Express decimal numbers with a suitable precision. See APA pp. 122-129.

- Use punctuation to support meaning.
  → See Section Punctuation and APA pp. 78-88.

1.4.3 Compact

- Say only what needs to be said!

- Short words and short sentences are always easier to comprehend.

- Weed out too detailed descriptions. E.g., when you describe previous work, avoid unnecessary details. Give a reference to a general survey or a review if available.

- Don't describe irrelevant or trivial observations (i.e., don't mention obvious things)

- Avoid wordiness, e.g.

  "based on the fact that" → "because"
  "at the present time" → "now"
  "for the purpose of " → "for/to sg."

  Notice: "reason" and "because" have the same meaning → don't use together!

- Use no more words than are necessary. Redundant words and phrases (which have no new information) should be omitted.

- Avoid too long sentences and paragraphs.

1.4.4 Smooth

- Verbs: Stay within the chosen tense! No unnecessary shifts in verb tense within

  - the same paragraph
  - in adjacent paragraphs

  → See Section Verbs.

- Use verbs rather than their noun equivalents.

- Prefer active to passive voice.

- Avoid long noun strings!

  Hint: sometimes you can move the last word to the beginning and fill in with verbs and prepositions.

- Each pronoun should agree with the referent in number and gender.

- Transitional words help to maintain the flow of thought

  – time links: then, next, after, while, since
  – cause-effect links: therefore, consequently, as a result
  – addition links: in addition, moreover, furthermore, similarly
  – contrast links: but, however, although, whereas

- Notice: some transitional words (while, since) can be used in several meanings → limit their use to their temporal meaning! (Use "because" instead of "since"; "although", "whereas" or "but" instead of "while", when there is no time connection.)

- Use abbreviations sparingly, especially the abbreviations which you define yourself for technical terms.

  → See Section Abbreviations.

- Do not use emphasis (italics) when it is not needed. Use syntax to provide emphasis.

- Metaphors can sometimes help to simplify complex ideas. However,

  – Don't overuse them
  – Don't mix several metaphors in one sentence
  – Avoid cliches

1.4.5 Objective

- Use the 3rd person rather than the 1st person.

- Use emotionally neutral expressions, e.g., "Students suffering from dyslexia" → "students who have dyslexia"

- Use words which are free from bias (implied or irrelevant evaluation). Especially, be careful when you talk about

  - gender
  - marital status
  - racial or ethnic groups
  - disability
  - age

  → See Section Gender-neutral language.

Hints:

- Select an appropriate degree of specificity. When in doubt, prefer the more specific expression. E.g.,

  - Instead of ”man” use ”men and women” or ”women and men” to refer to all human beings
  - Instead of ”old people” define the age group ”ages 65-83”
  - Instead of ”Asian” mention the nationality ”Chinese”

- Differences should be mentioned only when relevant. Careless use of biassed words can create ambiguities.

  E.g., avoid the use of ”man” as a generic noun or an ending for an occupational title. Otherwise it can imply incorrectly that all people in the group are male.

Chapter 2

Searching, reading, and referring literature

2.1 Need for references

In scientific writing, we use a lot of references!

- All text must be justified, either based on previous research or your own results.
- It must be clear what the information is based on!
- Sometimes the entire master's thesis is based on systematic study of existing literature. The information is just analyzed and organized from a new point of view.
- The sources for scientific writing must also be scientific!

2.2 Source types

The literature sources can be divided into three groups:

1. **Primary sources**: articles in conferences and journals
   - original sources
   - the papers should have appeared in a refereed journal/conference (i.e., reviewers have checked their correctness)
   - also technical reports and other theses
2. **Secondary sources**: textbooks, encyclopedias, glossaries

- sometimes useful analysis or interpretation, but not original sources
- you can use these in the master's thesis, but only as supplementary material

3. **Bibliographies**

   - support information retrieval
   - lists of articles + references
   - scientific search engines are on-line bibliographies

Task: Can you trust the information you find in wikipedia? Why or why not? Why wikipedia cannot be used as a reference in a scientific text?

2.3 Collecting literature

Starting point: your preliminary topic.

- goal
- central concepts, theories and themes

How to proceed?

- Begin from familiar: lecture notes, textbooks
- Check if there are tutorial or survey papers on the topic – they save a lot of work!
- Check references in useful papers or books. Some bibliographies allow also to check papers that cite the given paper (e.g., in google scholar "cited by").
- Make keyword queries in scientific bibliographies (e.g., ACM, IEEE, Springer, Elsevier).
- If you make an internet query, prefer scholar google. Check always that the paper has been published!
- Write down the references – they can be hard to find afterwards! (especially store the bibtex files)

2.4 Reading

- You cannot read everything throughout!
  ⇒ Read only as much as is needed to
  - recognize that the article is useless
  - get the useful information
- Often an iterative process: important articles are read several times!
  - Title and abstract
  - Scan through introduction and conclusions/summary
  - Check references: new good references?
  - Important or useful sections and subsections (the organization is usually described in the introduction)
  - In the beginning, don't get stuck in details; don't check individual words or references; believe the arguments
  - If the article is important, then try to understand it properly, and check the referred sources
- Ask yourself:
  - What is the main idea?
  - What is the contribution (the new or interesting thing)?
  - What is important for you? Where it is presented?
- If you don't understand the article
  - Try to invent examples or simulate the solution yourself
  - Ask your fellows, supervisor, experts
  - Ask (yourself and others) specified questions: Where this equation comes from?, What is the relationship between these algorithms? Can you give an example for this definition?
  - Often understanding happens as a background process!

2.5 Citations and references

Citation is a key for finding source details in the reference list.

2.5.1 Citations in the text

- The citation is usually immediately after the cited theory, algorithm, author, etc.

  ”According to Dijkstra [Dij68], goto statement should be avoided...”

  ”*Bloom filters* [Ref03] solve this problem...”

- The citation is in the end, if the citation covers the whole sentence or a paragraph. (before full stop, if it refers only to the previous sentence, otherwise after the full stop)

  ”Goto statement should be avoided [Dij68].” Notice the difference: now you agree with Dijkstra!

- Sometimes there is no one ”original” source, but a new concept or theory has developed little by little. In this case, you can mention a couple of example references where the reader can find more information.

  ”*Context-aware computing* (see e.g. [DeA99,CaK00]) is a style of computing...”

Other examples

”Minsky and Papert [MiP69] showed that...”

”Version spaces were introduced by Mitchell [Mit77].”

”Nonparametric methods are described by Randles and Wolfe [RaW79].”

”The principles of CART were first described in Breiman et al. [BrF84].” or
”The principles of CART were first described in [BrF84].”

”Prolog was primarily used for writing compilers [VRo90] and parsing natural language [PeW80].”

”The general procedure for skolemization is given by Skolem [Sko28].”

”Other methods are summarized in e.g., [Bro92,Woo96].”

"The problem is NP-complete [Coo00].

2.5.2 Citation format: main styles

- Vancouver style: Arabic numbers (like [1] or (1)). Popular in physical sciences, quite often in cs (also theses), sometimes in biology and medical science.
  - compact, but not informative
  - many bibtex styles, prefer "plain"
- Harvard style: surname + year. E.g., (Hämäläinen, 2006) Common in humanistic sciences, sometimes in biology and medical science, occasionally in cs.
  - informative, but interrupt reading
  - many bibtex styles, try e.g., "apalike"
- Alpha-numeric labels: three letters from the authors' names + the last digits from the year. E.g., [Ham06]. Quite often in cs theses.
  - compromise between the two (at least years seen)
  - bibtex style "alpha"

Notes

- If you refer to a book, give the chapter or the page numbers! Otherwise, it is nearly impossible to find the place you are referring.
- If you summarize an entire chapter in a book, you can give the chapter number: [WMB94, chapter 2].
- If you refer to a certain page or pages: "[Bro92, pp. 3–7]"
- If you have several references, list them together: [Bro92, Woo96]
- The bibtex style affects also how references are presented (if bibtex is used). Styles plain and alpha produce similar entries.

2.5.3 Reference list

The last chapter in your thesis (or section in a paper) is called References. For each source, give

- The authors: surname and initials of the first names (in either order). If you have ≥ 3 authors, you can give only the first one, and replace the others by ”et al.” E.g., ”Mitchell T.M. et al.” or “T.M. Mitchell et al.”
- The title
- Publisher, place (publisher’s address) and year.
- Page numbers, if the source is a paper or a chapter in a collection written by several people.
- The title and the editors of the collection, if the paper has appeared in a collection (e.g., conference proceedings).
- The volume (always!) and the issue number after a comma or in parentheses, if the source is a journal paper.
- Series, if the book has appeared in some series. (E.g., Lecture Notes in Computer Science + number)

Examples

Benjamini Y. and Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B, 57(1):289–300, 1995. (a journal paper)

Edgington E.S. Randomization Tests. Marcel Dekker Inc., New York, third edition, 1995. (a book)

Lallich S., Teytaud O., and Prudhomme E. Association rule interestingness: Measure and statistical validation. In Guillet F. and Hamilton H.J., editors, Quality Measures in Data Mining, volume 43 of Studies in Computational Intelligence, pages 251–275. Springer, Berlin, Heidelberg, 2007. (an article in a collection)

- Precise formatting depends on the style (here bibtex style plain)

- More examples in the exercises!
- Notice that the journal and book titles are written with capital letters!

2.5.4 In latex:

- Latex creates the reference notations automatically!
- You can select the style by setting the style parameter for the bibliography environment
- Just invent a unique label string for each source, which you use in references by command `\cite`. E.g., `\cite`{whamalai}, or if you want to refer page 3, `\cite`[3]{whamalai}
- In the References, define what the label refers
- If you have a lot of sources, it is recommended to use `bibtex` to generate the reference list automatically (we will return to bibtex later in this course)

We will practise these in the computer class!

2.6 Quotations

Direct quotations are seldom used in cs texts. Always try to explain the ideas by your own words! (paraphrasing and summarizing)

If you use quotations, make clear who is responsible for what!

- If you express somebody else's ideas by your own words, then put the reference immediately after the idea.
- If you express somebody's ideas by her/his own words, then it is a quotation!
- If quotation marks "..." are missing, it is called plagiarism!
- How many consecutive words make up a quote? It depends on the context!
  - Variable suggestions: more than 3–4 consecutive words (excluding stop words like "a" "the", "and") or at least 5 words

  - Exception: conventional expressions (like definitions) that are widely used. E.g., in information retrieval, recall is usually defined as the “fraction of relevant documents that are retrieved” (no quotation marks needed).

- If the citation is translated, then mention also the translator in the reference.

- If you add or drop words, show it by [] or ....

- If you emphasize words, mention it.

- An example:
  Nykänen [Nyk03] would allow even longer quotations (translation and emphasis by the author): ”If you quote more than *seven* words ... from a text it [quotation] is called *literary theft.*”

- Be aware of **nearly quoting** (too little paraphrasing), where you make minor changes to the original text (divide and combine sentences or replace words by synonyms)! It is also considered a form of plagiarism or at least bad writing style (if appropriate references given).

2.7 Your own opinions?

By default: no opinions, everything must be based on facts!

If you have to express your own opinions, then

- *In principle*, everything without references is your own interpretation.

- However, make clear, what is borrowed and what are your own opinions!

- Often clearer to write a separate section called ”Discussion”.

