Information Visualization - Project Report

| Maxwell Júnior  79457  Group 15  IST - Alameda | Margarida Morais  86473  Group 15  IST- Alameda | Yasser Zacarias  88647  Group 15  IST - Alameda |
| --- | --- | --- |

# INTRODUCTION

# One of the first things we are taught in school is how to read. And it is this knowledge that allows us to get most of the information in our lives. When learning, most of the time the way we gather the information we need is either through reading books, or in the present times, by reading articles on the internet and searching information through websites (i.e. Wikipedia).

# Not only is reading important when studying, but also, when reading books in our daily lives, and this is because reading helps us develop skills that are important to our wellbeing. It helps to improve your self-expression capabilities since you are extending your vocabulary, it can teach you how to deal with certain obstacles you find in life or help you to learn a new skill.

# Our desired outcome is understanding how different these habits are between different countries throughout Europe, and also try to find whether are correlations between them or not. if it has an impact in the overall picture of the demographics indicators such as the level of dropout or even high achievement in education system.

# Initially there were many questions about the subject that we wanted to answer through the visualizations, but after further study of the topic, we realized that there was a need to narrow our scope that would be feasible given our data limitations, and only to focus on the most important questions.

With this in mind, we decided to come out with the questions below, that should be answered by correlating reading habits metrics and some demographic indicators.

Possible Tasks

This visualization will allow to:

*Search*

* Search a country and visualize their current information about time spent reading books in minutes.

*Identify*

* Observe countries metrics like percentage early leavers by countries through the years.

*Explore*

* Visualize the evolution of the metrics of population average income by the education level in a set and specific countries through the years.

*Compare*

* Add information related to tasks available in the idioms

**Example Questions**

1. How many hours, in average, do the countries in EU spend reading?
2. What is the average income for education for level?
3. What is country reading habit and dropout rate?
4. Household expenditure in books?

## RELATED WORK

Regarding the inspiration and motivation for this work, it all began with our desire to present a work that could corelate important subjects and perhaps, untapped trends. However, theoretical lectures were also very helpful to gather useful information that guide us to structure the idea of how to encode the data, and what types of idioms to use in order to get the most out of available tools.

For our research subject specifically, we use different sources as websites, and scientific articles, to gather the highest amount of information, and data to support our subject, but most precisely reading habits across Europe. This led us to information like, time spent reading books.

Uma imagem com captura de ecrã

Descrição gerada automaticamente

Figure 1. Example of visualization of time spent reading books in Europe

In regard to implementation, we had good hints and useful examples from Eurostat, which in addition to the data, provides some visualizations on it. For example they had implemented a visualization of time spent reading books in bar chart, so it did not allow for a good assessment of the difference in terms of minutes, as the interval between minutes was too small, and it did not make a good differentiation between the colors used. However, it was a good example of what we should not do for our project.

We also like to mention that during laboratory presentations, whenever others groups presented their works, we learn and look carefully to their approach and sometimes we asked them to share the source of the code, so we could implement something similar.

**THE DATA**

Our main source of was the site Eurostat, which is the statistical office of the European Union. Its mission is to provide high quality statistics for Europe. It also provides allowed us to get the necessary information to fulfill our purpose of comparisons between countries data.

**Raw** **Data**

Originally our dataset was composed by about 10 different .csv and .xlsx files (approximately 614 KB of crude data), related to multiples countries from a time period of roughly from 2002 to 2016 – however, the most complete period of time in terms of data was approximately from the 2009 to 2018.

We face few challenges of acquiring and processing the data. Because at the beginning we wanted to compare time spent reding books with other metrics like rate of dropout, underachievement of students in reading, mathematics and science, average incoming by education level. But the lack of data with variables that could relate these domains, in way that the comparisons could make a really sense, and provide valuable information for the project, made us change a bit and focus only on the accurate data acquired and exploit them.

**Missing Values, and Data Cleaning**

Assuming that we had multiple data files with heterogenous format, gathered from the same source, but with different domains. It was challenging to standardize them.

To ensure the quality of the data needed to answer raised questions, it was necessary to attend a process of cleaning, that involved eliminate unwanted columns with unnecessary attributes to keep the coherency of the data. Also, there was a lot of empty cells, that we had to decide the best approach to address this problem. Therefore, for the case of some countries that were in a conflict or got divided along the years, we assign value “0” to cells that were empty and when we could not add a value, we erased the country.

