CIS 530 Advanced Data Mining

Course Project Instructions

Shuwen Wang, Computer & Info Science Department

Project Choice 1 - Overview (100 points)

- Team: 1 to 4 people
  - We will have a really high expectation to teams of 4 people → Nearly a top conference submission quality
- Presentation: In-person on April 23, April 28, April 30 (if needed)
  - No presentation, No credit
  - Each team will have 10 minutes to present
- Report Due **11:59pm May 5** No report, No credit
  - At least 4 pages
    - Double-column, 11 pt
    - Roughly 2.5-3 thousand words + figures, tables, and equations
- Code Due **11:59pm May 5** No code, No credit
  - A GitHub repo
  - Or A PDF

Report Template

- ACM Proceedings Templates
  - https://www.acm.org/publications/proceedings-template
  - Overleaf: https://www.overleaf.com/latex/templates/acm-conference-proceedings-master-template/pnrfvrrdbfwt

The Name of the Title is Hope

Ben Trovato*
G. K. M. Tobin *
trovato@corporation.com
webmaster@marysville-ohio.com
Institute for Clarity in Documentation
Dublin, Ohio

Lars Thørväld
The Thørväld Group
Hekla, Iceland
larst@affiliation.org

Valerie Béranger
Inria Paris-Rocquencourt
Rocquencourt, France

Aparna Patel
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India

Huifen Chan
Tsinghua University
Haidian Qu, Beijing Shi, China

Charles Palmer
Palmer Research Laboratories
San Antonio, Texas
cpalmer@prl.com

John Smith
The Thørväld Group
jsmith@affiliation.org

Julius P. Kumquat
The Kumquat Consortium
jpkumquat@consortium.net

Figure 1: Seattle Mariners at Spring Training, 2010

ABSTRACT

A clear and well-documented LATEX document is presented as an article formatted for publication by ACM in a conference proceedings or journal publication. Based on the "acmart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

CCS CONCEPTS

• **Computer systems organization** ! **Embedded systems**; *Redundancy*; Robotics; • **Networks** ! Network reliability.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

ACM Reference Format:
Ben Trovato, G.K.M. Tobin, Lars Thørväld, Valerie Béranger, Aparna Patel, Huifen Chan, Charles Palmer, John Smith, and Julius P. Kumquat. 2018 The Name of the Title is Hope. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/1122445.1122456

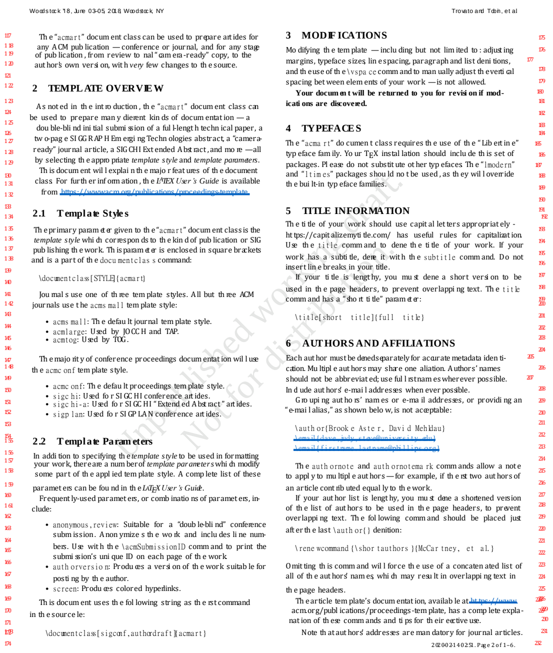
1 INTRODUCTION

ACM's consolidated article template, introduced in 2017, provides a consistent LATEX style for use across ACM publications, and incorporates accessibility and metadata-extraction functionality necessary for future Digital Library endeavors. Numerous ACM and SIG-specific LATEX templates have been examined, and their unique features incorporated into this single new template.

If you are new to publishing with ACM, this document is a valuable guide to the process of preparing your work for publication. If you have published with ACM before, this document provides insight and instruction into more recent changes to the article template.

[*] Both authors contributed equally to this research.

2020-02-14 02:51. Page 1 of 1–6.



Woodstock '18, June 03-05, 2018, Woodstock, NY — Trovato and Tobin, et al.