Chapter 3

Use of tables, figures, examples, and similar elements

3.1 Figures and tables

3.1.1 General rules

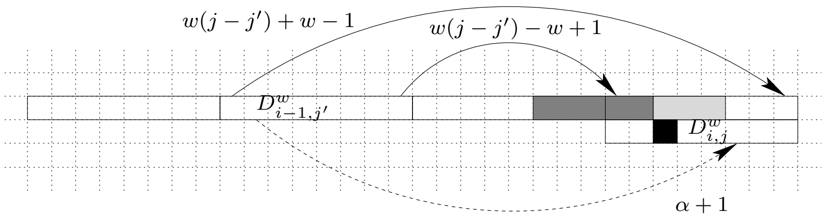
- Notice: all charts, graphs, pictures and drawings are called **figures**.
- Figures illustrate the models or the results, and tables give summaries. Usually there are never too many figures and tables, but remember two rules

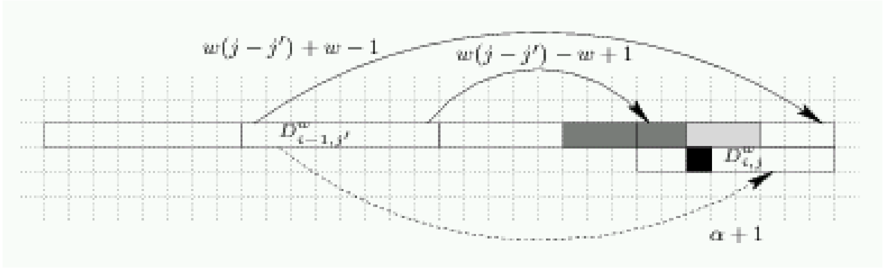
  1. All figures and tables must be referred in the text.
  2. There is no sense to express trivial things as a figure or a table (e.g. a table, which contains only two lines).

- If there is no need to refer to a figure/table in the text, the figure/table is probably not needed!
- Avoid repeating the same information in several places. An informative table or figure supplements rather than duplicates the text. Refer to all tables/figures, and tell the reader what to look for.
- Discuss only the most important items of the table in the text.
- A figure should be easy to understand. Do not present any unnecessary details.
- If two tables/figures should be compared, position them next to each other.

3.1.2 Vector graphics

Draw the figures by a tool which uses vector graphics, not raster graphic (bitmaps)! There is a big difference in quality:





(The bitmap file was also about 30 times larger!)

3.1.3 Captions

- Each table or figure should be understandable by its own. Give a brief but clear explanation or a title in the caption.

- Explain all special abbreviations, symbols, special use of bold font or parentheses, etc.

- Use the same style in all tables. If you use abbreviation *stdev* for standard deviation in one table, then do not use *sd* in another table.

- If you want to present a figure or table from some source and the authors have given permission, mention the source in the caption: “Table 5. An example ontology. Note. Reprinted from [ref].” (A page number is needed, if the table or figure is from a book.)

3.1.4 Tables and figures in latex

- Notice: Refer to tables and figures by numbers. Do not write ”the table below”. In latex this is implemented by using labels.

- The tables are encapsulated between `\begin{table}` and `\end{table}` commands. Similarly, the figures are encapsulated between `\begin{figure}` and `\end{figure}` commands.
- Inside table or figure environment you can write the caption for the figure/table and define a label (after the caption). E.g., define a figure label: `\label{fig-cmap}`.
- In the text, use `\ref{label}`, e.g., "An example concept map is shown in Figure `\ref{fig-cmap}`."

3.1.5 Expressions

When you refer to figures and tables you can use the following expressions:

- The results are summarized/reported in Table 1
- The results are represented in
- Figure 2 illustrates
- In the Figure we observe
- The model is given in Figure 7
- etc.

Notice the capital letters!

3.2 Lists

- Lists are not separate objects, and they are introduced in the text.
- Use list only when they are necessary! E.g.
  "The main criteria of X are (the following):"
  - Criterion 1
  - Criterion 2
  - ...

  Or "The method consists of five steps:" + a list
- If you list only a couple of items, you can usually write them without a list. Use lists when they clarify things!

3.3 Referring to chapters or sections

- The following chapters and sections can be referred easily in latex, even if you don't know their numbers yet.
- You just have to define a unique label name for the referred chapter.
- In the beginning of the referred chapter, you write

  ```
  \chapter{Conclusions}
  \label{concl}
  ```

  And when you want to refer it you write

  "The final conclusions are drawn in Chapter `\ref{concl}`."
- Notice that you can invent the labels yourself, if they are just unique and not reserved words in latex. E.g., above the label could be simply "c", but then there is a danger that you will give the same name for another object.

Useful expressions when you refer to chapters or sections

- The problem is discussed in Chapter X
- We will return to this topic in Section Y
- This problem is analyzed in ...
- etc.

Notice the capital letters!

3.4 Algorithms

- Give only the main algorithms in the text, in an appropriate abstraction level (pseudocode)
- Fix the pseudocode notation and use it systematically

- Simple methods can be described by a numerated list of steps
- Logical and set operations are often useful when you describe algorithms in an abstract level (for all $x_i \in X$, $T = T \cup \{p_i\}$, find such $S \subsetneq T$ that $q(S)$,...)
- If you writer longer algorithms, insert them into a figure or an environment of their own. Now they can be referred like tables and figures: "The *EM* algorithm for probabilistic clustering in given in Alg. 1"
- Later in this course, we will introduce a latex environment for writing algorithms.

3.5 Examples and definitions

3.5.1 Definition

A good definition

- explains the defined concept
- is not a circular argument (where x is defined by y and y by x)
- is not expressed by negative terms, if possible (Sometimes you cannot avoid this. E.g., statistical dependence is defined by statistical independence, because independence can be defined unambiguously.)
- doesn't contain unclear, vague, or descriptive language (i.e., is exact)
- defines only what is needed (i.e. the scope is restricted)

3.5.2 In latex

In latex, you can easily define environments for writing examples or definitions in a systematic way. The examples or definitions are numbered automatically and you can refer to them without knowing the actual number.
In the header you define `\newtheorem{example}{Example}`
In text you write
"The problem is demonstrated in the following example:"

```
\begin{example}
\label{example:bayes}
Write the example here.
\end{example}
```

When you want to refer to the example afterwards, you can write
"Let the problem be the same as in Example `\ref{example:bayes},...''`

3.5.3 Expressions for referring to a definition

- The definition of ... is the following:
- The definition of ... is as follows:
- Formally, we define...

3.6 Equations

3.6.1 Without equation numbers

If you don't need equation numbers (equations are not referred later), you can write the equations simply between double $ characters: $$<equation>$$. E.g., "The prior probability of X is updated by *Bayes rule*, given new evidence Y:

$$P(X|Y)\frac{P(X)P(Y|X)}{P(Y)}.$$

Remember the full stop in the end of the equation, if the sentence finishes! If the sentence continues, then you need comma:
"The dependency is described by equation

$$< equation >,$$

where a is sg. and b is sg."

3.6.2 With equations numbers

If you want to give an equation a number that can be referred later, you have to enclose it between commands `\begin{equation}` and `\end{equation}`.

$$P(X|Y)\frac{P(X)P(Y|X)}{P(Y)} \tag{3.1}$$

Now the equation is written in the math mode, and you don't need $ characters.
If you want to refer to some previous equation, you have to give it a unique label like for examples.

3.6.3 Text inside equations

Often you need also text inside an equation. To write text, you have to change to the text mode by `\textrm{text}` command.
For example, writing

```
$$A=\{(x,y)~|~x \in X, y \in Y \textrm{ and for all even } x, y
\textrm{ is odd}\}$$
```

produces the following:

$$A=\{(x,y)~|~x \in X, y \in Y \textrm{ and for all even } x, y \textrm{ is odd}\}$$

Chapter 4

Grammar with style notes

4.1 Verbs

Remember two important rules when you use verbs:

1. The number of subject determines the number of verb

2. Do not mix inconsistent tenses

4.1.1 Number and person

- When the subject is singular third person (she/he/it), the verb needs suffix -s (in the present, positive sentence). The auxiliary verbs have their own special forms (is, can, has, does).

- Be careful with special phrases:

  ”A number of new experiments were done” (plural)
  ”Plenty of time was spent...” (singular)
  ”A few data points belong to cluster X” (plural)

- Notice: when the subject is composed of a singular and a plural noun by ”or” or ”nor”, the verb agrees with the noun that is closer.

- If the number of the subject changes, retain the verb in each clause. E.g., ”The positions in a sequence were changed and the test rerun” → ”The positions in the sequence were changed, and the test was rerun.”

4.1.2 Tenses (temporal forms)

- Default: the present
- Past or present prefect (but not both) when you describe previous research (literature review)
- Past tense to describe the experiments and their results

- In scientific writing, the default is present (is). With present, you can combine perfect (has been) (and future, will be) if needed, but not the other tenses.
- Use past tense (was) only for good reasons. It expresses that something belongs to the past and has already finished. E.g., when you report your experiments.
- Past perfect (had been) is seldom needed. It is used, when you describe something in the past tense, and you refer to something which has happened before it. E.g.,
  ”We tested the system with data which had been collected in *Programming 1* course.”
- Notice: Use ”would” with care! It expresses a conditional action. E.g., ”it would appear” → ”it appears”.

4.1.3 Active or passive voice, which person?

Use of passive voice

- In active voice the actor is known, while in passive voice it is unknown.
- In the basic form of passive (”sg is done”), you can express also the actor (”sg is done by sy”). Expressing the actor is always more informative!
- It is often recommended to prefer active voice, but in scientific writing passive voice is sometimes convenient. It allows us to draw the reader's attention to the phenomenon or the event, instead of the actor. E.g., ”The probabilities are updated by Bayes' rule”, ”The values are recorded every minute”.

- Often the purpose determines the voice. Usually we want to begin with a familiar word and put the new information in the end. E.g., before an equation or a definition, we can say ”The model is defined as follows.”
- However, do not overuse passive, and do not chain passive expressions. As a rule of thumb, use only one passive per sentence
- Read Section 11 in Strunk: ”Elements of style”! (link in the course page)

”It is” and ”There is/are”

- A formal subject ”it” is sometimes used in passive expressions: ”It is often recommended [reference] that...”
- Typical verbs in this expression are: say, suppose, consider, expect.
- ”There is/there are” is a similar expression, but now we don’t need the passive. This expression is used when the real subject (what is somewhere) comes later and we haven’t mentioned it before.

  E.g., ”There was only one outlier in data set D.” vs. ”The outlier was in data set D.”
- The verb is nearly always ”be” (sometimes ”exist” or something else)
- Notice that the verb follows the real subject’s number.

  E.g., ”There were a lot of outliers in data set D.”
- ”There is” expression is seldom needed in scientific writing, and often you can circumvent it:

  ”Data set D contained a lot of outliers.”

Other passive expressions

- ”We” can be used as passive. E.g., ”In Chapter X, we define the basic concepts.” However, it is better to say ”The basic concepts are defined in Chapter X.”
- ”You” is sometimes used as passive, especially in manuals. Don’t use it in scientific text!
- ”People” when you refer generally to people. Quite a vague expression, not recommendable!

Person?

- Basic rule: avoid the first person (no opinions, but facts). However, sometimes we can use ”we” as a passive expression. Problem: whom you are referring to, if you write alone?

- Referring to yourself: you can talk about ”the author”. E.g., ”All programs have been implemented by the author.” (Notice that it is not guaranteed that your supervisor likes this! Some supervisors prefer ”I”.)

- Gender-neutral language: when you refer to an unknown user, student, etc. try to use gender-neutral language.

  - The most common way is to say ”she/he” or ”he or she”. Some authors are careful about the order of her/him, as well! E.g., you can use every second time ”she or he” and every second time ”he or she”. Remember to put the other pronouns in the same order (”She/he tries her/his best”)
  - ”One” is neutral, but sounds often awkward. ”The learner can define one’s own learning goals”
  - Sometimes you can avoid the problem by using plural.

4.1.4 Other notes

- Do not use short forms ”isn’t”, “can’t”, “doesn’t”, but ”is it”, “cannot”, “does not”.

- ”to be” + -ing form of the verb when something is currently happening or takes some time. E.g., ”Thread 2 can be started in the same time when thread 1 is still running”

- Some verbs require that the following verb is in the -ing form:

  {enjoy, avoid, succeed in, finish, keep, mind, practice, risk} + verb + ing

  E.g., ”Students enjoyed learning new things”

- Special phrases: ”be used to”, ”be (un)likely to”

4.1.5 Noun syndrome

"Noun syndrome" = use of common verbs {be, do, have, make, ...} + a noun

E.g., "We can get better understanding...","Different people have different responses to the methods"
⇒ Prefer illustrative verbs!

Task: How would you correct the previous sentences?

