



Universität Hamburg

DER FORSCHUNG | DER LEHRE | DER BILDUNG



CHAI

Humanities-Centered AI

# Understanding Data vs. Machine Training

**Malte Luttermann – Institute for Humanities-Centered  
Artificial Intelligence (CHAI)**

November 14, 2025

# Overview of Contents

- (1) Programming language Python
  - (a) Introduction and first steps
  - (b) Basics
  - (c) Advanced
- (2) Markup languages
  - (a)  $\text{\LaTeX}$
  - (b) Markdown
- (3) Development environments
  - (a) Jupyter notebooks
- (4) Version control
  - (a) Git and GitHub
- (5) Scientific computing
  - (a) NumPy and SciPy
- (6) Data processing and visualisation
  - (a) Pandas, matplotlib, and NLTK
- (7) Machine learning (scikit-learn)
  - (a) Basics (datasets, analysis)
  - (b) Simple methods (clustering, ...)
- (8) Deep learning
  - (a) PyTorch

# Topics for Today

## (1) Markup languages

- Content, structure, and form
- Semantic markup
- Markdown
- L<sup>A</sup>T<sub>E</sub>X



The L<sup>A</sup>T<sub>E</sub>X Project



# Acknowledgement

- The upcoming slides are taken from the following lecture and have been translated and partially modified:
  - Dr. Magnus Bender: »[Python für Machine Learning und Data Science](#)« as part of the German course »Werkzeuge für das wissenschaftliche Arbeiten«

# Content, Structure, and Form

## ■ Content

- Meaning of a text
- The words (and sentences)

## ■ Structure

- Document composition
- Paragraphs, chapters, and headings

## ■ Form

- Appearance of the document
- Presentation (colors, font, highlighting, b

▼ General information  
Announcements  
▼ Content  
Lecture  
Seminar  
▼ Seminar topics

▼ General information

**Instructors:** Prof. Dr. Ralf Möller; Dr. Marcel Gehrke; Malte Luttermann

**Event type:** Lecture + seminar (Intellectics students) or lecture only (others)

**Hours per week:** 2 (lecture) + 2 (seminar)

**Credits:** 8 (for the whole course)

- 3 (lecture)
- 5 (seminar)
  - 3 (for active seminar attendance and presentation)
  - 2 (for term paper at the end of the semester)

---

**Place and time**

**Lecture:** Friday 10:15 - 11:45, room ESA K (Albrecht-Mendelssohn-Bartholdy-Hörsaal)

**Seminar:** Friday 12:15 - 13:45, room Phil A 12006

**Start of the lecture and seminar:** Friday, October 17, 2025

- 3 (for active seminar attendance and presentation)
- 2 (for term paper at the end of the semester)

---

**Place and time**

**Lecture:** Friday 10:15 - 11:45, room ESA K (Albrecht-Mendelssohn-Bartholdy-Hörsaal)

**Seminar:** Friday 12:15 - 13:45, room Phil A 12006

**Start of the lecture and seminar:** Friday, October 17, 2025

# Semantic Markup

- Separation between content and form
- Specification of the content with structure
  - E.g., Heading »My Exercise«
- Afterwards formatting of the structure
  - E.g., Headings should be **large and bold**

*Structure*

HTML, Markdown,  $\text{\LaTeX}$

Word, Libre Office, Pages

PDF, Vector graphics

Pixel graphics

*Form*

# Markdown



```
1 # Markdown
2
3 > From [Wikipedia](https://en.wikipedia.org/wiki/Markdown), the free
   encyclopedia
4
5 ## Article
6
7 Markdown is a lightweight markup language for creating formatted text
   using a plain-text editor.
8 John Gruber and Aaron Swartz created Markdown in 2004 as a markup
   language that is appealing to human readers in its source code
   form.
9
10 Paragraphs are separated by a blank line.
11
12 Two spaces at the end of a line
13 produce a line break.
14 Text can be styled italic, bold, or monospace.
```

Content and structure  
but no form

# Markdown



## Markdown

From [Wikipedia](#), the free encyclopedia

## Article

Markdown is a lightweight markup language for creating formatted text using a plain-text editor. John Gruber and Aaron Swartz created Markdown in 2004 as a markup language that is appealing to human readers in its source code form.

Paragraphs are separated by a blank line.

Two spaces at the end of a line produce a line break. Text can be styled *italic*, **bold**, or `monospace`.

One possible form

# Markdown



```
1 # Heading
2 ## Subheading
3
4 > Quote
5
6 - **Bold**
7 - *Italic*
8 - `Code`
9
10 1. [Links](http://www.example.com)
11 2. ![Images](http://www.example.com
12   /image.jpg)
13
14 ----
```

## Heading

### Subheading

#### Quote

- **Bold**
- *Italic*
- `Code`

#### 1. Links



#### 2.

# Markdown



```
1 Footnotes are also possible [^1].
```

```
2  
3 | A | B | C |  
4 |---|---|---|  
5 | 1 | 2 | 3 |
```

```
6  
7 ```python  
8 print("Hello!")  
9 ```
```

```
10  
11 - [x]  $\frac{1}{2}^2$   
12 - [ ]  $\frac{x_1^2 + 5x_2}{x_1}$ 
```

```
13  
14 [^1]: This is extended Markdown.
```

Footnotes are also possible[^1].

A	B	C
1	2	3

```
print("Hello!")
```

- [x]  $\frac{1}{2}^2$
- [ ]  $\frac{x_1^2 + 5x_2}{x_1}$

[^1]: This is extended Markdown.