The "acmart" document class can be used to prepare articles for any ACM publication — conference or journal, and for any stage of publication, from review to final "camera-ready" copy, to the author's own version, with *very* few changes to the source.

2 TEMPLATE OVERVIEW

As noted in the introduction, the "acmart" document class can be used to prepare many different kinds of documentation — a double-blind initial submission of a full-length technical paper, a two-page SIGGRAPH Emerging Technologies abstract, a "camera-ready" journal article, a SIGCHI Extended Abstract, and more — all by selecting the appropriate *template style* and *template parameters*.

This document will explain the major features of the document class. For further information, the *LATEX User's Guide* is available from https://www.acm.org/publications/proceedings-template.

2.1 Template Styles

The primary parameter given to the "acmart" document class is the *template style* which corresponds to the kind of publication or SIG publishing the work. This parameter is enclosed in square brackets and is a part of the documentclass command:

```
\documentclass[STYLE]{acmart}
```

Journals use one of three template styles. All but three ACM journals use the acmsmall template style:

- acmsmall: The default journal template style.
- acmlarge: Used by JOCCH and TAP.
- acmtog: Used by TOG.

The majority of conference proceedings documentation will use the acmconf template style.

- acmconf: The default proceedings template style.
- sigchi: Used for SIGCHI conference articles.
- sigchi-a: Used for SIGCHI "Extended Abstract" articles.
- sigplan: Used for SIGPLAN conference articles.

2.2 Template Parameters

In addition to specifying the *template style* to be used in formatting your work, there are a number of *template parameters* which modify some part of the applied template style. A complete list of these parameters can be found in the *LATEX User's Guide.*

Frequently-used parameters, or combinations of parameters, include:

- anonymous,review: Suitable for a "double-blind" conference submission. Anonymizes the work and includes line numbers. Use with the \acmSubmissionID command to print the submission's unique ID on each page of the work.
- authorversion: Produces a version of the work suitable for posting by the author.
- screen: Produces colored hyperlinks.

This document uses the following string as the first command in the source file:

```
\documentclass[sigconf,authordraft]{acmart}
```

3 MODIFICATIONS

Modifying the template — including but not limited to: adjusting margins, typeface sizes, line spacing, paragraph and list definitions, and the use of the \vspace command to manually adjust the vertical spacing between elements of your work — is not allowed.

Your document will be returned to you for revision if modifications are discovered.

4 TYPEFACES

The "acmart" document class requires the use of the "Libertine" typeface family. Your TEX installation should include this set of packages. Please do not substitute other typefaces. The "lmodern" and "ltimes" packages should not be used, as they will override the built-in typeface families.

5 TITLE INFORMATION

The title of your work should use capital letters appropriately - https://capitalizemytitle.com/ has useful rules for capitalization. Use the title command to define the title of your work. If your work has a subtitle, define it with the subtitle command. Do not insert line breaks in your title.

If your title is lengthy, you must define a short version to be used in the page headers, to prevent overlapping text. The title command has a "short title" parameter:

```
\title[short title]{full title}
```

6 AUTHORS AND AFFILIATIONS

Each author must be defined separately for accurate metadata identification. Multiple authors may share one affiliation. Authors' names should not be abbreviated; use full first names wherever possible. Include authors' e-mail addresses whenever possible.

Grouping authors' names or e-mail addresses, or providing an "e-mail alias," as shown below, is not acceptable:

```
\author{Brooke Aster, David Mehldau}
\email{dave,judy,steve@university.edu}
\email{firstname.lastname@phillips.org}
```

The authornote and authornotemark commands allow a note to apply to multiple authors — for example, if the first two authors of an article contributed equally to the work.

If your author list is lengthy, you must define a shortened version of the list of authors to be used in the page headers, to prevent overlapping text. The following command should be placed just after the last \author{} definition:

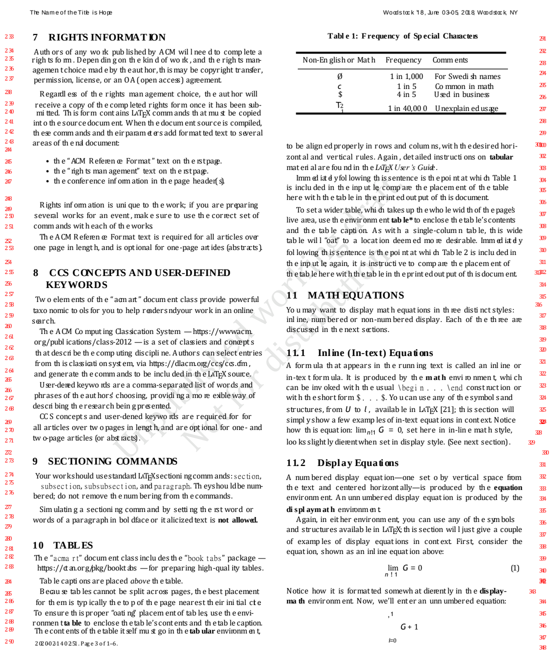
```
\renewcommand{\shortauthors}{McCartney, et al.}
```

Omitting this command will force the use of a concatenated list of all of the authors' names, which may result in overlapping text in the page headers.

The article template's documentation, available at https://www.acm.org/publications/proceedings-template, has a complete explanation of these commands and tips for their effective use.

Note that authors' addresses are mandatory for journal articles.

2020-02-14 02:51. Page 2 of 1–6.



The Name of the Title is Hope — Woodstock '18, June 03-05, 2018, Woodstock, NY

7 RIGHTS INFORMATION

Authors of any work published by ACM will need to complete a rights form. Depending on the kind of work, and the rights management choice made by the author, this may be copyright transfer, permission, license, or an OA (open access) agreement.

Regardless of the rights management choice, the author will receive a copy of the completed rights form once it has been submitted. This form contains LATEX commands that must be copied into the source document. When the document source is compiled, these commands and their parameters add formatted text to several areas of the final document:

- the "ACM Reference Format" text on the first page.
- the "rights management" text on the first page.
- the conference information in the page header(s).

Rights information is unique to the work; if you are preparing several works for an event, make sure to use the correct set of commands with each of the works.

The ACM Reference Format text is required for all articles over one page in length, and is optional for one-page articles (abstracts).

8 CCS CONCEPTS AND USER-DEFINED KEYWORDS

Two elements of the "acmart" document class provide powerful taxonomic tools for you to help readers find your work in an online search.

The ACM Computing Classification System — https://www.acm.org/publications/class-2012 — is a set of classifiers and concepts that describe the computing discipline. Authors can select entries from this classification system, via https://dl.acm.org/ccs/ccs.cfm, and generate the commands to be included in the LATEX source.

User-defined keywords are a comma-separated list of words and phrases of the authors' choosing, providing a more flexible way of describing the research being presented.

CCS concepts and user-defined keywords are required for for all articles over two pages in length, and are optional for one- and two-page articles (or abstracts).

9 SECTIONING COMMANDS

Your work should use standard LATEX sectioning commands: section, subsection, subsubsection, and paragraph. They should be numbered; do not remove the numbering from the commands.

Simulating a sectioning command by setting the first word or words of a paragraph in boldface or italicized text is **not allowed.**

10 TABLES

The "acmart" document class includes the "booktabs" package — https://ctan.org/pkg/booktabs — for preparing high-quality tables.

Table captions are placed *above* the table.

Because tables cannot be split across pages, the best placement for them is typically the top of the page nearest their initial cite. To ensure this proper "floating" placement of tables, use the environment **table** to enclose the table's contents and the table caption. The contents of the table itself must go in the **tabular** environment, to be aligned properly in rows and columns, with the desired horizontal and vertical rules. Again, detailed instructions on **tabular** material are found in the *LATEX User's Guide*.

Table 1: Frequency of Special Characters

Non-English or Math	Frequency	Comments
Ø	1 in 1,000	For Swedish names
c	1 in 5	Common in math
$	4 in 5	Used in business
T_1^2	1 in 40,000	Unexplained usage

Immediately following this sentence is the point at which Table 1 is included in the input file; compare the placement of the table here with the table in the printed output of this document.

To set a wider table, which takes up the whole width of the page's live area, use the environment **table*** to enclose the table's contents and the table caption. As with a single-column table, this wide table will "float" to a location deemed more desirable. Immediately following this sentence is the point at which Table 2 is included in the input file; again, it is instructive to compare the placement of the table here with the table in the printed output of this document.

11 MATH EQUATIONS

You may want to display math equations in three distinct styles: inline, numbered or non-numbered display. Each of the three are discussed in the next sections.