Useful verbs:

represent, analyze, compare, demonstrate, illustrate, summarize, conclude, list, define, report, model, implement, design, consider, involve, simplify, generalize, perform, be based on sg., take into account sg., depend on sg, increase, decrease, evaluate, predict, assign, require, satisfy, ...

Task: What is the difference between the following concepts? Give examples when they are used!

evaluate – assess
compute – calculate
derive – infer
approximate – estimate
discover – find

4.1.6 Often needed irregular verbs

The following list contains irregular verbs which are sometimes needed in computer science expressions, excluding the most common ones (which all of you know!):

choose – chose – chosen
find – found – found
hide – hid – hidden
hold – held – held
lead – led – led
lose – lost – lost
rise – rose – risen
seek – sought – sought
show – showed – shown

spin – spun – spun
split – split – split
spread – spread – spread
stick – stuck – stuck

In addition, the last consonant can be doubled before -ed, if

- the spell is short and stressed: planned, dropped,
- the consonant is 'l': travelled, modelled, biassed

Notice: American English is not so strict, and `ispell` can complain about correct spelling!

Exercise

Read the given text part and underline useful expressions. Search especially for the following kinds of expressions:

- Useful verbs and their prepositions in computer science texts.
- How to list advantages or disadvantages without repetition (usually in the beginning of sentences).
- How to compare approaches?
- Any other useful expressions!

The same text is given to two people. Thus, you can discuss with your pair, if you don't understand something. However, it is not important if you don't understand all words.

4.2 Nouns

Nouns are usually easy. If you don't know a word, you can check it from a dictionary – just be careful that the meaning is what you want.

Often a better way is to move a term from your passive vocabulary to the active one – then you know also the context!

4.2.1 Plural forms

Irregular plural forms

half	–	halves
life	–	lives
axis	–	axes
matrix	–	matrices
child	–	children
person	–	people
automaton	–	automata
vertex	–	vertices
analysis	–	analyses
thesis	–	theses
basis	–	bases
series	–	series
medium	–	media
criterion	–	criteria
phenomenon	–	phenomena

Data is originally the plural form of **datum**, but nowadays it is frequently used as a singular word. The same holds for **hypermedia**. ”The data is biassed”, ”Hypermedia offers a new way to implement learning environments”

Notice also:

- If the suffix is {-s, -ss, -sh, -ch, -x, -z} in singular → -es in plural, e.g., research – researches, approach – approaches, quiz – quizzes
- The same happens with most words which have suffix -o, unless the word is abbreviated or of foreign origin. E.g.,
  cargo – cargoes, but photo – photos, dynamo – dynamos
- After **consonant** -y changes to -ies in plural. E.g., floppy – floppies.

Singular words which look like plural forms

The names of disciplines: mathematics, statistics, physics.
”Statistics is the predecessor of data mining.”

news is also singular!
”Good news is that the algorithm works in $O(n)$ time”

4.2.2 Countable and uncountable nouns

Countable nouns (C-words) refer to things which can be counted, while things referred by **uncountable** nouns cannot be counted.

Uncountable nouns (U-words) can be divided into three groups:

1. Words expressing material: water, air, wood, ...

2. Abstract words: life, time, work, strength, ...

3. Exceptional: advice, information, news, equipment, money

Notes

- Uncountable words are missing the plural form!

- Notice that sometimes a noun can be either a countable or an uncountable word depending on the meaning. E.g., science (when you refer generally to natural sciences) – a science (when you refer to a discipline).

- The words in group 3 are grammatically *singular* but they have also plural meaning. If you want to refer to a singular piece you have to express it in another way: "a piece of information", "an item of news", "a bit of advice".

"This information **is** important"! "All advice **is** good!

4.2.3 Extra: differences between British and American English

Some nouns have different spelling in British and American English. Try to use systematically either British or American forms!

We will return to other differences in the end of the course.

British	**American**
colour	color
neighbour	neighbor
behaviour	behavior
favour	favor
honour	honor
metre (unit)	meter
meter (device)	meter
centre/center	center
analogue	analog
dialogue	dialog
encyclopaedia	encyclopedia
arguement	argument
judgement	judgment
programme (academic, tv)	program
program (computer)	program
defence	defense
practice (noun)[1]	practise
maths	math
speciality	specialty

4.3 Compound words

Spelling practices vary, and it is hard to give exact rules when compound words should be written together, with a hyphen −, or separately (i.e., if they are *closed, hyphenated, or open compounds*).

- If the words have become one concept, they are usually written together, e.g., ”software”, ”keyboard”, ”database”

- If the independent meaning of words is emphasized, they are hyphened, e.g., ”non-smoker” (cs example?)

- Hyphen is often used when the concept consists of more than two words: ”depth-first search”, ”between-cluster variation”, ”feed-forward neural network”, ”first-order logic”

- Multiple word adjectives are usually hyphened, e.g., ”data-driven”, ”model-based”, ”class-conditional”

- If the first part is a symbol or an abbreviation, the word is hyphened, e.g., ”NP-complete”, ”k-nearest neighbour method”, ”3-dimensional”

- Some common phrases have become compound words in American English, but remained as phrases in British English. E.g., in American English you can spell ”trademark”, but in British English ”trade mark” or ”trade-mark”. (cs example?)

- Notice that many words which are compound in your mother tongue are written separately in English: ”data set”, ”density function”, ”wave length” (this is typical especially for long words)

Problem: how should we spell the following computer science terms?
overfitting, nondeterministic, time demanding, drop-out, EM-algorithm

4.4 Articles

4.4.1 Position

Basic rule: before the noun phrase (a noun + preceding attributes)
Exceptions:

1. {what, such, quite, rather, half} + a/an + noun phrase

   ”Half an hour”, ”quite a fast system”

   (In American English the rules are not so strict concerning quite, rather, and half.)

2. {too, as, so, how, however} + adj. + a/an + noun

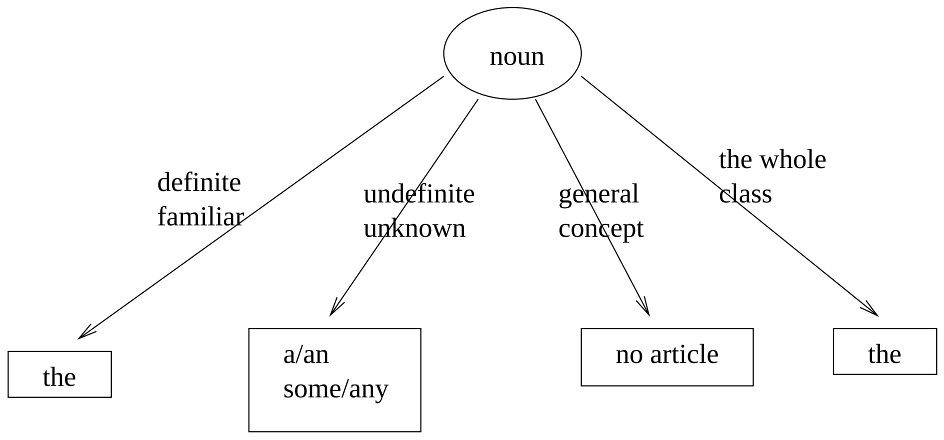
   ”Too great a distance”, ”so long a time”, ”as big a difference”

3. {all, both, double, twice, half} + the + noun

   ”All the methods”, ”twice the time”, ”double the amount”

4.4.2 Use of articles

Basic rules:



Definite and indefinite concepts

A concept is indefinite, when you mention it the first time, and it is not clear from the context. Usually this kind of expressions are describing: "There was a time delay between processes A and B."
It is definite, when

- you mention it again ("The time delay was about 10 ms")
- the context defines what you mean ("The left-most bit is always 1.", "The results of process A were correct.")
- the concept is familiar to everybody (the Earth, the sun, the moon)

Usually this kind of expressions are defining: "The delay between two processes P_1 and P_2 is $t_{end}(P_1) - t_{start}(P_2)$."

When you refer to an indefinite concept

a singular C-word → a/an
a plural C-word + positive clause → some
a plural C-word + negative or interrogative clause → any
a U-word + pos. clause → some
a U-word + neg. or interr. clause → any

When you refer to something generally

a plural C-word or a U-word → no article

"Students need time to process new information"

When you refer to the whole class

Definite article + a C-word:

”The computer cannot solve all problems”
(which means that none of the computers can solve all problems, the property concerns the class of all computers)

Exceptional expressions

Sometimes you can use a/an article with an abstract word:

- when the word is proceeded by a describing relative clause ”There is a danger that the model overfits”
- expressions ”a /short/long time”, ”a while”

The article with ordinal numbers and some adjectives

Definite article ”the” is used

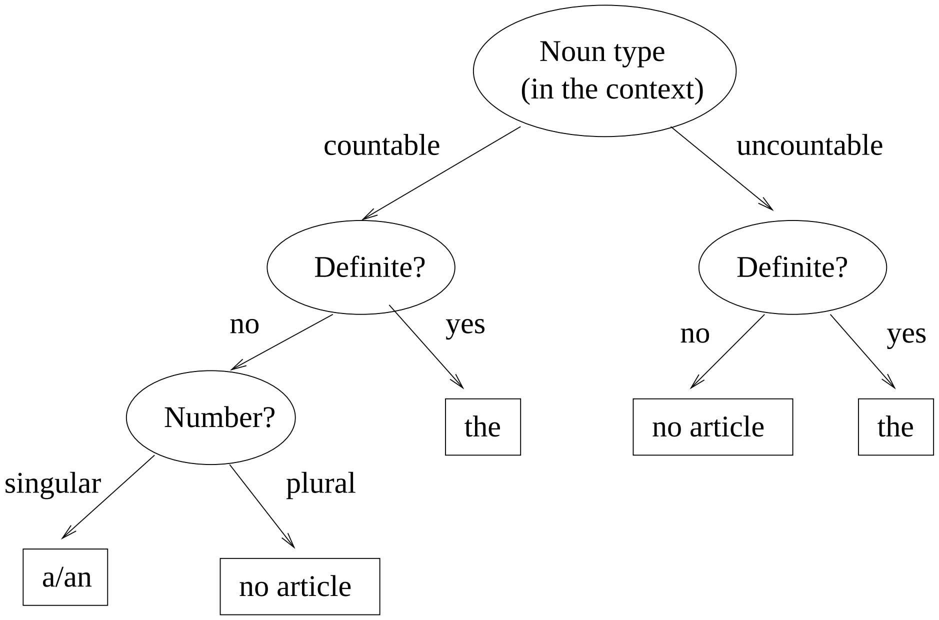
- when the noun is preceded by an ordinal number (”The first attribute describes...”)
- when the noun is preceded by an adjective expressing order (”the next attribute”, ”in the following chapter”)
- with adjectives “same”, “only”, “right”, “wrong” (”The results were the same”, ”The only model which has this property is X”)

Notice: ”the” is not used with ordinal numbers or adjective ”last”, when you refer to the performance in a competition (”Program X came first and program Y was last when the programs were compared by the Z test.)

Task: Try to draw a complete decision tree for choosing articles

4.4.3 Hints

A better decision tree for articles:



When a noun can be used as a countable or an uncountable concept

The use of articles depends on the concept which is meant in the **current context**. For example, word *memory* can have at least three meanings:

1. The store of things learnt or the power or process of recalling (in our brains) → generally uncountable. "Memory can be divided into two classes: short-term memory and long-term memory. The short-term memory..." However, you can say: "I have a good memory".

2. The object of recall → countable. "My earliest memories"

3. The capacity of a computer to store information → uncountable. In the cs context, you can suppose it as a known concept and use article **the** (always?). "The data is loaded into the main memory"

Time is another word which can be used in different ways. It can mean a limited period or interval, an indefinite period or duration, or it can express an occasion of repeated actions. In addition, it occurs in several phrases. By default, time is uncountable (either no article or article "the").

1. Without any article:
   "Time will show..."
   "It is time to do sg."
   "It takes time..."
   "on time" (or "in time")

2. Article "the"
   "all the time"
   "at the same time"

3. Article "a":
   "It is a long time..."
   "one at a time" (i.e., one by one)

4. Plural:
   "many times"
   "modern times"

Hint: could you use "any" or "some"?