# Markdown



- By now a standard for simple text formatting
  - Communication (e.g., Moodle, Discord, Slack)
  - Code Editors (e.g., VS Code)
  - Version control platforms (e.g., GitHub, GitLab)
  - Collaboration platforms (e.g., HedgeDoc)

# Markdown Example: HedgeDoc



- Collaborative
- Markdown with extensions
  - Equations ( $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ )
  - Diagrams
- Features and Demo

The screenshot shows the HedgeDoc web interface. The left pane displays the source markdown code, and the right pane shows the rendered HTML output.

```

# Features

## Introduction

**HedgeDoc** is a real-time, multi-platform collaborative markdown note editor.

This means that you can write notes with other people on your desktop, tablet or even on the phone.

You can sign-in via multiple auth providers like Facebook, Twitter, GitHub and many more on the [*homepage*]().

If you experience any issues, feel free to report it on [*GitHub*](https://github.com/hedgedoc/hedgedoc/issues). Or meet us on [*Matrix*](https://chat.hedgedoc.org) for dev-talk and interactive help.

Thank you very much!

## Workspace

### Modes

#### Desktop & Tablet

<i class="fa fa-eye fa-fw"></i> View: See only the result.
<i class="fa fa-columns fa-fw"></i> Both: See editor and result at the same time.
<i class="fa fa-pencil fa-fw"></i> Edit: See only the editor.

#### Mobile

<i class="fa fa-eye fa-fw"></i> View: See only the result.
<i class="fa fa-pencil fa-fw"></i> Edit: See only the editor.

### Night Mode

When you are tired of a white screen and like a night mode, click on the little moon <i class="fa fa-moon-o"></i> and
  
```

The rendered preview on the right shows the following structure:

- Features**
  - Introduction**

HedgeDoc is a real-time, multi-platform collaborative markdown note editor. This means that you can write notes with other people on your **desktop**, **tablet** or even on the **phone**. You can sign-in via multiple auth providers like **Facebook**, **Twitter**, **GitHub** and many more on the [homepage](#).

If you experience any **issues**, feel free to report it on [GitHub](https://github.com/hedgedoc/hedgedoc/issues). Or meet us on [Matrix](https://chat.hedgedoc.org) for dev-talk and interactive help. **Thank you very much!**
  - Workspace**
    - Modes**
      - Desktop & Tablet**
        - View*: See only the result.
        - Both*: See editor and result at the same time.
        - Edit*: See only the editor.
      - Mobile**
        - View*: See only the result.
        - Edit*: See only the editor.
    - Night Mode**

When you are tired of a white screen and like a night mode, click on the little moon and turn on the night view of HedgeDoc.

The editor view, which is in night mode by default, can also be toggled between night and day view using the little sun.

# L<sup>A</sup>T<sub>E</sub>X

- Pronunciation: »LAH-tek« or »LAY-tek«
- A powerful typesetting system
  - Suitable for: Exercise sheets, reports, summaries, term papers, theses, ...
  - Less suitable for: Notes, lecture transcripts (better: Markdown)
- In particular, support for equations, bibliographies, tables of contents



# The First L<sup>A</sup>T<sub>E</sub>X Document

```
1 \documentclass[
2   12pt, % Font size
3   a4paper, % Paper size
4   parskip=full % Paragraph style
5 ]{scrartcl}
6
7 % File encoding
8 \usepackage[utf8]{inputenc}
9 % English language (e.g., for hyphenation)
10 \usepackage[english]{babel}
11 % Font
12 \usepackage[T1]{fontenc}
13 \usepackage{lmodern}
14
15 \begin{document}
16 My first \LaTeX{} document!
17 \end{document}
```

Preamble

Content

# The First L<sup>A</sup>T<sub>E</sub>X Document

```
1 \documentclass[
2   12pt, % Font size
3   a4paper, % Paper size
4   parskip=full % Paragraph style
5 ]{scrartcl}
6
7 % File encoding
8 \usepackage[utf8]{inputenc}
9 % English language (e.g., for hyphenation)
10 \usepackage[english]{babel}
11 % Font
12 \usepackage[T1]{fontenc}
13 \usepackage{lmodern}
14
15 \begin{document}
16 My first \LaTeX{} document!
17 \end{document}
```

My first L<sup>A</sup>T<sub>E</sub>X document!

1

# Another L<sup>A</sup>T<sub>E</sub>X Document

```
1 \begin{document}
2 \title{LaTeX Example}
3 \author{Malte Luttermann}
4 \date{\today}
5 \maketitle
6
7 \begin{center}
8   My \textbf{first} \LaTeX{} \textsc{document}!
9 \end{center}
10
11 \tableofcontents
12
13 \section{Section}
14   Lorem ipsum dolor sit amet, consetetur sadipscing elitr, ...
15
16   \subsection{Subsection}
17     Lorem ipsum dolor sit amet, ...
18
19 \section{Another Section}
20   Lorem ipsum dolor sit amet, consetetur sadipscing elitr, ...
21
22   Lorem ipsum dolor sit amet, ...
23 \end{document}
```

**LaTeX Example**

Malte Luttermann  
November 3, 2025

My first L<sup>A</sup>T<sub>E</sub>X document!

**Contents**

<b>1 Section</b>	<b>1</b>
1.1 Subsection .....	1
<b>2 Another Section</b>	<b>1</b>

**1 Section**

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, ...

**1.1 Subsection**

Lorem ipsum dolor sit amet, ...

**2 Another Section**

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, ...  
 Lorem ipsum dolor sit amet, ...

1

# Compilation of $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ Documents

- How to compile the document?
  - `pdflatex document.tex` (multiple times to resolve references)
  - `pdflatex document.tex`  
`bibtex document`  
`pdflatex document.tex`  
`pdflatex document.tex`
  - `latexmk -pdf document.tex`
  - If you use an IDE (e.g., VS Code + LaTeX Workshop Extension): Just press the compile button!

# L<sup>A</sup>T<sub>E</sub>X Template for the Written Term Paper

```
1 \documentclass[12pt]{scrarticle}
2
3 \newcommand{\coursetitle}{Understanding Data vs.\ Machine Training}
4 \newcommand{\courseyear}{Winter Term 2025/26}
5 % TODO: Insert your name here
6 \newcommand{\authorname}{Malte Luttermann}
7 % TODO: Insert your student ID here
8 \newcommand{\studentId}{0000000}
9
10 \input{res/preamble.tex}
11
12 \begin{document}
13 ...
14 \end{document}
```

# L<sup>A</sup>T<sub>E</sub>X Template for the Written Term Paper

## Term Paper

### Understanding Data vs. Machine Training

Winter Term 2025/26

Institute for Humanities-Centered Artificial Intelligence (CHAI Institute)  
of the University of Hamburg

Submitted by

**Malte Luttermann**  
Student ID: 000000

## 1 Introduction

This L<sup>A</sup>T<sub>E</sub>X template aims to assist you in writing your term paper for the course "Understanding Data vs. Machine Training". By using this template, you don't have to worry about formatting issues and can focus on the content of your term paper. In the following, we provide a brief overview on how to use this template.