11.1 Inline (In-text) Equations

A formula that appears in the running text is called an inline or in-text formula. It is produced by the **math** environment, which can be invoked with the usual \begin . . . \end construction or with the short form $ . . . $. You can use any of the symbols and structures, from U to l, available in LATEX [21]; this section will simply show a few examples of in-text equations in context. Notice how this equation: $\lim_{n!1} G = 0$, set here in in-line math style, looks slightly different when set in display style. (See next section).

11.2 Display Equations

A numbered display equation—one set off by vertical space from the text and centered horizontally—is produced by the **equation** environment. An unnumbered display equation is produced by the **displaymath** environment.

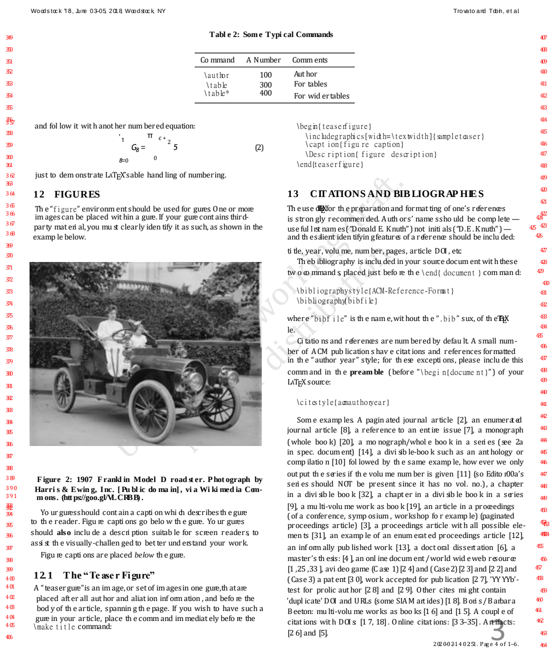
Again, in either environment, you can use any of the symbols and structures available in LATEX; this section will just give a couple of examples of display equations in context. First, consider the equation, shown as an inline equation above:

$$\lim_{n!1} G = 0 \tag{1}$$

Notice how it is formatted somewhat differently in the **displaymath** environment. Now, we'll enter an unnumbered equation:

$$\sum_{i=0}^{1} G + 1$$

2020-02-14 02:51. Page 3 of 1–6.



Woodstock '18, June 03-05, 2018, Woodstock, NY — Trovato and Tobin, et al.

Table 2: Some Typical Commands

Command	A Number	Comments
\author	100	Author
\table	300	For tables
\table*	400	For wider tables

and follow it with another numbered equation:

$$\sum_{8=0}^{1} G_8 = \int_{0}^{\pi c+2} 5 \tag{2}$$

just to demonstrate LATEX's able handling of numbering.

12 FIGURES

The "figure" environment should be used for figures. One or more images can be placed within a figure. If your figure contains third-party material, you must clearly identify it as such, as shown in the example below.

Figure 2: 1907 Franklin Model D roadster. Photograph by Harris & Ewing, Inc. [Public domain], via Wikimedia Commons. (https://goo.gl/VLCRBB).

Your figures should contain a caption which describes the figure to the reader. Figure captions go below the figure. Your figures should **also** include a description suitable for screen readers, to assist the visually-challenged to better understand your work.

Figure captions are placed *below* the figure.

12.1 The "Teaser Figure"

A "teaser figure" is an image, or set of images in one figure, that are placed after all author and affiliation information, and before the body of the article, spanning the page. If you wish to have such a figure in your article, place the command immediately before the \maketitle command:

```
\begin{teaserfigure}
  \includegraphics[width=\textwidth]{sampleteaser}
  \caption{figure caption}
  \Description{figure description}
\end{teaserfigure}
```

13 CITATIONS AND BIBLIOGRAPHIES

The use of BIBTEX for the preparation and formatting of one's references is strongly recommended. Authors' names should be complete — use full first names ("Donald E. Knuth") not initials ("D. E. Knuth") — and the salient identifying features of a reference should be included: title, year, volume, number, pages, article DOI, etc.

The bibliography is included in your source document with these two commands, placed just before the \end{document} command:

```
\bibliographystyle{ACM-Reference-Format}
\bibliography{bibfile}
```

where "bibfile" is the name, without the ".bib" suffix, of the BIBTEX file.