Try if you could use words **any** or **some** before the noun. If you can, it is indefinite. This means that you cannot use article "the".
"The grammar is not strict in (any) spoken language"
"The disk contains (some) space for back-up files"
"There is some reason for this behaviour" → "There is a reason for this behaviour".

Hint: are you referring to sg particular?

If you have need to say "This particular x", say "The x", where x is a noun. "This particular" hints that you have already talked about x and it is known (definite).
Don't use pronouns, if you mean article "the"! "This x" can often be replaced by "the x" (where x is a noun).

Hint: could you use $\exists$ or $\forall$?

Imagine the concept C as a set (universe) of all its instances. E.g., concept "computer" is a set of all possible computers.
If you want to express $\exists x \in C$ such that $P(x)$ (there is some x in C for which holds property P), use article **a/an**. "A computer could solve this problem faster." (maybe not all of them, but some computers can)

If you want to express $\forall x \in CP(x)$ (for all x in C property P holds; i.e., it holds for the whole set C), use article **the**. Now you refer to the whole class of xs in C, which is definite. "The computer can solve only mechanical problems." (all computers can do this)
Notice that this technique suits only for countable concepts!

Articles before variable names?

In cs, we often use the names of variables, data sets, models, etc.

- When you use the name without any modifying word → no article
  "X is independent from Y", "S contains no outliers"

- When you use a modifying word like "set", vector", "model" etc. before the name →
  by default, no article

- Problem: in literature, article "the" is sometimes used in such cases??

Exercises

Task 1: Add the correct articles to the following sentences or mark the absence of articles by $-$!

1. _____true positive rate was higher in _____method X than _____method Y.

2. _____method X had _____higher true positive rate than _____method Y.

3. _____memory means _____power or _____process of recalling.

4. X is _____algorithm which solves _____Travelling Salesman problem. _____algorithm X is _____fastest among all _____known TSP algorithms.

5. _____data set X follows _____Normal distribution with _____parameters μ and σ^2. _____parameter μ is _____mean of _____set X and _____parameter σ^2 is _____variance of _____X.

6. _____problem X belongs to _____class P, if it has _____polynomial time algorithm Y. _____time complexity of _____algorithm Y is $O(p(n))$ where n is _____size of input and p is _____polynomial function.

7. In _____next section we introduce _____theory of _____Bloom filters.

8. To assess _____students' program codes, we construct _____bug library. _____bug library contains all _____errors which have occurred in _____students' programs.

9. _____infinite time Turing machines extend _____idea of _____traditional Turing machines.

10. In _____pattern extraction we produce _____set of _____new attributes from _____original ones. _____goal is to find such _____set of attributes which describes _____data _____best. _____goodness of representation depends on _____modelling purpose, and in _____practice we have to define _____appropriate goodness measure.

11. In _____clustering analysis we divide _____data points into _____clusters such that all _____data points in one cluster are similar to each other but different from _____data points in _____other clusters.

12. _____episode is _____set of _____events which occur together. If _____order of _____events is fixed, _____episode is called serial.

13. There is always _____danger that _____model overfits. _____danger that _____model overfits is unavoidable.

14. _____main parts of _____computer are _____central unit, _____hard disk, and _____i/o devices. _____central unit is responsible for all _____computation.

Task 2: Are the following words countable or uncountable? Which articles can you use with them? Give example sentences!

- space
- requirement
- model
- program
- computation
- power
- capacity
- data
- information
- knowledge
- recognition
- software
- hardware
- code
- value
- property
- strength

- weakness
- use
- usability

4.5 Pronouns

Two important rules when you use pronouns:

1. When a pronoun refers to a noun in the preceding sentence, make sure that the **referred is obvious**!
2. Each pronoun should agree with the referent in number and gender.

4.5.1 Unclear references

- The simple pronouns – it, they, this, that, these, those – do often create ambiguities.
- Goal: the reader should not have to scan the previous sentence to understand what you mean.
- Recommendation: **Avoid them, when possible! If you use them, always check twice that the meaning is not ambiguous!**
- Never use "those" – it is usually a sign that the sentence is foggy.
  "There was no difference in the accuracy of models between those which belonged to group A and those which belonged to group B. → "The models in groups A and B were equally accurate."
- Do not use "it" to begin a sentence, if it is not absolutely clear, what it refers! (Exception: expressions like "It is difficult to estimate..." require "it" as a formal subject.)
- Hint: often you can replace "this/these" + noun by "the" + noun!
  "This experiment demonstrated..." → "The experiment demonstrated..."

4.5.2 Pronouns which require singular verb form

{everybody, anybody, nobody, everyone, anyone, no one} → verb is singular

4.5.3 Every vs. all

every	**all**
+ singular noun	+ singular or plural noun
when you talk generally	when you mean sg certain

4.5.4 Many vs. several

several < many
several ≈ some

4.5.5 Phrases

one – the other (singular)
some – the others (plural)

each other, e.g. "X and Y affect each other"

This kind of + singular noun, e.g., "This kind of system..."
Problem: What the plural means? "This kind of systems" (one kind with many systems belonging to it) vs. "systems of this kind..." (equivalent?) vs. “these kinds of systems” (many kinds)?

on one’s own, e.g., "The students solved the task on their own".

"**All but one** point belong to cluster 1"
"**First of all**, we have to initialize the parameters"
"**On the one hand**, the system is stable, **on the other hand**, it has poor accuracy"
"The initialization phase is time demanding. **Otherwise** the program is very efficient."

4.5.6 Relative pronouns

Relative pronouns (who, which, that) are used in **relative clauses**. To understand their use we have to study also relative clauses.
→ Section Relative clauses.

4.5.7 Extra material: Tricks for gender-neutral language

Trick	Incorrect	Correct
Use plural	The student returned his solution.	The students returned their solutions
Article "the"		The student returned the solution.
Drop the pronoun	The user himself defines the preferences.	The user defines the preferences.
Special expressions	man, mankind man-machine interface Researchers' wives mothering chairman Mrs. Smith housewife	people, human beings, humankind user-system interface, human-computer interface Researchers' spouses parenting, nurturing chairperson, chair, head Jane Smith homemaker

4.6 Adjectives

These seem to be well mastered, just two notes:

1. **Avoid vague adjectives!**

2. How to derive and use comparative and superlative forms?

4.6.1 Vague adjectives

- Do not use vague adjectives. Especially the adjectives which describe amounts (large, small, huge) are very context-sensitive!

- E.g., for statisticians, a data set of 500 rows is quite large, while for a data miner it is extremely small → numbers are more exact!

- The expressions become even vaguer, when you add modifiers "quite", "rather", "very", etc. Skip them always when possible!

4.6.2 Comparative and superlative

Basic rule: use -er/-est for short adjectives, and more/most for longer ones.

Adjective type	**Comparative**	**Superlative**	**Examples**
1-syllable adjectives	-er	-est	strong, stronger, the strongest
2-syllables adjective with suffix **-y, -ow, -er**	-er	est	narrow, narrower, the narrowest
2-syllables adjective with suffix **consonant + le**	-er	est	noble, nobler, noblest
all other adjectives	more + adj.	most + adj.	efficient, more efficient, the most efficient
participles verb+{**-ed, -ing**} when used as adjectives	more	most	interesting, more interesting, the most interesting
irregular adjectives			good, better, the best bad, worse, the worst

Notice the spelling:

- the consonant is doubled in a short stressed syllable: big, bigger, the biggest

- **-y** becomes **-ie**: easy, easier, easiest

4.6.3 When you compare things

When you use the comparative, make clear what you are referring!

"Problem X is easier to solve" (than what?)

Basic structure:

X is **as** efficient **as** Y (X and Y are equally efficient)
X is more efficient **than** Y

Exceptional expressions:

X is **different from** Y
X is **similar to** Y
X is **the same as** Y
X is **inferior/superior to** Y
X is **equal to** Y (Notice: use "X equals Y" only in math, for $X = Y$)

4.7 Other word groups

Verbs, nouns, pronouns, numerals, and adjectives compose the skeleton of sentences. The additional stuff consists of

- adverbs,
- prepositions, and
- conjunctions.

Adverbs modify verbs, adjectives, or other adverbs, while conjunctions join words, clauses or sentences together. Some words can be used either as adverbs or as conjunctions. Prepositions are always connected to other words (nouns, pronouns, or verbs in -ing form). Prepositional phrases ("in the beginning", "through a gateway") are used in the same way as adverbs.

4.8 Adverbs

Adverbs answer questions When? Where? What? Why? How?
They express

- time (immediately, now, soon, later, next)
- place (here, there, everywhere)
- manner (easily, temporarily, well, poorly)
- degree (very, quite, ...) → Avoid in scientific texts!
- frequency (often, seldom, usually, sometimes)
- speaker's attitude "Fortunately, the data set is small, and function f can be computed in real time." → use sparsely!

Notes:

- **Recommendation**: Use expressive verbs and nouns which express the most of message, and as few adverbs/prepositional phrases as possible!
- Use introductory adverbs like "fortunately, similarly, conversely, certainly" carefully, as a synonym to expressions "it is fortunate" or "in a similar manner". Drop them if they are not needed.

- Notice that ”importantly” and ”interestingly” are not proper adverbs. E.g.,
  ”More importantly, the accuracy can actually increase when the complexity is reduced”
  → ”More important, the accuracy can actually increase when the complexity is reduced.”

  ”Interestingly, we found that...”
  → ”An interesting finding was that...”

4.8.1 The position of adverbs in a sentence

The adverb can be

1. in the beginning, when you express time or attitude. E.g., ”Evidently, the students’ learning outcomes depend on their effort”, ”Later, we realized that...”
2. in the end, when you express way, time or place. E.g., “This problem occurs frequently in sparse data.”
3. in the middle, when you express frequency or attitude. Notice that *already* behaves in the same way. E.g., ”In knowledge discovery, we assume that the features have been already extracted”

An adverb should clearly refer to the word it modifies!

4.8.2 Special cases

still and **yet**

- Still (mostly in positive sentences): before the main verb, but after be-verb. ”These enlargements are still unimplemented”
- Yet (mostly in negative or interrogative sentences): in the end. ”These enlargements have not been implemented yet.
- If still or yet is used in the beginning, it means ”however”.

so and **such**

- So: before adjectives or adverbs which are **not** succeeded by nouns. E.g., ”The proof is not so difficult”

- Such: when an adjective is succeeded by a noun. E.g.,”Such a brute-force approach is infeasible”

- Notice the article ”a/an”, if the noun is countable: ”such a system”, ”such an algorithm”

4.8.3 Extra: How to derive adverbs from adjectives?

Basic rule

Basic rule: by **-ly** suffix:adverb = adj. + ”ly”

E.g. poor – poorly

Exceptions

Adjective suffix	Adverb	Examples
-y	-ily	easy – easily
-e	-ly	whole – wholly, true – truly
-ic	-ally	automatic – automatically, systematic – systematically Exception: public – publicly
-able/-ible	-l disappears	sensible –sensibly
-ly	in a <adj.> way	in a friendly way

If you are not sure how to derive an adverb, check it from a dictionary!

Adverb = adjective

fast, hard, late, straight, low, wrong, right, long

Notice the difference in meaning (both can be used as adverbs):

deep	vs.	deeply
hard	vs.	**hardly**
high	vs.	highly
most	vs.	**mostly**

Task: Draw a decision tree for deriving adverbs from adjectives!

4.8.4 Comparing adverbs

Adverb type	Comparative, superlative	Examples
-ly suffix	more <adv.>, most <adv>	more carefully, most carefully
like adjective	-er, -est	faster, fastest

Exceptions: well, badly, much, little, far

"This is a less desirable solution", "The X algorithm performs worse/better than the Y algorithm"

Notice:

- far, farther, farthest, when you express distance, E.g.,
  "Point x lies farthest from the centre."
- far, further, furthest, when you express distance, time, or in an abstract context. E.g.,
  "In Chapter X, we will analyze this problem further" or "This problem is further analyzed in Chapter X"

4.9 Parallel structures

Conjunctions and some special phrases are used to combine words, word groups (phrases), clauses or sentences. Here we concentrate on combining parallel elements. A different structure is needed for combining a main clause and a subordinate clause. → Section Sentences.