If you have any questions regarding this particular L<sup>A</sup>T<sub>E</sub>X template or questions regarding the course "Understanding Data vs. Machine Training" in general, please don't hesitate to contact the instructors of the course.

## 2 Get Started with this L<sup>A</sup>T<sub>E</sub>X Template

This L<sup>A</sup>T<sub>E</sub>X template comes with a predefined structure and consists of multiple files. Specifically, you don't need to modify any of the files located in the directory `res/`, as these files contain the formatting settings for this template. The content of your term paper should be written within the environment `\begin{document}` ... `\end{document}`, located in the file `main.tex`. Currently, the file `main.tex` contains this explanatory text as a placeholder and looks like this (document preamble omitted for brevity):

```
\begin{document}
\input{res/titlepage.tex}
\clearpage
\setcounter{page}{1}
```

...

```
\bibliographystyle{res/named.bst}
\bibliography{literature.bib}
\end{document}
```

To get started with writing your term paper, just remove the placeholder text and place your content in the environment `\begin{document}` ... `\end{document}`. Please note that both the title page generated by

```
\begin{document}
\input{res/titlepage.tex}
\clearpage
\setcounter{page}{1}
```

as well as the bibliography section generated by

```
\bibliographystyle{res/named.bst}
\bibliography{literature.bib}
\end{document}
```

are required and should not be removed or changed. Before starting to write the actual content of your term paper, please make sure to customize your personal information by adjusting the following commands at the top of the file `main.tex`:

Left	Center	Right
1	2	3
4	5	6

Table 1: This is an example for a table.

```
% TODO: Insert your name here
\newcommand{\authorname}{Malte Luttermann}
% TODO: Insert your student ID here
\newcommand{\studentId}{0000000}
```

Simply replace the placeholder values with your actual name and student ID. If you wish to load any additional packages, feel free to do so (e.g., by putting `\usepackage{<package>}` commands in the preamble of the file `main.tex`).

The additional file `literature.bib` is used to store your bibliography entries in the Bib<sub>T</sub>E<sub>X</sub> format. You can add entries as needed and then cite them in your paper using the `\citep{<key>}` or `\citet{<key>}` commands. For example, `\citet{Russell2020a}` produces Russell and Norvig [2020] whereas the command `\citep{Russell2020a}` produces the citation [Russell and Norvig, 2020]. The bibliography section of your term paper will then automatically be generated based on the entries you included in the file `literature.bib`.

Hence, in total, there are two files that are relevant for you: `main.tex` for your actual content and `literature.bib` for citations.

## 3 Further Information About this L<sup>A</sup>T<sub>E</sub>X Template

You can create sections, subsections, and so on via the `\section{<title>}`, `\subsection{<title>}`, `\subsubsection{<title>}`, and `\paragraph{<title>}` commands. All illustrations (figures, drawings, tables, etc.) should be placed into floating environments, preferably floated to the top of a page (via the optional argument `[t]`). For instance, Table 1 shows an example table.

You can refer to any table, figure, or section using the `\ref{<label>}` command (where `<label>` is the label you assigned via the `\label{<label>}` command). Formulas can be typeset using `\$ ... \$` for inline math such as  $\gamma = 2 \cdot \alpha$ , or via the `\begin{align} ... \end{align}` environment for aligned equations, e.g.,

$$x = \prod_{i=1}^n \sum_{j=1}^m f_i^j \quad (1)$$

$$y = \sqrt{\frac{a}{b}} + \sqrt{\frac{b}{a}} \quad (2)$$

$$z = \sum_{i=1}^n (a_i + b_i)^2 \quad (3)$$

# L<sup>A</sup>T<sub>E</sub>X Beamer

- Beamer is a powerful tool to create presentations with L<sup>A</sup>T<sub>E</sub>X
- The environment `frame` defines slides
- Various themes are available to design the look of your presentation
- Separation between content and form

# L<sup>A</sup>T<sub>E</sub>X Beamer (Preamble)

```
1 \documentclass{beamer}
2
3 % File encoding
4 \usepackage[utf8]{inputenc}
5 % English language
6 \usepackage[english]{babel}
7 % Font
8 \usepackage[T1]{fontenc}
9 \usepackage{lmodern}
10
11 \title{LaTeX Example}
12 \author{Malte Luttermann}
13 \date{\today}
14
15 \usetheme{Luebeck}
```

# L<sup>A</sup>T<sub>E</sub>X Beamer (Content)

```
17 \begin{document}
18 \frame{\titlepage}
19
20 \begin{frame}
21   \begin{center}
22     My \textbf{first} \alert{presentation} using \LaTeX!
23   \end{center}
24 \end{frame}
25
26 \begin{frame}
27   \tableofcontents
28 \end{frame}
29
30 \section{Section 1}
31 \begin{frame}{Slide 1}
32   Lorem ipsum dolor sit amet, ... \pause
33   At vero eos et accusam et justo duo dolores et ea rebum.
34 \end{frame}
```

Section 1  
Section 2

Slide 1

Lorem ipsum dolor sit amet, ... At vero eos et accusam et justo duo dolores et ea rebum.

Malte Luttermann LaTeX Example

# L<sup>A</sup>T<sub>E</sub>X Beamer (Content Continued)

```
36 \section{Section 2}
37 \begin{frame}{Slide 2}
38   \begin{block}{Note}
39     A text...
40   \end{block}
41
42   \begin{alertblock}{Important}<2->
43     Another text...
44   \end{alertblock}
45 \end{frame}
46
47 \section{Section 3}
48 \end{document}
```

Section 1  
Section 2

Slide 2

Note  
A text...

Important  
Another text...