In the other hand, most of data initially had column named value, which represented a value, for example, a percentage or minutes spent reding. It was necessary to rename each of value attribute with a name that identifies the domain that we were addressing.

None less important, we standardize the name of countries with the abbreviations of them, having in mind that we would use it to represent in graphs.

The process above described was possible by using Pentaho Data Integration as the main tool. Using transformations and other resources available on Pentaho Data Integration. The entire data in different formats were processed, and transformed to .json files, for later use in D3.

## VISUALIZATION

Final page presented below

Uma imagem com captura de ecrã

Descrição gerada automaticamente

**Overall Description**

Our final visualization solution consists in four different idioms, a choropleth map on top right, heatmap on lower right corner, stacked bar chart on top left and finally a line chart on lower left corner. The first part is the top right which have the choropleth map, that can be navigated with mouse interacting with user presenting a tooltip with information regard selected country. Note that each country has associated a color that represents/encodes the respective country.

On the left side of choropleth map, stacked bar chart with a selection box, that allow interactively to select a year to observe the stats. Following the idea of selection box, also for the line chart it is possible to select a country from country selection box, the respective country related line will be drowned in the chart.

In order to provide more and particulars details of the functioning of the page, and the elements represented in them, bellow we provide some more information for each of representation.

*Choropleth Map*

Uma imagem com texto, mapa

Descrição gerada automaticamenteThe choropleth map (figure..) allows the user to take some conclusions regarding the how much time some countries spent reading books.

From the color scale it is possible to understand the differentiation of the countries that spent more time, by the strongest colors meaning more time spent and less strong less time.

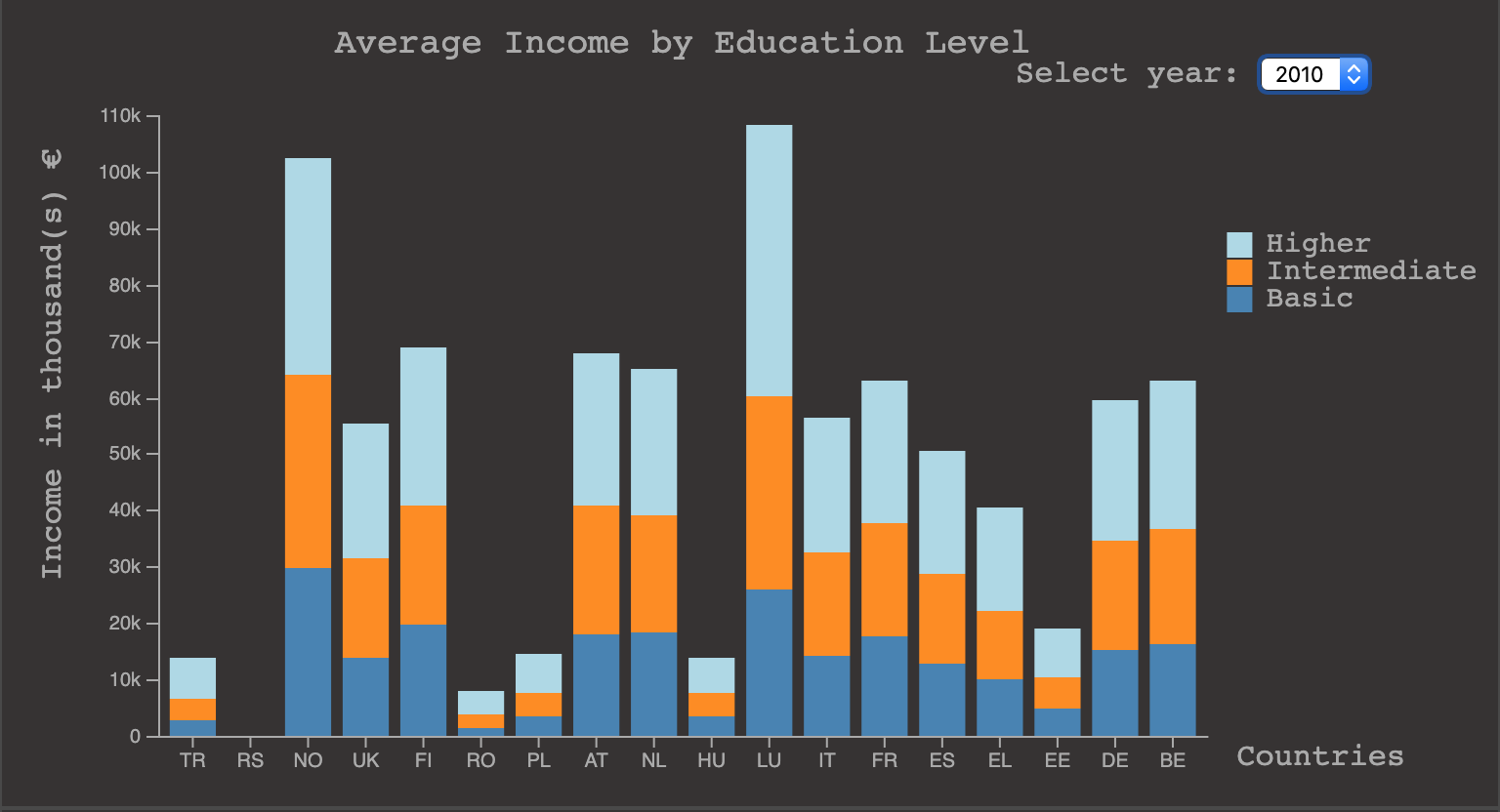
To ensure interaction, when mouse over the country a pop-up tooltip shows country name and of course minutes, corresponding the evaluated country.