Parallel structures are used to present parallel ideas.

Parallel structure = words, phrases, clauses or sentences combined by commas and/or conjunctions. Here we call the combined items as **parallel items**.

- Parallel items are combined by parallel conjunctions (and, or, but, ...).
- Notice that lists are also parallel structures!

- Often the parallel structure lists alternatives or makes some kind of comparison: the items belong to the same or similar classes or to two opposite classes.
- E.g.,
  "Method X has several advantages: it is easy to implement, it works in polynomial time, and it can use both numeric and categorical data."

  contains two parallel structures: three advantages ("it is, it works, it can") in a list and "both numeric and categorical data"

4.9.1 Basic rules

The parallel structure should be consistent in two ways

- **Semantically**: the concepts referred by parallel items should be comparable, i.e., the comparison should make sense.
- **Syntactically**: the items should have similar grammatic structure. All of them should be either nouns, noun phrases, verb phrases, or clauses. In addition, they should be in the same form, e.g., you cannot combine "to" + verb and a verb without "to".
  "The problem is both hard to define and solve"
  → "The problem is both hard to define and **to** solve"

4.9.2 Parallel items combined by conjunctions and, or, but

The most common form of parallel structures!

"The method has low space **but** high time requirement"
→ "The method has low space requirement **but** high time requirement.

"The students were told to make themselves comfortable, to read the instructions, **and** that they should ask about anything they did not understand"
→ "The students were told to make themselves comfortable, to read the instructions, **and** to ask about anything they did not understand"

"The results show that X did not affect the error rate **and** the model overfitted the data"
→ "The results show that X did not affect the error rate **and** that the model overfitted the data"

4.9.3 Lists

Notice that elements in a list should be in a parallel form!

Example 1

"Boud [Bou89] has listed general characteristics which are typical for problem-based courses:

- Acknowledgement of learners' experience.
- Emphasis on students taking responsibility of their own learning.
- Crossing of boundaries between disciplines.
- Focus on the processes of knowledge acquisition rather than the products of such processes.
- Change in staff role from instructor to facilitator.
- Students' self- and peer assessment of learning.
- Focus on communication and interpersonal skills."

Example 2

"The clustering methods can be divided into three categories:

1. *Hierarchical methods* construct a hierarchy of (typically) nested clusters.
2. *Partitioning methods* try to find optimal partitioning into a specified number of clusters.
3. *Probabilistic model-based clustering* tries to find the underlying probabilistic model which has produced the data."

Example 3

"The whole procedure is following:

1. Determine the number of clusters k.
2. Choose parametric models (density functions f_j) for each of the clusters.

3. Determine the component probabilities π_k and parameters θ_k from data.
4. Assign each point to the most probable cluster."

4.9.4 Example 4

"According to O'Shea [OSh00], an intelligent tutoring system should be

- robust,
- helpful,
- simple,
- transparent,
- flexible,
- ...
- sensitive, and
- powerful."

Notice! The previous kind of list should be avoided, because it can be written as a normal sentence. A list was used above, because 13 items were listed (and they were analyzed later). If you list only a couple of items (e.g., less than 5), write them as a normal sentence!

4.9.5 Parallel items combined by conjunction pairs

Sometimes the parallel expression consists of two conjunctions like

- **between...and**,
- **both...and**,
- **either...or**,
- **neither...nor**, and
- **not only...but**.

The first conjunction should be immediately before the first part of the parallelism.

between – and

"between 20-22 years of age" → "between 20 and 22 years of age"

"We recorded the difference **between** the students who completed the first task **and** the second task"
→ "We recorded the difference **between** the students who completed the first task **and** the students who completed the second task."

both – and

"The task is **both** easy to solve **and** efficient." (Doesn't make any sense!)
→ The task is **both** easy to solve **and** can be solved efficiently."

Or another structure:
"The task is easy **and** the solution is efficient."

either – or

"The students **either** gave the worst answer **or** the best answer."
→ "The students **either** gave the worst answer **or** gave the best answer." or
"The students gave **either** the worst answer **or** the best answer."

neither – nor

In negative clauses → less often needed in sciwri! (Say things in a positive way, when possible.)

"X solves the problems of traditional clustering algorithms. **Neither** outliers **nor** missing values affect the clustering quality."

(Grammatically correct, but better to say: "X solves the problems of traditional clustering algorithms. It is not sensitive to outliers or missing values.")

not only – but (also)

"The task is **not only** easy to solve **but also** efficient"
→ "The task is **not only** easy to solve **but** the solution is **also** efficient" or
"The task is **not only** easy to solve **but** it can **also** be solved efficiently"

Once again: say in a positive way, when possible – clearer!

On the one hand – on the other hand

- A special expression: can combine either clauses or parallel sentences!
- An affective way to describe opposite points, like advantages and disadvantages!

"**On the one hand**, a complex model can describe the data well, but **on the other hand**, it overfits easily."

"There is always a wrestling between the descriptive power and the generalization ability. **On the one hand**, too complex a model describes the data well, but it does not generalize to any new data. **On the other hand**, too simple a model generalizes well, but it does not describe the essential features in the data."

4.9.6 The comparative – the comparative

The comparative forms of adjectives can used in a parallel way in the following structure:

> the + comparative + x + comma + the + comparative + y, where x and y complete the clauses.

"**The more complex** the model is, **the better** it describes the training data."

If x and y are missing, then no comma:
"The sooner the better."

Notice: Use sparsely!

4.9.7 Parallel sentences

Numerating properties or ideas is an efficient way to create logical structures into paragraphs. The sentences in the list begin by ordinal numbers "First, Second, Third". (Notice: you can say "Firstly", but there is no need for that!)

"*X* model has three important properties: First, the model structure is easy to understand. This is a critical feature in adaptive learning environments, as we have noted before. Second, the model can be learnt efficiently from data. There are feasible algorithms for both numeric and categorical data. Third, the model tolerates noise and missing values."

4.10 Prepositions

- Be careful with prepositions. A wrong preposition can give a totally different meaning!
- Hint: When you use a preposition, visualize the direction it is signalling and ask yourself if it is appropriate.
- If you are unsure about the use of a preposition, ask yourself what a cat would do! (Fedor's sciwri book)

  Cats sit **on** mats, go **into** rooms, are part **of** the family, roam **among** the flowers.

4.10.1 Expressing location

- Usually **in**, e.g., "in set X"
- If an exact location, then **at**, e.g., "at point (x, y)"
- If the location can be imagined as a line or a surface, then **on**, e.g., "on the x-axis", "on a time line"

Notice: "**on** page 3", "**on** line 5"

"A file is loaded **from** the hard disk **into** main memory."
"results **from** the survey suggest..."
over – under/beneath
above – below

"X's points were **below** the average points"
"The task is to optimize f **under** the given constraints"

4.10.2 Expressing time

- Exact time: **at**, e.g., "at the moment", "at four o'clock", "at the same time"
- Longer period of time: **in**, e.g., "in the 1970's", "in the future", "in five minutes", "events occur close in time"

Notice: "In the beginning/end" vs. "At the beginning/end of sg"

4.10.3 Expressing the target or the receiver: to or for?

Basic rules:

- When direct receiver, then **to**
  ”The values are assigned to variables”

- When the final receiver (for whom sg is meant) then **for**
  ”I gave the book for Belinda to Tersia”
  ”The messages for nodes F and G are transferred to node D for rerouting”

- When sg is good or bad for sg, then **for**
  ”Problem-based learning is good for students”

Some verbs require either for or to:

1. If the verb is {bring, give, take, show, offer} → **to**

2. If the verb is {be, get, keep, make} → **for**

Sometimes the preposition can be missing, depending on the word order:

i) verb + receiver + object
ii) verb + object + to/for + receiver
iii) verb + to/for + receiver (no object)

- If the verb is **tell**, then always case i.

- If either object or receiver is pronoun, then the pronoun becomes before the noun (case i or ii)

- If both are pronouns, then the object becomes first (case ii)

- If the verb is {belong, describe, explain, introduce, reply, say, speak, suggest} → always **to** (cases i–iii)

Task: Draw a decision tree for deciding when to use ”to” or ”for”!

4.10.4 Special phrases

Some prepositional phrases just have to be remembered! (or checked)

constraint **on** sg (e.g., constraints on the order)
independent **from** sg but dependent **on** sg
different **from** sg but similar **to** sg
difference **between** sg and sg
prefer sg **to** sg
impact of sg **on** sg
influence **on** sg
effect **on** sg (but to affect sg)
a discussion **about/on** sg (but to discuss sg)
research **on** sg but a study **of** sg
reason **for** sg
opportunity **of/for** sg
in spite of sg (but despite sg)
take into account
in relation to sg
a proportion **of** sg. ("a large proportion of data")
in proportion **to** sg, proportional **to** sg ("The time complexity of f proportional to n is...")
the ratio **of** a **to** $b = a/b$
$x\%$ **of** y
by default

Problem: compare with or to?
Depends on the meaning!

From Kdict:
"Usage: Things are **compared with** each other in order to learn their relative value or excellence. Thus we compare Cicero with Demosthenes, for the sake of deciding which was the greater orator. One thing is **compared to** another because of a real or fanciful likeness or similarity which exists between them. Thus it has been common to compare the eloquence of Demosthenes to a thunderbolt, on account of its force, and the eloquence of Cicero to a conflagration, on account of its splendor. Burke compares the parks of London to the lungs of the human body."

4.11 Sentences

4.11.1 Terminology

- A **sentence** consist of one or more clauses
- A **clause** contains always a subject and a predicate, and usually an object
  - An **independent clause** (main clause) can make a sentence alone.
  - A **dependent clause** (subordinate clause) needs an independent clause for support.

4.11.2 Sentence types

The sentence type depends on the type of its main clause. The main types are following:

1. Statement (ends by a full stop: "*x* is *y*.")
2. Question (ends by a question-mark: Is *x y*?")
3. Imperative (order, request, command; may end with an exclamation mark: "Be *x y*!)

In scientific writing the default type is the statement. Direct questions and orders are seldom used.
Questions suit best to the introduction where you state your main research questions clearly and concretely, e.g.,
"The main research questions are the following:

1. What is the relationship between X and Y?
2. When X can be applied?
3. Can we apply X in Z?
4. How X can be extended?"

Imperatives can be useful in pseudo code, when you describe some method. E.g., "Divide data into two parts." "Search such c_i that $d(x, c_i)$ is minimal."

Dependent clauses can be divided into the following types:

1. Clauses beginning by sub-ordinating conjunctions (when, if, because, while, ...)
2. Relative clauses (begin by relative pronouns which, who, that)
3. Indirect questions (begin by question words or if/whether)

Examples:
"The dependency is trivial, **because** $Y = f(X)$."
"X and Y are linearly independent, **if** the correlation coefficient, $corr(X, Y)$, is zero"
"Let c_i be the cluster **which** is closest to x.
"We select the first model **that** fits the data."
"First we should study **what** is the relationship between X and Y."
"The main problem is **whether** X can be applied in Z."
"We analyze the conditions **under which** X can be applied."

4.11.3 Sentence length?

Recommendations:

- always less than 30 words, preferably less than 20 words!
- 1-3 clauses
- expresses one idea

If you tend to write too long sentences, try the following:

1. Identify the main subject-predicate-object section
2. Prune or compress everything else, which is not needed
3. Check the verb structures and ask yourself if they could be shorter

E.g., verb structure "has been shown" can often be replaced by "is".

Notice! Don't go into the other extreme when you shorten sentences! If the clarity suffers, then a longer sentence is better.

Analogue: A good model of data does not overfit nor underfit, i.e., it is simple enough but still expresses all essential features. Now the sentence is a model of the idea you want to express.

4.11.4 Word order

The order of words has a strong impact on the meaning!

E.g., ”There is, however, currently no information about the limitations of quantum computers.” →
”However, there is no current information about the limitations of quantum computers.” →
”However, the limitations of current quantum computers are not known.”