Malte Luttermann LaTeX Example

# L<sup>A</sup>T<sub>E</sub>X Beamer Themes

```
40 % ...  
41 \usetheme{CHAI}  
42 % ...
```

```
45 \begin{frame}[plain]  
46   \titlepage  
47 \end{frame}
```

```
1 \ProvidesPackage{beamerthemeCHAI}  
2 % ...
```

beamerthemeCHAI.sty

U+H Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CHAI  
Humanities-Centered AI

## Understanding Data vs. Machine Training

Malte Luttermann – Institute for Humanities-Centered Artificial Intelligence (CHAI)  
17.10.2025

Version from November 3, 2025

# L<sup>A</sup>T<sub>E</sub>X Beamer Themes

```
40 % ...  
41 \usetheme{CHAI}  
42 % ...
```

```
49 \begin{frame}[t]{Example Slide}  
50   \begin{itemize}  
51     \item First item  
52     \item Second item  
53       \begin{itemize}  
54         \item First subitem  
55       \end{itemize}  
56     \item Last item  
57   \end{itemize}  
58 \end{frame}
```

```
1 \ProvidesPackage{beamerthemeCHAI}  
2 % ...
```

beamerthemeCHAI.sty

U+H Universität Hamburg  
DER FORSCHUNG | DER LEHRE | DER BILDUNG

CHAI  
Humanities-Centered AI

## Example Slide

- First item
- Second item
  - First subitem
- Last item

Malte Luttermann      Understanding Data vs. Machine Training      2/3

# L<sup>A</sup>T<sub>E</sub>X Beamer Themes

```
40 % ...  
41 \usetheme{CHAI}  
42 % ...
```

```
60 \begin{frame}[t]{Two Columns}  
61   \begin{columns}[t]  
62     \begin{column}{0.55\textwidth}  
63       \begin{enumerate}  
64         \item First enum left  
65       \end{enumerate}  
66     \end{column}  
67     \begin{column}{0.40\textwidth}  
68       \begin{itemize}  
69         \item First item right  
70       \end{itemize}  
71     \end{column}  
72   \end{columns}  
73 \end{frame}
```

```
1 \ProvidesPackage{beamerthemeCHAI}  
2 % ...
```

beamerthemeCHAI.sty

Two Columns

(1) First enum left

■ First item right

Malte Luttermann Understanding Data vs. Machine Training 3/3

# Formulas in L<sup>A</sup>T<sub>E</sub>X

```

1 \begin{document}
2 Let  $x^2 + 5x - \alpha = 12$ .
3
4 We assume that the Gaussian sum\footnote{see, e.g., Wiki
5  $\sum_{i=1}^n i = \frac{n(n+1)}{2}$ }
6 is known.
7
8 The product can also be written in a compact form:
9  $y_1 \cdot \dots \cdot y_n = \prod_{i=1}^n y_i$ 
10
11 \begin{align*}
12 3x^2 &+ 4x &= 0 \\
13 2x^2 &+ 10x &= 0 \\
14 4x^2 &+ &= 0 \\
15 \end{align*}
16 \end{document}

```

Let  $x^2 + 5x - \alpha = 12$ .

We assume that the Gaussian sum<sup>3</sup>

$$\sum_{i=1}^n i = \frac{n(n+1)}{2}$$

is known.

The product can also be written in a compact form:

$$y_1 \cdots y_n = \prod_{i=1}^n y_i$$

$$\begin{array}{rcl}
 3x^2 + 4x & & = 0 \\
 2x^2 + 10x & & = 0 \\
 4x^2 + & & = 0
 \end{array}$$

<sup>3</sup>see, e.g., Wikipedia

## Formulas in $\text{\LaTeX}$

- The  $\text{\LaTeX}$  formula syntax is used in various places (e.g., Moodle, HedgeDoc)
- Inline math: Enclose with dollar signs  $\$ \dots \$$
- Displayed math (centered): Use  $\$ \$ \dots \$ \$$  or  $\backslash [ \dots \backslash ]$
- Aligned equations: Use  $\backslash \text{begin}\{\text{align}*\} \dots \backslash \text{end}\{\text{align}*\}$
- An overview of symbols, parentheses, etc. is given at <https://tug.ctan.org/info/undergradmath/undergradmath.pdf>





# Further Elements in L<sup>A</sup>T<sub>E</sub>X

```
1 \begin{enumerate}[i)]
2   \item Python
3   \item Java
4   \item \LaTeX
5     \begin{itemize}
6       \item Lists
7       \item in enumerations
8       \item are also possible
9       \item with \LaTeX.
10    \end{itemize}
11 \end{enumerate}
12
13 \begin{description}
14   \item[A term] that should be highlighted.
15 \end{description}
```

i) Python

ii) Java

iii) L<sup>A</sup>T<sub>E</sub>X

- Lists
- in enumerations
- are also possible
- with L<sup>A</sup>T<sub>E</sub>X.

**A term** that should be highlighted.

Package	Version	Number
Numpy	1	12
Scipy	1.7	200
Gensim	2.4	30

## Further Elements in L<sup>A</sup>T<sub>E</sub>X

```
1 \begin{tabular}{l|rr}  
2 \textbf{Package} & Version & Number \\ \hline  
3 Numpy & 1 & 12 \\  
4 Scipy & 1.7 & 200 \\  
5 Gensim & 2.4 & 30  
6 \end{tabular}  
7  
8 \includegraphics[width=6cm]{logo-chai.pdf}  
9 % \includegraphics[width=6cm]{logo-chai.png}  
10 % \includegraphics[width=6cm]{logo-chai.jpg}
```

i) Python

ii) Java

iii) L<sup>A</sup>T<sub>E</sub>X

- Lists
- in enumerations
- are also possible
- with L<sup>A</sup>T<sub>E</sub>X.