Uma imagem com texto, mapa

Descrição gerada automaticamente

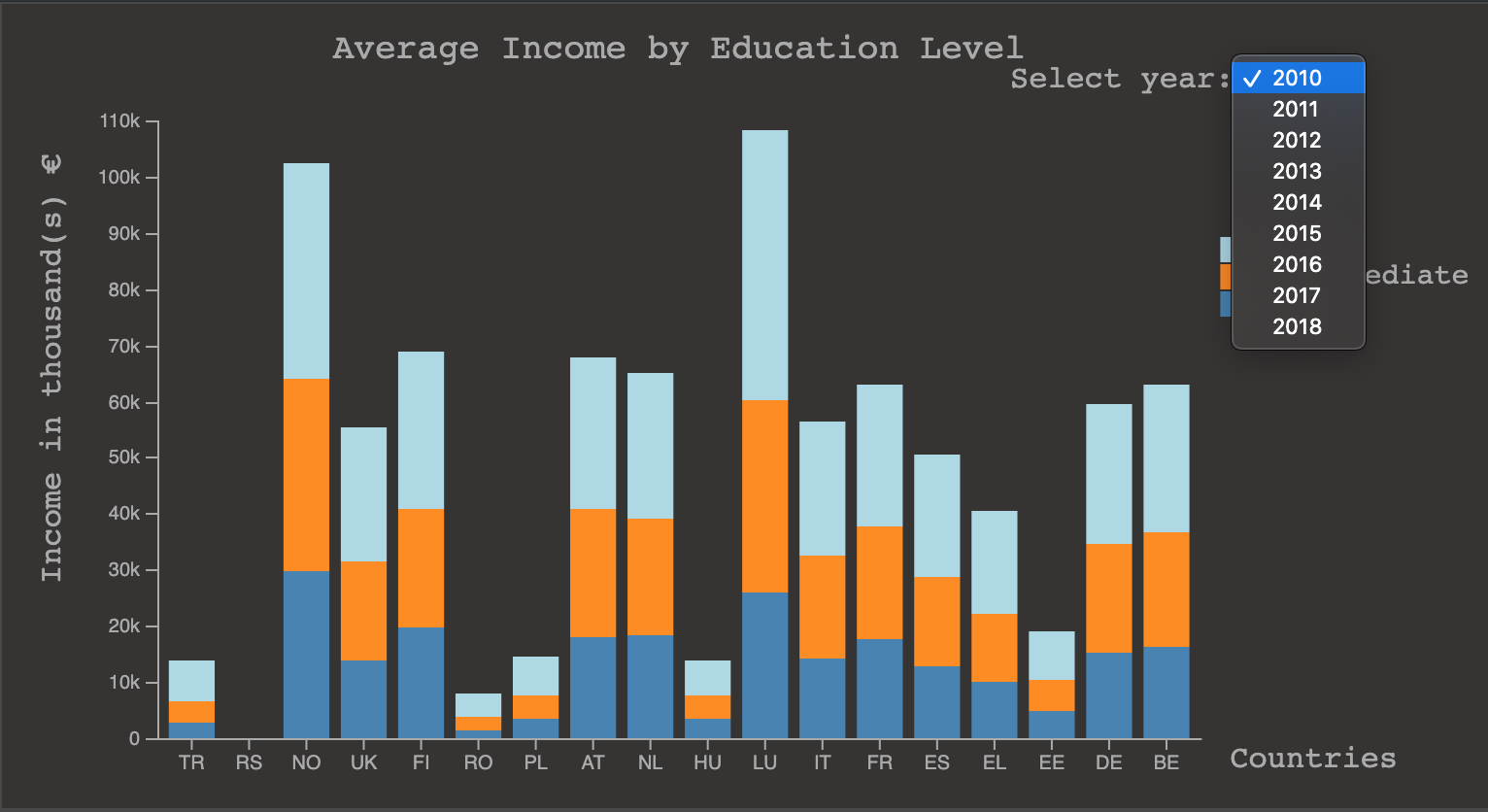
*Stacked Bar Chart*

The stacked bar chart (figure..) allows user to compare education levels, average income by levels of education through the years, showing how it grows.



As is possible to observe, each axis is labeled with the metric it represents, and each bar encodes a country. Three different colors encode level of education.

Since we always want to compare values of incoming, Y axis represents the scale of values.



The figure above shows when mouse hovers the select box, the years are displayed and if a specific year is selected, the bars assume different values representing the year.

Uma imagem com captura de ecrã

Descrição gerada automaticamente

It might hard to know the values representing income only by looking to Y axis. Tooltip is displayed with country name, education level, and values every time we hover above the bar.