The basic word order: subject–predicate–object

Recommendation: use the basic format **subject-predicate-object** in your sentences. You can add attributes, phrases and clauses, but don’t deviate too far from the basic format.

Why?

- Goal: put the most important information to the beginning of a sentence! ”X is a new algorithm for the TS problem.”
- Or begin by a familiar thing and put the new information to the end ”The probabilities are updated by the Bayes’ rule:” + the equation.
- Often the sentence is most informative, if you express the most important topic by the subject.
- This format helps to write clear and compact sentences.

The adverbs and prepositional phrases occur in order: **way, place, time**.

”The nearest neighbours can be identified **efficiently** (way) **in a dendrogram** (place)”.
”The values can be updated **easily** (way) **in linear time** (time)”.

Verb modifiers: in the middle of clause

Words which modify the predicate (the main verb) are located in the middle of the clause:

- Adverbs which express frequency: always, ever, never, often, seldom, sometimes, usually.

- Adverbs which express degree: almost, quite, certainly, completely, hardly, just, only, quite, really.
- Other words which modify the verb: already, also, still.

Hint: always consider if the word modifies the verb (the action) or the object (the target).

Verb modifiers are located

- before the predicate, if the verb consists of one word and is not the "be"-verb.
- after the first auxiliary verb, if the verb consists of several words.
- always after the "be"-verb.

E.g.,
"X often implies Y."
"The method gets sometimes stuck at a local optimum."
"The data was probably biased."

Task: Draw a decision tree for deciding the position of adverbs!

Problem: some words like "only" can modify also other words!
→ Put the word "only" before the word or phrase it modifies!
E.g., (notice the different meaning):

"X was the only method which could parse the $LL(1)$ grammar"
"X was the method which could only parse the $LL(1)$ grammar"
"X was the method which could parse only the $LL(1)$ grammar"

Adverbs which can begin the clause

If the adverb expresses time, it can be also in the beginning:
"Next, the data is loaded to the main memory."
This gives more emphasis to the word. It is also used, when there are other adverbs/prepositional phrases in the end of the clause.

Introductory adverbs like "obviously", "fortunately", etc. are always set to the beginning (if they are needed).

4.11.5 Combining clauses

Say the main message in the independent clauses! Use dependent clauses only to add details.

Combining two independent clauses

A **compound sentence**= two or more independent clauses which are combined by co-ordinating conjunctions or (rarely) by semicolons.

- In principle, you can combine several independent clauses, but in practice, combine only two main clauses (unless the clauses have the same subject which is mentioned only once).
- The ideas expressed in the clauses must be closely connected (otherwise separate sentences).
- The most common co-ordinating conjunctions are **and** and **but**.
  - **and** just links one idea to another (doesn't describe the relationship – typical for the children's language and dreams where things just happen). E.g., "The data is sparse and the model overfits easily."
  - **but** establishes an interesting relationship between the ideas → a higher level of argument. E.g., "The data was sparse, but the model did not overfit." (="Even if the data was sparse, the model did not overfit.")
- Commas? If the clauses have the same subject, then no commas. Otherwise usually a comma, unless the clauses are very short.

4.11.6 Combining clauses by sub-ordinating conjunctions

The basic form: an independent clause + a sub-ordinating conjunction + a dependent clause.

The most common sub-ordinating conjunctions express

1. a chronological order: **when, as, as soon as, while, after, before, until, since**

2. a conditional relationship: **if, unless**. **If**-clauses can also begin the sentence: "If the order is fixed, the episode is called serial." Notice: unless = if... not

3. a reason: **because** (Recommendation: reserve word "since" to express chronological order)

4. a purpose: **so that** (You can also use **in order to** + infinitive verb.)

5. an admission: **although, even if**

Examples:
"The search can be halted as soon as min_{fr} proportion of data is checked"
"The method is time-efficient, because all the parameters can be updated in one loop"

When you combine

- an independent clause + a dependent clause → sometimes but not always a comma (e.g., before **but**, but not before **that**).

- a dependent clause + an independent clause → always a comma.

4.11.7 Relative clauses

Correlate = referred word or clause, e.g., "An outlier is **a data point** which lies outside the clusters", "**Students** who solved the task...", "**The time complexity is quadratic**, which is sometimes undesirable..."

Restrictive and non-restrictive relative clauses

Relative clauses can be divided into two categories:

1. **Restrictive** or essential relative clause

    - defines the correlate
    - is necessary for understanding the sentence correctly
    - no commas
    - "X is an algorithm which solves the Travelling Salesman problem in $O(n^k)$ time."

2. **Non-restrictive** or non-essential relative clause

- gives only additional information about the correlate
- is separated from the main clause by commas
- "X, which solves the TS problem, works in $O(n^k)$ time."
- can refer to the previous clause

Notice that in spoken language we can drop the relative pronoun, if it is not the subject of the clause. In scientific writing, it is better to write all pronouns, because it should always be clear what you are referring.

Which relative pronoun to select?

1. Who, whose, whom

- **who** refers to a person.
- **whose** is the genitive form, e.g., "The student, **whose** solution was correct, got extra points."
- Notice that **whose** can refer to things and objects, too!
  "X is an example of problems **which** belong to class NP and **whose** known solutions are exponential."
- **whom** is used as an object and with prepositions, e.g., "The student, **about whom** I told you yesterday, wants to speak to you."

2. which

- refers to things and objects.
- can be used as a subject or an object or with prepositions.
  "X is a trick **which** helps to estimate the parameters more accurately."
  "Let X be the variable **which** Y depends **on**."
- when you refer to an entire clause, e.g., "The time complexity is quadratic, **which** is sometimes undesirable...", "All students cannot study themselves, **which** means that tutors are needed."
- In most cases, the genitive form can be either **of which** or **whose**. E.g.,
  "A computer, the cache **of which** is disabled, is less efficient..." = "A computer, **whose** cache is disabled, ..."

- However, when you refer to abstract nouns, use **of which** structure: "N.N. has introduced a new method the complexity **of which** is exponential." (**Problem**: is this rule still valid? Could we nowadays use also "whose"?)

3. that

- **can be used only in restrictive relative clauses!** → never use comma before it!
- can refer to people or things,
  "The student that has solved the task", "The task that was solved"
- can be used both as a subject and an object.
- If you need prepositions, they have to be in the end of the clause!
  "The problem **that** we talked **about**..."
- **that** is often used, when
  - the correlate is {**all, little, much**}, e.g., "all that we know"
  - with **superlative**, "the best solution that we can invent"
  - with ordinal numbers (**first**, **second**,...), **last**, and **only**,
    "the only algorithm that is comparable with X is Y"

4. what

- **what** contains also the correlate
- Only in special expressions!
- E.g.,
  "WYSIWYG means 'What you see is what you get'."
  "This is what we know so far."

Extra: When to use who and when whom?

Problem: Is the relative pronoun a subject or an object (who or whom)?

Hint: turn the subordinate clause around and substitute the relative pronoun by a personal pronoun. If you can use "she" or "he", it is subject (who), if "her" or "him", then it is an object (whom).

E.g., "N.N. is the student who/whom I mentioned earlier"
→ "N.N. is the student. I mentioned her/him earlier"
→ "whom" is the correct choice.

4.11.8 Indirect questions

The dependent clause begins by a question word **what, why, when, where, how** or **if/whether** when the corresponding direct question begins by a verb.

"First we should study **what** is the relationship between X and Y."
"The main problem is **whether** X can be applied in Z."

- The word order is direct!
- No auxiliary word **do**
- No comma!
- No question mark

4.12 Paragraphs

How to combine sentences? How to begin paragraphs? How to link paragraphs to each other? Introductory paragraphs (at the beginning of a chapter)

4.12.1 Combining sentences in a paragraph

1. Use (but do not overuse!) conjunctions or transitional words:
    - Time links, when you describe a process: then, next, first-second-third, while, ...
    - Cause-effect links, when you describe reasons or results: therefore, as a result, thus, ...
    - Addition links, when you add points: in addition, moreover, similarly, ...
    - Contrast links, when you describe two sides of one thing: however, despite (=in spite of sg), ...

- Other: For example,...

2. Link the beginning of a sentence to the end of the previous sentence. E.g., the subject of sentence 2 is the object of sentence 1.
   ”A model consists of a model structure and model parameters. The model structure defines...”

3. Repeat the key terms throughout the paragraph. However, do not repeat the same word twice in one sentence.

Task: Search useful expressions from the text excerpt given to you!

4.12.2 Dividing a section into paragraphs

Logical structure

Logically structured disposition (topic outline) is the most important thing in writing!

Analogy: In software engineering, the earliest errors (in specification and design phases) are the most expensive, if they are not recognized in the beginning. If you don’t plan, you write awful spaghetti code which nobody understands or can debug. Similarly, writing an illogical or a poorly organized disposition can cause serious problems. In the worst case, you have to write everything again!
→ Spend time and write the disposition carefully!

An iterative process:

1. The main structure of the whole thesis: the main chapters and their contents in a couple of sentences or key words. The order of chapters.

2. For each chapter (or a paper), the main sections + key words, introductory sentences or phrases. The order of sections.

3. In each section, the subsections or paragraphs. The introductory sentences, key words, and the order of paragraphs. List the related tables and figures.

Mark the points you wish to emphasize!

Suggestion: put your disposition aside for a while, before you begin writing.

A paragraph

The topic for each paragraph must be clearly stated – usually in the first sentence = **topic sentence**.

- Helps the reader: tells what the paragraph is about.
- Helps the writer: forces you to organize the material logically.
- In an ideal case, you get a summary of the whole section by reading the topic sentences.
- If you cannot write a clear topic sentence, ask yourself whether the paragraph is needed at all!

Other good advice:

- Never begin with unimportant words. The beginning of a paragraph is the most important.
- Omit superfluous phrases like
  ”First let us consider...”
  ”An interesting example which must be mentioned in this context is...”
  ”Next it must be noted that...”
- Emphasize important things by
  - telling them in the beginning of a paragraph or beginning of a sentence,
  - expressing them in short sentences,
  - repeating the key words, or
  - numbering.
- Keep the same verb tense (change it only for good reasons).
- Express parallel things in parallel structures.

If it is hard to divide a section into paragraph, list the things in a bullet list. Arrange the items and give them mini-subheadings. All items under one such heading belong to one paragraph. Tell the topic (expressed in the heading) in the topic sentence.

4.12.3 Introductory paragraphs

In the beginning of each chapter or a section having subsections, give 1–2 introductory paragraphs. These paragraphs tell what the chapter or section is about, i.e., it introduces the topics of sections or subsections. In the beginning of a chapter you can also introduce the main theme or problem and motivate the reader.

Suggestion: one brief paragraph (a few of sentences) in the beginning of a section, a longer paragraph or possibly two in the beginning of a chapter.

E.g., for the section ”Correlation analysis”:

”In the following, we recall the most common measure for correlation, Pearson correlation coefficient. We discuss restrictions and extensions of the common correlation analysis. Finally, we analyze the *ViSCoS* data by Pearson correlation and correlation ratios to reveal linear and non-linear dependencies.”

In the beginning of chapter ”Modelling dependencies between attributes” (could be briefer!):

”The main goal of predictive modelling is to predict a target variable Y from a set of other variables $X = \{X_1, ..., X_k\} \subseteq R$. Variables X are called *explanatory*, because they explain Y. The existence of such model requires that Y depends on X. Thus, the first step of modelling process is the descriptive analysis of dependencies between Y and X. The task is two-fold: First, we should select an attribute set X which best explains Y. Then we should analyze the type of dependency. Given this information, we can select the appropriate predictive modelling paradigm and define restrictions for the model structure.

In the following, we define the main types of dependencies for categorical and numeric data. We introduce three techniques (correlation analysis, correlation ratios, and multiple linear regression) for modelling dependencies in numeric data and four techniques (χ^2 independence test, mutual information, association rules, and Bayesian networks) for categorical data. In both cases we begin by analyzing pair-wise dependencies between two attributes, before we analyze dependencies between multiple attributes $X_1, ..., X_k$ and the target attribute Y. This approach has two benefits. First, we can avoid testing all 2^k dependencies between subsets of $\{X_1, ..., X_k\}$ and Y, if Y turns out to be independent from some X_i. Second, this analysis can reveal important information about suitable model structures. For example, in some modelling paradigms, like multiple linear regression and naive Bayes model, the explanatory variables should be independent from each other. Finally, we analyze the suitability of described modelling techniques for educational domain."