**A term** that should be highlighted.

Package	Version	Number
Numpy	1	12
Scipy	1.7	200
Gensim	2.4	30



# Floating Environments in $\text{\LaTeX}$

```

1 \begin{example}[Collinearity of Vector Representations]
2   Take a look at \ref{fig:aacp_example_vector_representation},
3   which shows the vector representations  $\vec{\phi}_1 = (8, 2)$ ,
4   ...
5 \end{example}
6
7 \begin{figure}
8   \centering
9   \input{example_vector_representation.tex}
10  \caption{A visualisation of the vector representations of
11  exemplary factors  $\phi_1$ ,  $\dots$ ,  $\phi_4$ . ...}
12  \label{fig:vector_representation}
13 \end{figure}
  
```

- The figure environment is a floating environment
- The figure is placed automatically by  $\text{\LaTeX}$  (placement may differ from its position in the .tex file)

Chapter 8 Lifted Model Construction without Normalization

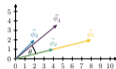


Figure 8.1: A visualisation of the vector representations of exemplary factors  $\phi_1, \dots, \phi_4$ . For the sake of this example, every factor has a potential table containing two potentials. The potential tables of the factors are encoded as vectors, which are given by  $\vec{\phi}_1 = (8, 2)$  (i.e.,  $\phi_1$  maps its first assignment to the potential 8 and the second assignment to the potential 2),  $\vec{\phi}_2 = (4, 1)$ ,  $\vec{\phi}_3 = (2, 2)$ , and  $\vec{\phi}_4 = (4.4, 3.6)$ .

**Example 8.3.2** (Collinearity of Vector Representations). Take a look at Figure 8.1, which shows the vector representations  $\vec{\phi}_1 = (8, 2)$ ,  $\vec{\phi}_2 = (4, 1)$ ,  $\vec{\phi}_3 = (2, 2)$ , and  $\vec{\phi}_4 = (4.4, 3.6)$  for exemplary factors  $\phi_1, \dots, \phi_4$ . To allow for a two-dimensional visualisation, every factor has a potential table containing two potentials (e.g., due to having a single Boolean argument). The angle between  $\vec{\phi}_1$  and  $\vec{\phi}_2$  is exactly zero, indicating that  $\phi_1$  and  $\phi_2$  are collinear and hence exchangeable, which can be verified as  $\phi_1(\mathbf{r}) = 2 \cdot \phi_2(\mathbf{r})$  holds for all assignments  $\mathbf{r}$ . At the same time, the angle between, e.g.,  $\phi_1$  and  $\phi_3$  is much larger than zero, indicating that  $\phi_1$  and  $\phi_3$  are not exchangeable. Moreover, the angle between  $\vec{\phi}_3$  and  $\vec{\phi}_4$  is not exactly zero but close to zero, indicating that  $\phi_3$  and  $\phi_4$  are not exchangeable but approximately equivalent (e.g.,  $\epsilon$ -equivalent for a sufficiently large  $\epsilon$ ).

By using vector representations and computing the cosine similarity between them, we are able to efficiently detect exchangeable factors independent of a scaling factor. The cosine similarity between two vector representations of factors lies within the interval  $[0, 1]$  (because potentials are always positive numbers) and reaches its maximum value of one if the angle between the vectors is zero. To obtain a distance measure, we define the cosine distance as one minus the cosine similarity.

**Definition 8.3.2** (Cosine Distance [Gehrke et al., 2020]). Let  $\phi_1(R_1, \dots, R_n)$  and  $\phi_2(R'_1, \dots, R'_n)$  denote two factors. The cosine distance between  $\phi_1$  and  $\phi_2$  is defined as

$$D_{\cos}(\phi_1, \phi_2) = 1 - \frac{\sum_{\mathbf{r} \in \Sigma_{\text{range}(R_1)} \times \dots \times \Sigma_{\text{range}(R_n)}} \phi_1(\mathbf{r}) \cdot \phi_2(\mathbf{r})}{\sqrt{\sum_{\mathbf{r} \in \Sigma_{\text{range}(R_1)} \times \dots \times \Sigma_{\text{range}(R_n)}} \phi_1(\mathbf{r})^2} \cdot \sqrt{\sum_{\mathbf{r} \in \Sigma_{\text{range}(R'_1)} \times \dots \times \Sigma_{\text{range}(R'_n)}} \phi_2(\mathbf{r})^2}} \quad (8.3)$$

In case  $\phi_1$  and  $\phi_2$  are defined over different function domains, we define  $D_{\cos}(\phi_1, \phi_2) = \infty$ .

**Example 8.3.3** (Cosine Distance). Consider again the factors  $\phi_1$  and  $\phi_2$  with corresponding vector representations  $\vec{\phi}_1 = (8, 2)$  and  $\vec{\phi}_2 = (4, 1)$ , respectively, from Figure 8.1.

# Floating Environments in L<sup>A</sup>T<sub>E</sub>X

```

1 \begin{example}[Collinearity of Vector Representations]
2   Take a look at \ref{fig:aacp_example_vector_representation},
3   which shows the vector representations  $\vec{\phi}_1 = (8, 2)$ ,
4   ...
5 \end{example}
6
7 \begin{figure}
8   \centering
9   \input{example_vector_representation.tex}
10  \caption{A visualisation of the vector representations of
11  exemplary factors  $\phi_1$ , \ldots, \phi_4. ...}
12  \label{fig:vector_representation}
13 \end{figure}

```

- `\label{...}` assigns a label to the figure
- `\ref{<label-name>}` refers to the figure and inserts the correct figure number

Chapter 8 Lifted Model Construction without Normalization

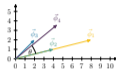


Figure 8.1: A visualisation of the vector representations of exemplary factors  $\phi_1, \dots, \phi_4$ . For the sake of this example, every factor has a potential table containing two potentials. The potential tables of the factors are encoded as vectors, which are given by  $\phi_1 = (8, 2)$  (i.e.,  $\phi_1$  maps its first assignment to the potential 8 and the second assignment to the potential 2),  $\phi_2 = (4, 1)$ ,  $\phi_3 = (2, 2)$ , and  $\phi_4 = (4.4, 3.6)$ .