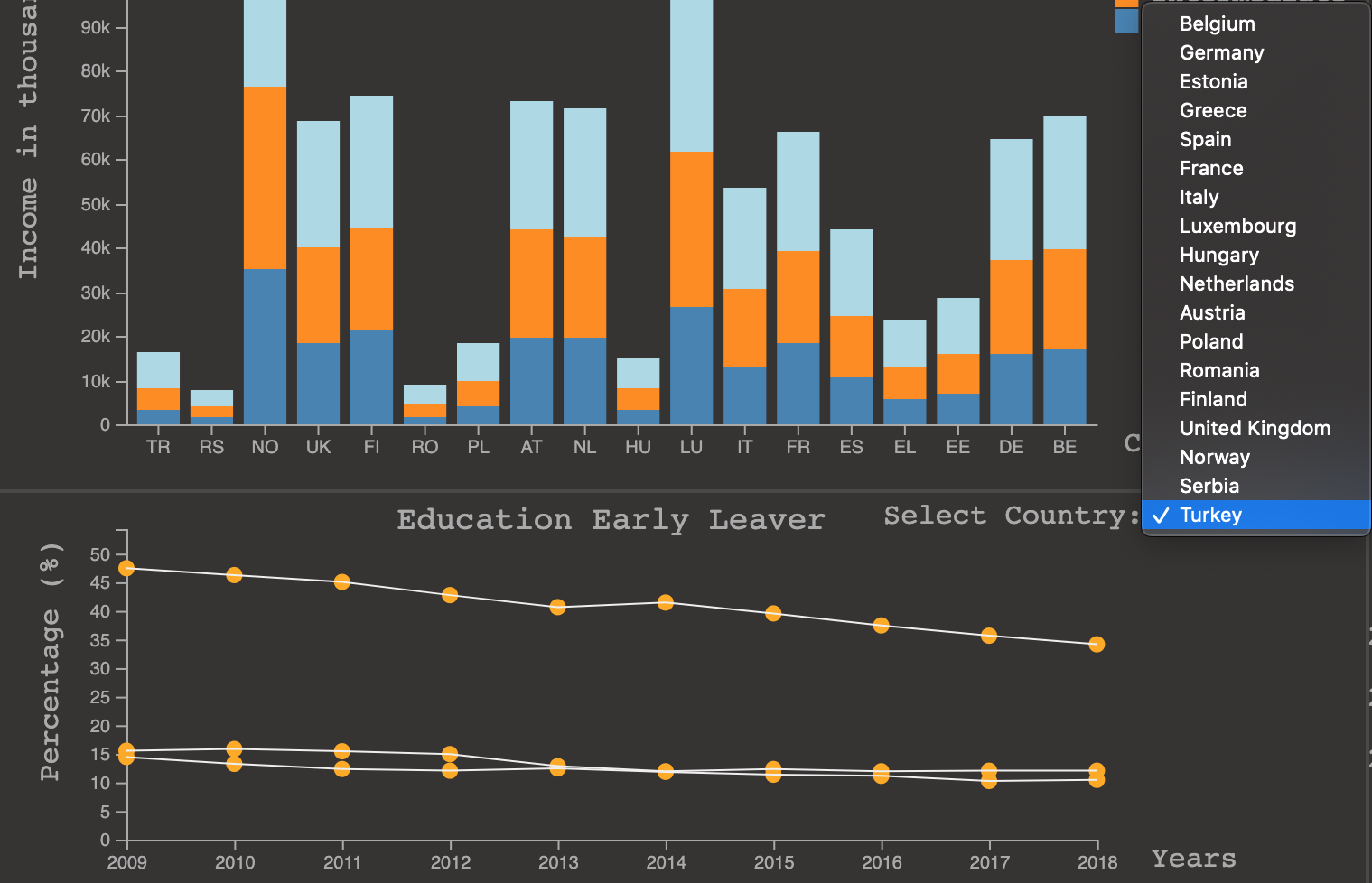
*Line Chart*

The line chart (figure..) allows the user to compare education early leaver percentage of different countries.

Uma imagem com interior, monitor, captura de ecrã, preto

Descrição gerada automaticamente

Each axis is labeled with the metric it represents, and each line represents a country. However, each circle encodes specific percentage value corresponding the country.



From the figure above, a select box displays country names when we hover the mouse on it and if we select one country, a line will be plotted in the line chart.

Uma imagem com monitor, interior, captura de ecrã, ecrã

Descrição gerada automaticamente

## After line added, it is possible to interact with line chart by hover through the circles and a tooltip is displayed with country name, year, and percentage of education early leaver.

**Rationale**

## When we started to work on our sketch out layout and selecting from amount of options the idioms that could visually represents the data, we dive into a healthy discussion, because all ideas could be valid but not necessarily accepted. Since we were working with mostly related with countries, it makes all sense to use a map as idiom to represent countries and their region, although whole countries were from Europe.

## We whose choropleth map as it represented a great idiom to overview a metric value for a specific country, as well as comparing it with other countries.

## Regarding interaction of choropleth map, when we hover the mouse in each of countries it underlines, and a tooltip with information related with data encoded is displayed.

The choice of stacked bar chart was in function of the variables that we had to encode in only one idiom, so as not create difficulties for the user to understand the information that we were trying to address. We also had a good feedback and ideas from the lab Professor.

We end up implementing stacked bar chart to encode country, average income value, and education level. Which became the best option for this visualization.

When it comes to interaction with a selection box it is possible to navigate through the years, being able to distinguish the income levels by the color that represents them and the values that could be seen by hovering the mouse on the bars, providing a pleasant user experience.