4.13 Punctuation

Goal: to make the text clearer. Unfortunately, the English punctuation rules (especially the use of comma) do not always coincide with the rules of your mother tongue.

Usually you manage with just two marks: **full-stop** and **comma**! The basic rules for other marks are:

- Use colon ':' only when needed.
- Avoid semicolon ';' and dash '–'.
- Avoid unnecessary parentheses '('...')'.

4.13.1 Full-stop

Full-stop ends a full sentence. Do not use comma instead of full-stop to separate independent clauses which are not logically related.

4.13.2 Comma

Comma is used

1. To separate introductory phrases and conjunctions (however, thus, similarly, etc.):
   "Ideally, all references are entered into a bibtex database."
   "Theorem 1 is important for two reasons. First, it allows us to... Second, it ..."
   "Despite its high time complexity, X is often used..."
   "For example, we can search episodes in www log data..."

2. When the sentence begins with a dependent clause:
   "Since $\overline{x}$ is a statistic, it is also a random variable."
   "If this condition is not satisfied, then the confidence bounds cannot be used."

3. When a non-restrictive relative clause is embedded into an independent clause or ends a sentence:
   ”X, which is responsible for data preprocessing, initializes Y.”

4. When two phrases with the same meaning are used side by side:
   ”One of the most useful statistics is $\overline{x}$, the sample mean.”

5. When the sentence begins by an infinitive structure (a clause substitute):
   ”To find the lower bound for the confidence interval, we isolate...”

6. To separate items in a list of three or more items. The ”Oxford comma” = the last comma before **and**, **or**, or **nor**, is optional:
   ”X is simple, fast**,** and easy to implement”

7. To avoid ambiguity (these could be said more clearly):
   ”Instead of hundreds, thousands of rows of data is required”
   ”Instead of 20, 50 students participated...”
   ”What the actual reason is, is not fully understood”
   (better: ”The actual reason is not fully understood”)

No comma is used

1. When an independent clause is followed by a restrictive relative clause or is embedded with a restrictive rel. clause (especially before **that**). Exception: ”It must be remembered, however, that...”

2. Between two independent clauses (in British English).

3. Before an indirect question.

4. When you begin by a prepositional phrase expressing the place. ”In this section we discuss...” ”In Chapter 3 we defined...”

4.13.3 Colon

Use colon between a grammatically complete introductory clause and a final phrase or clause that illustrates or extends it. If the following clause is a complete sentence, it begins with a capital letter.

”The formal definition of X is the following: (definition here)”
”X has several benefits: It is efficient, robust, and easy to implement.”

4.13.4 Dash

Dash is nearly always used in pairs. You can always use commas instead of dashes. Additional details can also be separated by parentheses. Notice that dash interrupts the continuity of a sentence!

Advice: Do not use dash, if you are not sure how to use it!

"The two students – one cs student and one maths student – were tested separately."

4.13.5 Semicolon

Semicolon separates two independent clauses. It is stronger than a comma but weaker than a full-stop. You can always replace it by a full-stop, and sometimes by a comma structure.

Advice: Save semicolons to program code!

Suits to separate independent clauses in a list:

"Metric d has three properties:

1. d is reflexive, i.e., $d(x, x) = 0$ for all x;
2. d is symmetric, i.e., $d(x, y) = d(y, x)$ for all x, y;
3. Triangular inequality holds for d, i.e., $d(x, z) \leq d(x, y) + d(y, z)$.

or to separate elements in a series which already contains commas:
"The colour order was red, yellow, blue; yellow, red, blue; or blue, yellow, red."

4.13.6 Quotation marks

Quotation marks are necessary when you use a direct citation!

You can use them also when you introduce a word or phrase used as an ironic comment, as slang, or as an invented expression. Use the quotation marks only when the new term is mentioned for the first time!

"Researchers have developed several measures to evaluate the "interestingness" of an association rule."

Notice: when you use a word or letter as an linguistic example, you can use a special font, e.g., italicize it (just be systematic with the font you select). "According to algorithm X, words *cat* and *God* were similar."

Similarly, when you mention variable names, values etc., use a special font (unless they mathematical symbols → $ characters (math mode). E.g., "X can have three values `low, medium, high`." "`Action1` is selected with the probability of 0.6 and `Action2` with the probability of 0.4."
(In latex: `{\tt Action}`)

4.13.7 Parentheses

Parentheses are used for two purposes:

- To introduce an abbreviation
  "*Minimum description length* (*MDL*) principle is often used to..."
- To add extra details. **Advice**: do not overuse them!
  "Two common choices are to represent a cluster by its centroid (central point) or by its boundary points."
  "The idea of *minimum edit distance* is to determine the minimum number of atomic operations (insertion, deletion, substitution) needed to transform one string to another."

Sometimes you can give extra references (extra reading) in parenthesis:

"To restrict the future development of adaptive learning environments as little as possible, we have adopted a wide (and visionary) view of *context-aware computing* (*ubiquitous computing* (see e.g., [14,17]), in which the whole context – the user, her/his actual situation and all relevant information – is used for determining the most appropriate action."

4.14 Genitive: 's or of?

Default: For animate things (people and animals) **'s**: **possessor's possessed** (in plural **possessors' possessed**).
For inanimate things **of** structure: **the possessed of possessor**.

Nowadays, **'s** genitive can be used also for inanimate things, especially in certain special cases (especially in American English!). However, never use **'s** genitive for abstract things!

"The meaning of life", "The time complexity of algorithm X".

4.14.1 Special cases where 's genitive is used for inanimate things

1. Temporal expressions: "two weeks' holiday", "an hour's work". However, in some expressions only **of** is possible: "in the middle of August".

2. Sometimes when the noun is geographical (country or city): "London's sights". However, if the target expresses place (town, city, kingdom, island), then **of**: "The city of Joensuu"

3. When the noun expresses place and is followed by superlative: "The world's best computer games".

4. When possessor is a collective noun, **'s** is often used, but **of** is also possible: "The board's decision".

5. When you express part–whole relation, **'s** is often used, especially in body parts "the car's doors", "the cat's whiskers".
   Hint: If the possessed necessarily belongs to the possessor → **'s**, if the possessed can exist alone → **of**.

6. Some special phrases: "For goodness' sake".

4.14.2 When "of" structure is necessary

's genitive makes the possessed noun definite, i.e., **possessor's possessed = the possessed of possessor**.
→ definite article **the** in the **of** genitive.
If you want to express that the possessed is indefinite (one of many), **of** genitive is the only choice (even if you refer to people): "a son of the mayor".

4.14.3 Possessive form of pronouns

When the possessor is a pronoun, use the possessive pronouns!

{my, your, her/his, its, our, your, their} + possessed.

If the possessive pronoun is not followed by noun, then special forms {mine, your, hers/his, ours, yours, theirs}. Seldom needed in scientific writing! (In spoken language e.g., "Whose cat is this? It is mine.")

In some special cases (rarely) you can use structure "of it" (referring to inanimate things) to emphasize the possessed. "I don't remember the name of it."

Chapter 5

Writing master's thesis

5.1 Parts of the master's thesis

5.1.1 Abstract

- Tells compactly the research problem, methods and results.
- At most 1 page, no literature references.
- In the end ACM classes + possibly key words.

5.1.2 Introduction

Typically 4-7 pages.

The introduction should define the problem clearly and give sufficient background information for the following chapters. However, no details, yet!

- What is the purpose of the research? Main research questions?
- What is the scope? Indicate explicitly all limitations and restricting assumptions!
- Why the topic is important or interesting?
- What methods are used?
- Briefly references to related research (just the main references – more references in the background chapters or throughout the thesis)
- Emphasize your own contribution: what is original or new?

Introduction can be divided into sections, e.g.,

1. Problem description, motivation and background (the heading could be ”Overview”, ”Problem”, ”Motivation and background” etc.)
2. Research objectives
3. Organization (contents of the chapters)

5.1.3 Main chapters

Usually 4-5 chapters (in addition to Introduction and Conclusions). A good idea is to begin from background theory or related research.

5.1.4 Conclusions

Just 1-3 pages!

- Summarize the main results in a general level.
- Tell what was your own contribution and what was based on other sources.
- Possibly also criticism (e.g., limitations), alternative approaches, topics for future research.
- No more new results and seldom any references (at most for alternative, unmentioned approaches).
- Conclusions are very short, if you have a separate chapter Discussion.

5.1.5 References

- A rule of thumb: at least 20 references, but no more than 50. 30–40 is often ideal.
- The number of references depends on the topic. More references are required in a literature review than in empirical research.
- The number of references is not a merit, but their quality is more important!
- The references should be relevant, up-to-date, and ideally represent different approaches or schools among researchers.

- **Important**: all sources (listed in References) must be cited in the text and the text should not contain any citations which are not listed in References!
  → Bibtex takes care of this automatically. If you type references manually, latex complains only about missing references, but not about extra references.

5.1.6 Appendices

- Additional material which is relevant to the research and is referred in the text, but not as important to be included in the main chapters. E.g., questionnaire forms or description of relevant implementation details. If you have a lot of result tables or graphs, consider presenting only the most important/summary tables in the main text and put others into an appendix.

- No chapter numbers, but enumerate the appendices (Appendix A, Appendix B,...). If you have only one appendix, then just "Appendix".

5.1.7 Examples of structuring a master's theses

The following examples give suggestions how to structure the main chapters between Introduction and Conclusions. Discussion may be a separate chapter (if long) or a section under Experiments (interpretation and reflection of results), preferably complemented by some more discussion in Conclusions.

A new application or a method

Here, a new method is developed to solve a problem or existing methods are applied in a new domain. The latter may be a comparison of existing methods to find an optimal one or a software that is tailored for a certain purpose and composes different features.

- Background theory and main concepts (1–3 chapters)

  - Related research (other existing solutions to the same or similar problems) can be one section here or under Introduction

- New method

  - not needed, if only comparison of existing methods

- Evaluation: comparison to other methods, empirical tests, or theoretical analysis

Typically, the first background chapter describes the problem area in detail, its special features and requirements for an ideal solution. If the problem is to apply some computational method in a new domain, the domain needs to be explained in such a level that the reader understands why certain method is chosen, how it is applied, and how its performance can be evaluated. Use some informative title (related to the specific domain/problem) or use generic titles like "Problem definition" or "Preliminaries".

The next background theory chapter or chapters describe generic methods that are used to solve the problem (both methods that are used in the work and their main competitors, to justify why certain methods were chosen for experiments). There may be a separate section describing how the problem has been solved in past (e.g., "Related research") unless all described methods concentrate on the specific problem. Sometimes, related research is described in one section under Introduction.

If no new method is introduced, there can be more theory chapters on existing methods (e.g., different approaches to data representation, feature extraction and selection, modelling) with careful analysis of their suitability to the given problem.

Literature review with theoretical analysis

A theory or a model is analyzed based on literature. Often one compares or analyses suitability of existing approaches to a new problem.

Your own contribution: how the methods or models are described in a uniform manner, analyzed and compared. The work can be deepened by inventing properties that explain the behaviour or pathological cases that lead to failure.

Now the existing literature is referred in all chapters and there is no need for a separate section "Related research".

- The problem + criteria for an ideal solution method
- Potential solution methods + analysis of their suitability (2-3 chapters)
- Possibly discussion (comparison, new solution ideas)

Empirical research

E.g., a new method or tool is tested with real users. Here you can consider variants of a classical structure:

- Background (main concepts and background theories)
- Methods (i.e., participants, data, evaluation protocol)
- Results
- Discussion

5.2 Master's thesis process

"The purpose of a thesis is to train the mind of the writer and to show how far it has been trained." [1, p. 141]

5.2.1 Reading literature

Problem: you should get a wide view of the existing research on the topic, but your time to search and read literature is limited!