**Example 8.3.2** (Collinearity of Vector Representations). Take a look at Figure 8.1, which shows the vector representations  $\phi_1 = (8, 2)$ ,  $\phi_2 = (4, 1)$ ,  $\phi_3 = (2, 2)$ , and  $\phi_4 = (4.4, 3.6)$  for exemplary factors  $\phi_1, \dots, \phi_4$ . To allow for a two-dimensional visualisation, every factor has a potential table containing two potentials (e.g., due to having a single Boolean argument). The angle between  $\phi_1$  and  $\phi_2$  is exactly zero, indicating that  $\phi_1$  and  $\phi_2$  are collinear and hence exchangeable, which can be verified as  $\phi_1(r) = 2 \cdot \phi_2(r)$  holds for all assignments  $r$ . At the same time, the angle between, e.g.,  $\phi_1$  and  $\phi_3$  is much larger than zero, indicating that  $\phi_1$  and  $\phi_3$  are not exchangeable. Moreover, the angle between  $\phi_3$  and  $\phi_4$  is not exactly zero but close to zero, indicating that  $\phi_3$  and  $\phi_4$  are not exchangeable but approximately equivalent (e.g.,  $\epsilon$ -equivalent for a sufficiently large  $\epsilon$ ).

By using vector representations and computing the cosine similarity between them, we are able to efficiently detect exchangeable factors independent of a scaling factor. The cosine similarity between two vector representations of factors lies within the interval  $[0, 1]$  (because potentials are always positive numbers) and reaches its maximum value of one if the angle between the vectors is zero. To obtain a distance measure, we define the cosine distance as one minus the cosine similarity.

**Definition 8.3.2** (Cosine Distance [Gehrke et al., 2020]). Let  $\phi_1(R_1, \dots, R_n)$  and  $\phi_2(R'_1, \dots, R'_n)$  denote two factors. The cosine distance between  $\phi_1$  and  $\phi_2$  is defined as

$$D_{\cos}(\phi_1, \phi_2) = 1 - \frac{\sum_{r \in \Sigma_{R_1} \times \dots \times \Sigma_{R_n}} \phi_1(r) \cdot \phi_2(r)}{\sqrt{\sum_{r \in \Sigma_{R_1} \times \dots \times \Sigma_{R_n}} \phi_1(r)^2} \cdot \sqrt{\sum_{r \in \Sigma_{R'_1} \times \dots \times \Sigma_{R'_n}} \phi_2(r)^2}} \quad (8.3)$$

In case  $\phi_1$  and  $\phi_2$  are defined over different function domains, we define  $D_{\cos}(\phi_1, \phi_2) = \infty$ .

**Example 8.3.3** (Cosine Distance). Consider again the factors  $\phi_1$  and  $\phi_2$  with corresponding vector representations  $\phi_1 = (8, 2)$  and  $\phi_2 = (4, 1)$ , respectively, from Figure 8.1.

# BibTeX

- BibTeX allows to create bibliographic references automatically in  $\text{\LaTeX}$
- Example BibTeX entry:

```
1 @article{Ahmadi2013a,  
2   author    = {Babak Ahmadi and Kristian Kersting and Martin Mladenov  
3     and Sriraam Natarajan},  
4   title     = {{Exploiting Symmetries for Scaling Loopy Belief  
5     Propagation and Relational Training}},  
6   journal   = {Machine Learning},  
7   volume    = {92},  
8   year      = {2013},  
9   pages     = {91--132},  
10  publisher = {Springer},  
11 }
```

# BibTeX

## ■ Usage in $\text{\LaTeX}$ document:

```
1 \begin{document}
2 The CompressFactorGraph algorithm \cite{Ahmadi2013a} solves the
   problem of constructing a lifted representation entailing
   equivalent semantics as a given input factor graph.
3
4 ...
5 \end{document}
```

# BibTeX

- Resulting bibliography in the document:

## Bibliography

Babak Alnadi, Kristian Kersting, Martin Mladenc, and Sriram Natarajan. Exploiting Symmetries for Scaling Loopy Belief Propagation and Relational Training. *Machine Learning*, 92:91–132, 2013.

Ragib Alsoan, David Arbour, and Elena Zheleva. Relational Causal Models with Cyclics: Representation and Reasoning. In *Proceedings of the First Conference on Causal Learning and Reasoning (CLaR-2022)*, pages 1–18. PMLR, 2022.

Ragib Alsoan, David Arbour, and Elena Zheleva. Learning Relational Causal Models with Cycles through Relational Acycclification. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-2023)*, pages 12164–12171. AAAI Press, 2023.

Steen A. Anderson, David Madigan, and Michael D. Perlman. A Characterization of Markov Equivalence Classes for Acyclic Digraphs. *The Annals of Statistics*, 25:505–541, 1997.

Udi Apsel and Ronen I. Brafman. Lifted MEU by Weighted Model Counting. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-2012)*, pages 1861–1867. AAAI Press, 2012.

David Arbour, Dan Garas, and David Jensen. Inferring Network Effects from Observational Data. In *Proceedings of the Twenty-Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2016)*, pages 715–724. ACM Press, 2016.

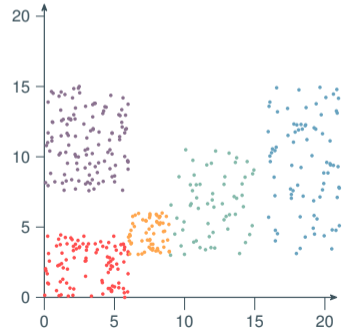
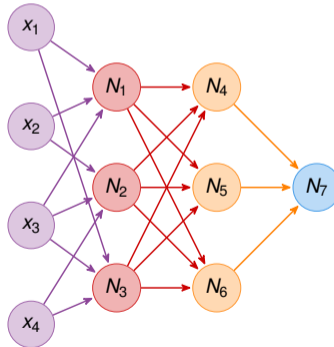
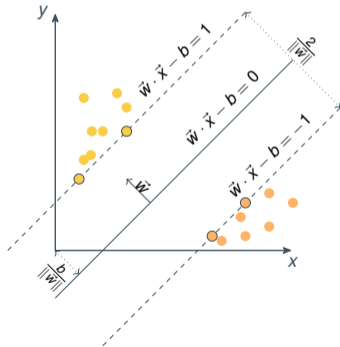
Sheldon Axler. *Linear Algebra Done Right*. Springer, 3rd edition, 2015.