Citations and references are numbered by default. A small number of ACM publications have citations and references formatted in the "author year" style; for these exceptions, please include this command in the **preamble** (before "\begin{document}") of your LATEX source:

```
\citestyle{acmauthoryear}
```

Some examples. A paginated journal article [2], an enumerated journal article [8], a reference to an entire issue [7], a monograph (whole book) [20], a monograph/whole book in a series (see 2a in spec. document) [14], a divisible-book such as an anthology or compilation [10] followed by the same example, however we only output the series if the volume number is given [11] (so Editor00a's series should NOT be present since it has no vol. no.), a chapter in a divisible book [32], a chapter in a divisible book in a series [9], a multi-volume work as book [19], an article in a proceedings (of a conference, symposium, workshop for example) (paginated proceedings article) [3], a proceedings article with all possible elements [31], an example of an enumerated proceedings article [12], an informally published work [13], a doctoral dissertation [6], a master's thesis: [4], an online document / world wide web resource [1, 25, 33], a video game (Case 1) [24] and (Case 2) [23] and [22] and (Case 3) a patent [30], work accepted for publication [27], 'YYYYb'-test for prolific author [28] and [29]. Other cites might contain 'duplicate' DOI and URLs (some SIAM articles) [18]. Boris / Barbara Beeton: multi-volume works as books [16] and [15]. A couple of citations with DOIs: [17, 18]. Online citations: [33–35]. Artifacts: [26] and [5].

2020-02-14 02:51. Page 4 of 1–6.

Five Components

- ❑ Dataset
- ❑ Predictive Task
- ❑ Model
- ❑ Literature
- ❑ Results

Dataset

- ❑ Identify a dataset
- ❑ Perform an exploratory data analysis
  - ❑ Basic Statistics
  - ❑ Properties
  - ❑ Interesting findings
- ❑ All these should motivate your model design/choice

- ❑ The dataset should be large enough (at least 10,000 instances in total)

Dataset - Example

- Grammy Winners and Nominees from 1965 to 2024
  - More than 25,000 records

- This dataset compiles historical Grammy Award winners and nominees, including details such as the award category, artist, song or album, and year of recognition. The data was sourced from grammy.com and can be used for music trend analysis, industry research, and historical insights.

- EDA
  - Most awarded categories?
  - Top Grammy Award winners?
  - …

Predictive Task

- ❑ Identify a predictive task based on your dataset
- ❑ How will you evaluate different models in this task?
- ❑ What are the baseline models you want to compare with?
  - ❑ Why do you think they are appropriate?
  - ❑ Why do you think your model can outperform them? Or say, what are their drawbacks?

Predictive Task - Example

- ❑ Grammy Winners and Nominees from 1965 to 2024
  - ❑ More than 25,000 records

- ❑ Whether a nomination is a winner or not
  - ❑ A classification problem!

Predictive Task - Example

- Whether a nomination is a winner or not

- Evaluation
  - Accuracy, F1, AUC

- Baseline
  - Logistic regression: Assume it’s a linear combination of selected features

Model

- ❑ What is the model that you propose to attack this task?
  - ❑ It’s fine to use models that were described in class here
  - ❑ Explain and justify your choice/proposal What are the features you designed for your model?
  - ❑ Any unsuccessful tries?
- ❑ How will you optimize your model?
  - ❑ It’s fine here to call any 3rd-party libs
- ❑ Did you encounter any troubles?
  - ❑ Scalability? Overfitting?

Literature

- Has your dataset/task been studied by others before?
- How the dataset was used?
- Are you working on a brand-new task?
- How are other people attacking the same/similar tasks?
- What is state-of-the-art method in this task or related tasks?
- Are your conclusions similar or different from existing work?
- What’s the major novelty of your work?
- …

Results

- ❑ Does your proposed method outperform the baselines?
  - ❑ Why your model can outperform?
  - ❑ Or why your model fails?
- ❑ Whether the gap is significant?
- ❑ Are all features you designed effective?
- ❑ How shall one set the hyper-parameters of your model?
- ❑ What are the major takeaways (i.e., conclusions)?
- ❑ …

Results - Example

- Performance comparison different methods
  - Baselines + Your proposed model
- Case Study
  - Some interesting cases when your model performs very well/poor
- Parameter Sensitivity
  - How do you decide hyperparameters?
  - Is the result sensitive to these hyperparameters?
- …