Even though the line chart did facilitate the comparisons between metrics, it might not work when we have to plot 18 lines for countries at the same. It turns to us a challenge to overcome. However, we decided to add a selection box for countries, which allow us to have room to encode more data in the same visualization.

In this case we end up encoding the early leaver value (in percentage) and adding a very useful tooltip that aggregates most of information of the visualization. Also, it added the necessary interaction with the idiom, which enabled the comparison of more than two countries.

Uma imagem com texto

Descrição gerada automaticamente

Initial we come out with different ideas and translate most of them in the first sketch, it was very ambitious, and it seems to be perfect. But when we dive in the project and the implementation technology, we get to understand that was a little bit hard to implement all of the ideas, because it was necessary spend a lot time and work to accomplish every task.

There was only one idiom that ended up being implemented, which is the choropleth map. Because, it was the one that fit in our presentation and for the type of attribute that we were exploiting it works.

For example, at beginning we were trying to implement average income by education level on a heatmap, but after evaluating the quantity of variables that we had to encode, it became clear that the idea did not work.

We also were stuck in the idea of use the parallel coordinates to encode data related with book and newspaper household expenditure, but it was not the best idiom to represent, because after we tried to implement we saw that it could not done with the structure of the lines, crossing each other.

## Why did you think your techniques would work? What visual encodings did you use and why (and why not others)? What alternatives did you consider, even if they turned out not to work? Especially, discuss how you managed the complexity of real data, and matters of scalability. Also, include in your discussion the evolution of the prototype, from the initial sketches to the last version highlighting what you learned from version to version and how that influenced your design.

**Demonstrate the Potential**

## In this section we will demonstrate our visualization implementation and discuss the idioms and the way they interact each other.

To demonstrate the potential of our visualization, we will pick some questions and tasks mentioned in the introduction and show how them can be achieved.