Tanya Braun and Marcel Gebcke. Explainable and Explorable Decision Support. In *Proceedings of the Twenty-Seventh International Conference on Conceptual Structures (ICCS-2022)*, pages 99–114. Springer, 2022.

Tanya Braun and Ralf Möller. Lifted Junction Tree Algorithm. In *Proceedings of the Thirty-Ninth German Conference on Artificial Intelligence (KI-2016)*, pages 30–42. Springer, 2016.

# TikZ

- »TikZ ist kein Zeichenprogramm«
- TikZ allows to create graphics, figures, diagrams, etc. directly in  $\text{\LaTeX}$



# Overleaf



- Collaborative cloud-based  $\text{\LaTeX}$  editor
  - <https://www.overleaf.com>
- Collaborate on  $\text{\LaTeX}$  documents in real-time
- Use  $\text{\LaTeX}$  without an installation on your local machine

# Overleaf

Menu Upgrade

Code Editor Visual Editor

Recompile

2 / 19

IJCAI2025\_Approximate Lifted Model Construction

```

132 \section{Background} \label{sec:eacp_background}
133
134 We first define  $\text{vACP}(fg)$  as propositional models and afterwards introduce the idea of lifted representations such as  $\text{vACP}(pfg)$ .
135 An  $\text{vACP}(fg)$  is a probabilistic graphical model to compactly represent a probability distribution over a set of  $\text{vACP}(rv)$  by factorising the distribution  $\text{vActep}$ [Frey1997a,Kschischang2001a].
136 \begin{definition}[Factor Graph]
137 An  $\text{vgraph}(\text{vACP}(fg))$   $\text{SM} = (\text{vboldsymbol } V, \text{vboldsymbol } E)$  is an undirected bipartite graph consisting of a node set  $\text{vboldsymbol } V = \text{vboldsymbol } R \cup \text{vboldsymbol } S$ , where  $\text{vboldsymbol } R = \text{v}(R_1, \dots, R_n)$  is a set of variable nodes ( $\text{vACP}(rv)$ ) and  $\text{vboldsymbol } S = \text{v}(S_1, \dots, S_m)$  is a set of factor nodes (functions), as well as a set of edges  $\text{vboldsymbol } E \subseteq \text{vsubsetsq } \text{vboldsymbol } R \times \text{vboldsymbol } S$ .
138 There is an edge between a variable node  $\text{v}R_i$  and a factor node  $\text{v}S_j$  if and only if  $\text{v}R_i$  appears in the argument list of  $\text{v}S_j$ .
139 A factor  $\text{v}S_j = \text{v}(\text{vboldsymbol } R_j)$  defines a function  $\text{v}S_j : \text{vboldsymbol } R_j \rightarrow \text{v}(0, 1)$ . The range of  $\text{v}S_j$  is the set of possible values  $\text{v}R_j$  can take.
140 The term  $\text{vrange}(R_j)$  denotes the possible values  $\text{v}R_j$  can take.
141 We define the joint potential for an assignment  $\text{vboldsymbol } r$  (where  $\text{vboldsymbol } r$  is a shorthand notation for  $\text{vboldsymbol } R = \text{vboldsymbol } r$ ) as
142 \begin{align}
\psi(\text{vboldsymbol } r) &= \prod_{\text{v}S_j \in \text{vboldsymbol } S} \text{v}S_j(\text{vboldsymbol } r_j)
\end{align}
143 where  $\text{vboldsymbol } r_j$  is a projection of  $\text{v}r$  to the argument list of  $\text{v}S_j$ .
144 The full joint probability distribution encoded by  $\text{vSM}$  is then given by the normalised joint potential
145 \begin{align}
P(\text{vboldsymbol } r) &= \frac{\psi(\text{vboldsymbol } r)}{\sum_{\text{v}r' \in \text{vboldsymbol } R} \psi(\text{v}r')}
\end{align}
146
147 \begin{align}
P(\text{vboldsymbol } r) &= \frac{\psi(\text{vboldsymbol } r)}{\sum_{\text{v}r' \in \text{vboldsymbol } R} \psi(\text{v}r')}
\end{align}
148

```

Files

- files
- defs.tex
- ijcai25.sty
- main.tex
- named.bst
- notes.tex
- references.bib

File outline

- Introduction
- Background
- Approximation of Indistinguishable...
- Finding and Grouping e-Equal...
- The e-Advanced Colour Passi...
- Bounding the Change in Query R...
- Experiments
- Conclusion
- Missing Proofs
- Full derivation of writing 'cref...
- Full derivation of writing 'cref...
- The Advanced Colour Passing Alg...
- Permutations of Factors' Argume...
- Approximate Symmetries Within ...
- Further Experimental Results

parameter  $\epsilon$  controls the trade-off between the exactness of the approximation and the size of the lifted representation. We prove that the approximation error induced by  $\epsilon$ -ACP is strictly bounded. In addition to the theoretical results, we empirically show that  $\epsilon$ -ACP significantly reduces run times for inference while at the same time keeping the approximation error close to zero.

The remaining part of this paper is structured as follows. We begin by introducing background information and notations in Sec. 2. Thereafter, we introduce the  $\epsilon$ -ACP algorithm to solve the problem of constructing an approximate lifted representation with a targeted approximation error. We then prove that the approximation error induced by  $\epsilon$ -ACP is strictly bounded and show that the given bound is optimal. Finally, we empirically demonstrate that in practice, the actual approximation error induced by  $\epsilon$ -ACP is well below the theoretical bounds before we conclude the paper.

## 2 Background

We first define factor graphs (FGs) as propositional models and afterwards introduce the idea of lifted representations such as PFGs. An FG is a probabilistic graphical model to compactly represent a probability distribution over a set of random variables (randomly by factorising the distribution [Frey 1997, Kschischang et al. 2001]).