- Try to find the most relevant articles. Begin from tutorial, review or survey papers, if available.
- To get a wider perspective, search papers by different authors/research groups. If there are several approaches to solve or study the problem, try to study something from all of them (or all of the main approaches).
- Use several digital libraries or bibliographies for searching – one collection may be biassed.
- Plan how much time you can spend for studying literature! In some point you have to stop collecting new material and begin to write.

5.2.2 Planning

Well planned is half done!

- Begin by brainstorming. Draw concept maps. Discuss with your friends or supervisors. Write down all ideas which come into your mind.

- Collect literature and scan through it. Select the most important sources.
- Try to write the disposition as early as possible. Process it with your supervisor until it looks good (logical structure and order).
- List the main research problems (in the form of questions).
- Fix your notations. When you read literature, you can translate notations to your own language. (You can also define notations with latex macros so that you can change them easily later.)

5.2.3 Difficulty to get started

Hints:

- Arrange a comfortable working place. Reserve time for writing every day. Try to make writing a routine for you!
- Set deadlines. Preferably fix them with your supervisor – it is always more effective.
- Work together with your friend. You can set the deadlines, discuss your topics, and read each other's texts. After good work you can reward yourself by doing something fun.
- Imagine that you are writing to your friend about your research topic!
- Summarize articles you have read. It is never waste of time – at least you learn!
- Begin to write immediately, when your disposition is finished.
- Write down ideas when they come – even in the middle of night.
- Invent good examples and write them down.
- If some part is difficult to write, begin from an easier one. Write the difficult parts, when you are in a good working mood.
- Draw figures which describe some method or model and write a description.
- Try to divide the problem or phenomenon into subproblems or parts and describe them separately.

- Collect main concepts and write definitions for them. Fix the notations and write a table where they are introduced.

How to write the beginning of chapters?

- Look at the opening sentences of similar work by other people
- Begin, for example, with a summary, a statement of the problem, a hypothesis, necessary and interesting background information, a new idea, an accepted procedure (then explain advantages of another procedure), ...
- Don't spend too much time trying to find an effective beginning – you can always modify it afterwards.
- Go straight to the point and, if possible, refer to things that you expect your readers to know (vs. constructivism).

5.2.4 Revising

"The time taken in planning, writing and revising is time for thought. It is well spent, for when the work is complete your understanding of the subject will have been improved." [1, p. 44]

- First of all, admit that the first draft(s) is not perfect! Ask criticism and respect it. Good criticism is really valuable.
- If possible, ask at least two people to read your thesis. Preferably one who is an expert on the subject, and one who is not. E.g., your supervisor and one of student colleagues.
- You can write and revise your work for ever, but in some point you have to stop! One trick is that you don't allow yourself to gather any more new literature.
- Have a break when your work is finished. At least, sleep one night before revising the text yourself.

Technical hints:

- Read text aloud and check if it sounds well.
- Try to misunderstand your own text, looking at only the literal meaning (like "Female students draw concept maps more often than male students.")

- Check all references. Especially, are names correctly spelled?

- Save old versions, you may need them afterwards.

5.2.5 Technical notes

Technical terms

If there is no widely accepted definition for it, then

1. Tell whose definition you follow and give this definition with reference, or

2. Give a definition yourself and tell that in this work the term is defined as given.

"If a technical term is used as a substitute for an explanation, it gives no more than an impression of knowledge. ... Unless a technical term can be defined clearly and then used with accuracy and precision, it may conceal our ignorance and obscure the need for further research, and it should have no place in scientific writing." [1, p. 62]

Symbols

- Don't use the same symbol for different things!

- Try to use also indexes in a uniform manner. E.g. if the $i = 1, ..., n$ is the number of rows and $j = 1, ..., k$ the number of attributes in one place, don't change them in another place.

- If some special notation is widely used in literature, follow it.

- If different sources use different notations, harmonize them. (Fix one notation and translate all notations to your own "language".)

- Do not use Greek or Hebrew letters unless there is a reason (e.g., a conventional notation, like α threshold for statistical significance).

- Try to use a similar font type for similar type of notations (e.g., regular font for single variables, boldface capital letters for sets, bold lowercase for vectors, calligraphy for collections of more complex objects)

Equations

Avoid listing mathematical equations! Try to integrate equations into sentences so that the results is readable.

Do not replace words by mathematical symbols (e.g. $\forall$) in the text.

Appendix: Literature references

In the thesis, you should usually give complete references (all fields, no abbreviations). In papers, you can use abbreviations and skip some fields (like editors of the proceedings or the publisher's address).

I Journal and conference articles

Most of your references should belong to these groups!

1. A journal article:

   <Authors>. <Title>. <Journal>, <volume> (<issue>): <pages>, <year>.

2. A conference article:

   <Authors>. <Title>. In <editors>, editors, <book title>, <pages>. <Publisher>, <publisher's address>, <year>.

Examples

A journal article:

V. Cheng, C.H. Li, J.T. Kwok and C.-K. Li. Dissimilarity learning for nominal data. *Pattern Recognition*, 37(7):1471–1477, 2004.

A conference article:

W. Hämäläinen. Efficient discovery of the top-K optimal dependency rules with Fisher's exact test of significance. In G.I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010)*, pages 196–205. IEEE

Computer Society, Los Alamitos, California, 2010.

H. Grosskreutz and S. Rüping. On subgroup discovery in numerical domains. In W. Buntine, M. Grobelnik, D. Mladenić, and J. Shawe-Taylor, editors, *Machine Learning and Knowledge Discovery in Databases, Proceedings of ECML/PKDD 2009, Part I*, volume 5781 of *Lecture Notes in Computer Science*, pages 210–226. Springer, Berlin, Heidelberg, 2009.
(Here "Lecture Notes in Computer Science" is the series name.)

Notes:

- When at least three authors or editors, you could give only the first one and replace the rest with "et al.", e.g., "G.I. Webb *et al.*"

- Delimiters between fields depend on the style (colon, comma or full stop).

- In papers (with tight page limits), you can use abbreviated references, like

  H. Grosskreutz and S. Rüping. On subgroup discovery in numerical domains. In *Machine Learning and Knowledge Discovery in Databases, LNCS 5781*, pp. 210–226. Springer, 2009.
  (Now the series abbreviation and volume are important, since the proceedings title is abbreviated.)

- Warning! In bibtex, use type *@incollection* instead of *@inproceedings* to get the publisher's address into a correct place (otherwise field "address" is interpreted as the conference location, which you don't need to give).

II Referring to books

1. A book:

   <Authors>. <Title>. <Publisher>, <publisher's address>, <year>.

2. A book chapter in a collection (an edited book):

   <Authors>. < Title>. In <Editors>, editors, <Book title>.
   <Publisher>, <publisher's address>, <year>.

Note: Give edition, when relevant (e.g., if the page numbers have changed and you cite to certain pages).

Examples

F.M. Lord. *Applications of item response theory to practical testing problems.* Lawrence Erlbaum Associates, New Jersey 1980.

D.W. Scott and S.R Sain. Multi-dimensional density estimation. In C.R. Rao and E.J. Wegman, editors, *Handbook of Statistics—Vol 23: Data Mining and Computational Statistics.* Elsevier, Amsterdam, 2004.

P. Smyth. *Data mining at the interface of computer science and statistics.* In R.L. Grossman, C. Kamath, P. Kegelmeyer, V. Kumar, and R. Namburu, editors, *Data Mining for Scientific and Engineering Applications*, volume 2 of *Massive Computing*, chapter 3, pages 35–61. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

(Here "Massive Computing" is the series name.)

III Technical reports and theses

Use technical reports only exceptionally, since they have not been reviewed. If the same report has been later published as a conference or journal paper, use it instead. Doctoral theses have usually gone through a careful review and are good sources. Master's theses have been checked by supervisors, but they may still contain errors – use cautiously.

1. A technical report:

   <Authors>. < Title>. <Report series> <report number>, <Institution>, <year>. (Optionally also url, if available online; consider also retrieval date, if the address may change.)

2. A master's thesis:

   <Author>. < Title>. Master's thesis, <Department>, <University or institution>, <year>.

Examples

A.K. Dey and G.D. Abowd. Towards a better understanding of context and context-awareness. GVU Technical Report GIT-GVU-99-22, College of Computing, Georgia Institute of Technology, 1999. Retrieved 1.1. 2006 from ftp://ftp.cc.gatech.edu/pub/gvu/tr/1999/99-22.pdf.

(Here the url and retrieval date are optional, since the report has been published also in a paper form.)

A. Norris. Multivariate analysis and reverse engineering of signal transduction pathways. Master's thesis, Department of Mathematics, Institute of Applied Mathematics, University of British Columbia, 2002.

P. Tuoresmäki. Chip-seq-piikkien koostaminen ydinestimoinnilla [Aggregation of ChIP-seq peaks using kernel density estimation]. Master's thesis, School of Computing, University of Eastern Finland, 2015.
(Here the thesis is written in Finnish, a translated title given in brackets.)

W. Hämäläinen. *Efficient Search for Statistically Significant Dependency Rules in Binary Data.* PhD thesis, Department of Computer Science, University of Helsinki, Finland, 2010. Series of Publications A, Report A-2010-2.

IV Referring to internet articles

Be default, all sources should have been published! You can cite internet articles if they have been published in an online journal or book. Other internet sources can be referred only for a good reason (i.e., if the information is not available elsewhere).

- If you refer to an article, which is available on the internet but has been **published in a paper form**, give the normal reference to the paper version. The url address is not necessary, but it can be given to help the reader to find the article.
- If an article has been **published only in an internet journal**, give the reference like to any common journal article, but replace the page numbers by the url address.
- If both URL and DOI are available, prefer DOI (it provides a persistent link to the location on the internet).
- If the article **exists only on the internet but is not published**, give the retrieval date and the URL address in the end of reference. E.g., ”Retrieved March 3, 2006, from `http:www.kissastan.edu/bnetworks/bnarticle.html`.
- If you refer to **an internet textbook**, give the normal book information if possible (author, book title, publisher, year; sometimes the

internet book has also a publisher like a company, institution, etc.). If it doesn't have any publication year, then give the date when the book was accessed by you. Always give the URL address.

- If the page address or contents may change, give the date when you retrieved it. E.g., "Retrieved <date> from <URL>" or "Available at: <URL> (Accessed: <date>)"

Examples

An unpublished internet source:

Fox, E.: Details of clustering algorithms (lecture notes). Virginia Tech, 1995-1996. Retrieved January 1, 2006, from `http://maya.cs.depaul.edu/~classes/ds575/clustering/CL-alg-details.html`.

An internet textbook (a special case, no author is mentioned, only the company – Xycoon – which has produced the book):

Xycoon: *Linear Regression Techniques*, Online Econometrics Textbook, chapter II. Office for Research Development and Education, 2000-2006. Retrieved January 1, 2006, from `http://www.xycoon.com/`.

Referring to software

- **Standard software tools and programming languages** like LaTeX, Matlab, and Java do not need any references.

- If you use **special tools or programs** with limited distribution it is recommendable to give the reference. E.g.,

  BCAT [A Bayesian network tool]. Retrieved March 3, 2006, from `http:www.kissastan.edu/bcat-tool/bcat3.0.html`.

- If you know the organization which has produced the work, give it in the publisher position (before retrieval information). If somebody has rights to the software, mention her/him as the author.

- Sometimes the home pages of tools or library packages tell the desired reference. E.g, there may be a reference to a paper where the underlying algorithm has been introduced.

Bibliography

[1] R. Barras: Scientists must write. A guide to better writing for scientists, engineers and students. 2nd edition, Routledge, London and New York, 2005

[2] J. Peat et al.: Scientific writing – easy when you know how. BMJ Books, London, 2002.

[3] Publication Manual of the American Psychological Association. Fifth Edition. American Psychological Association, Washington DC, 2002.

[4] Strunk, W.: Elements of Style. Priv. print, Ithaca, NY, 1918. On-line edition published July 1999 by Bartleby.com. `www.bartleby.com/141/`. Retrieved March 1, 2006.