**Definition 1 (Factor Graph).** An FG  $\text{FG} = (V, E)$  is an undirected bipartite graph consisting of a node set  $V = \{v_1, \dots, v_n\}$  and a set of edges  $E \subseteq \{v_1, \dots, v_n\} \times \{v_1, \dots, v_n\}$ . There is an edge between a variable node  $v_i$  and a factor node  $v_j$  if and only if  $v_i$  appears in the argument list of  $v_j$ . A factor node  $v_j$  defines a function  $v_j : \text{vrange}(R_j) \rightarrow \mathbb{R}^+$ , that maps the range of its arguments  $R_j$  to a positive real number from  $\mathbb{R}^+$  to  $\mathbb{R}^+$ . The range of  $v_j$  denotes the possible values  $R_j$  can take. We define the joint potential for an assignment  $r = (r_1, \dots, r_n)$  as

$$\psi(r) = \prod_{v_j \in V} v_j(r_j) \quad (1)$$

where  $r_j$  is a projection of  $r$  to the argument list of  $v_j$ . The full joint probability distribution encoded by  $\text{FG}$  is then given by the normalised joint potential

$$P(r) = \frac{\psi(r)}{\sum_{r' \in \text{vrange}(V)} \psi(r')} \quad (2)$$

where  $\text{vrange}(V) = \prod_{v_i \in V} \text{vrange}(v_i)$  is the normalisation constant.

**Example 1.** Consider the FG depicted in Fig. 1, which models the dependencies between the revenue  $\text{Rev}$  of a company and the salary of two employees, denoted as  $\text{SalA}$  and  $\text{SalB}$ . We have  $\text{Rev} = \text{v}(\text{high}, \text{low})$ ,  $\text{SalA} = \text{v}(\text{high}, \text{low})$ , and  $\text{SalB} = \text{v}(\text{high}, \text{low})$ . For the sake of the example, let  $\text{range}(\text{SalA}) = \text{range}(\text{SalB}) = \text{range}(\text{Rev}) = \{\text{high}, \text{low}\}$ . The potential values of  $\psi$

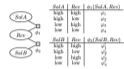


Figure 1: An FG modeling the dependency between the revenue of a company (Rev) and the salaries of two employees (SalA, SalB). The potential values of the factors are shown on the right.

and  $\psi_2$  are shown on the right. In particular, it holds that  $\psi_1(\text{high}, \text{high}) = \psi_1(\text{high}, \text{low}) = \psi_2$ , and so on, where  $\psi_1, \psi_2 \in \mathbb{R}^+$  are arbitrary positive real numbers.

Probabilistic inference describes the task of computing marginal distributions of random graph observations for other reasoning. In other words, probabilistic inference refers to query answering, where a query is defined as follows.

**Definition 2 (Query).** A query  $Q = (v_1, \dots, v_k, c_1, \dots, c_k)$  consists of a query atom  $Q$  and a set of events  $\{c_1 = \psi_1, \dots, c_k = \psi_k\}$  (called evidence), where  $Q$  and  $\psi_i$  are random variables of a specific probability instead of a probability distribution. The query atom is an event  $Q = \psi$ .

**Example 2.** Take a look at the FG shown in Fig. 1. The query  $Q(\text{Rev} = \text{high})$  asks for the probability distribution of  $\text{Rev}$  if the salary nodes of the company have a high revenue.

When considering relations between objects, there are often groups of indistinguishable objects that behave identically or at least similarly. Lifted representations such as PFGs exploit identical behavior to enable scalable probabilistic inference with respect to domain sizes of objects. To illustrate this idea behind lifting, consider the following example.

**Example 3.** Consider the FG depicted in Fig. 1 and the query  $Q(\text{Rev} = \text{high})$ . Then, it holds that

$$\begin{aligned}
 P(\text{Rev} = \text{high}) &= \sum_{\text{vrange}(\text{SalA}) \times \text{vrange}(\text{SalB})} P(\text{Rev} = \text{high}) \\
 &= \frac{1}{2} \sum_{\text{vrange}(\text{SalA}) \times \text{vrange}(\text{SalB})} \psi_1(\text{high}, \psi_2(\text{high}, \text{high})) \\
 &= \frac{1}{2} (\psi_1 \psi_2 + \psi_1 \psi_2 + \psi_1 \psi_2 + \psi_1 \psi_2)
 \end{aligned}$$

Employers A and B are indistinguishable, that is, if  $\psi_1$  holds, so does  $\psi_2$ , for all  $i \in \{1, \dots, n\}$ , we can simplify the computation and obtain

$$\begin{aligned}
 P(\text{Rev} = \text{high}) &= \sum_{\text{vrange}(\text{SalA})} \psi_1(\text{high}, \text{high}) \sum_{\text{vrange}(\text{SalB})} \psi_1(\text{high}, \text{high}) \\
 &= \sum_{\text{vrange}(\text{SalA})} \psi_1(\text{high}, \text{high})^2
 \end{aligned}$$

$\text{Rev}$	$R_1$	$\psi_1(R_1, R_1)$	$\psi_1(R_1, R_1)$	$\psi_1(R_1, R_1)$	$\psi_1(R_1, R_1)$
high	high	0.75	0.6	0.45	0.84
high	low	0.33	0.3	0.33	0.31
low	high	0.45	0.3	0.33	0.31
low	low	0.22	0.3	0.33	0.31

Example 3 illustrates that in case A and B are indistinguishable, we can select one representative (i.e., A) and reduce the number of factors to consider for computation.

The idea of exploiting representation can be generalized to

$$\begin{aligned}
 P(\text{Rev} = \text{high}) &= \sum_{\text{vrange}(\text{SalA})} \psi_1(\text{high}, \text{high}) \sum_{\text{vrange}(\text{SalB})} \psi_1(\text{high}, \text{high}) \\
 &= \sum_{\text{vrange}(\text{SalA})} \psi_1(\text{high}, \text{high})^2
 \end{aligned}$$

Example 3 illustrates that in case A and B are indistinguishable, we can select one representative (i.e., A) and reduce the number of factors to consider for computation.

The idea of exploiting representation can be generalized to



## Summary

- Markup languages
- Content, structure, and form
- Semantic markup
- Markdown
- L<sup>A</sup>T<sub>E</sub>X



The L<sup>A</sup>T<sub>E</sub>X Project

Overleaf