Describe for at least a couple of cases (from the questions you promised you’d answer before) where, step by step (illustrated with screenshots), you find the answers you seek. In short, demonstrate the potential of your solution! Does your visualization provide insights on data that *you were not expecting / that are not common knowledge*? ***These are pure gold!*** Be sure to include them!

**IMPLEMENTATION DETAILS**

**Overall Description**

hat challenges did you find and overcome? How did you implement the links between the views (incl. brushing, etc.)? What algorithms did you use? What techniques did you adapt, or implement, from scratch? (instead of just copying & pasting them from the D3 examples page...)

## References and Citations

Use a numbered list of references at the end of the article, ordered alphabetically by last name of first author, and referenced by numbers in brackets [1,3,4].

Your references should be published materials accessible to the public. Internal technical reports may be cited only if they are easily accessible (i.e., you provide the address for obtaining the report within your citation) and may be obtained by any reader for a nominal fee. Proprietary information may not be cited. Private communications should be acknowledged in the main text, not referenced

References should be in ACM citation format: <http://acm.org/publications/submissions/latex_style>. This includes citations to internet resources [1,4,8,1] according to ACM format, although it is often appropriate to include URLs directly in the text, as above.

# CONCLUSIONS

To conclude, unfortunately we cannot say that this project was successful to attend the expectations that we create around them, having in mind the potential of technologies available.

Although the difficulties that we face to manage our time and the tasks that purpose to do, we enjoy most of the steps of the project, since the concept definition, data collection, to drawing the sketches, and finally hands on implementation. We were also glad to learn a lot of the reading habits of some countries in Europe.

What did you learn? Were you able to address all the questions? If you were to start over, what would you have done differently? Also, if you now had 1 more month and €3000 do spend on this, what else would you do to enrich your solution?

The heading of a section should be in Arial 9-point bold, all in capitals (Heading 1 style). Sections should not be numbered.

## FUTURE WORK

Headings of subsections should be in Arial 9-point bold with initial letters capitalized (Heading 2 style). For sub-sections and sub-subsections, a word like *the* or *of* is not capitalized unless it is the first word of the heading.

### Sub-subsections

Headings for sub-subsections should be in Arial 9-point italic with initial letters capitalized (Heading 3 style).

# FIGURES/CAPTIONS

Place figures and tables at the top or bottom of the appropriate column or columns, on the same page as the relevant text (see Figure 1). A figure or table may extend across both columns to a maximum width of two columns, or 17.78 cm (7 in.).

Captions should be Times New Roman 9-point bold (Caption style). They should be numbered (e.g., “Table 1” or “Figure 2”), centered, and placed beneath the figure or table. The words “Figure” and “Table” should be spelled out (e.g., “Figure” rather than “Fig.”) wherever they occur.

All figures should also include alt text for improved accessibility. In Word, right click the figure, and select Format Picture | Layout | Alt Text). Papers and notes may use color figures, which are included in the page limit; the figures must be usable when printed in black-and-white in the proceedings.

The paper may be accompanied by a short video figure up to five minutes in length. However, the paper should stand on its own without the video figure, as the video may not be available to everyone who reads the paper.

## Inserting Images

Occasionally MS Word generates larger-than-necessary PDF files when images inserted into the document are manipulated in MS Word. To minimize this problem, use an image editing tool to resize the image at the appropriate printing resolution (usually 300 dpi), and then insert the image into Word using Insert | Picture | From File.



Figure 2. Sample of a wide figure. Be sure to place at the top or bottom of the page. Ensure that important information is legible in both black-and-white and color printing. Image: CC-BY-ND ayman on Flickr.

## Table Style

The text of tables will format better if you use the Table Text style (as in Table 1). If you do not use this style, then you may want to adjust the vertical spacing of the text in the tables. To adjust the spacing of text in a table in Word, use Home | Paragraph | Indents and Spacing. Generally, text in each field of a table will look better if it has equal amounts of spacing above and below it, as in Table 1. Table captions should be placed below the table. We recommend table lines be 1 point, 25% black. Minimize use of unnecessary table lines.

For improved accessibility, header rows of tables should be marked. In Word, right-click a header row, and select Table Properties | Row | Repeat as header…

# LANGUAGE, STYLE AND CONTENT

The written and spoken language of SIGCHI is English. Spelling and punctuation may use any dialect of English (e.g., British, Canadian, US, etc.) provided this is done consistently. Hyphenation is optional. To ensure suitability for an international audience, please:

* Write in a straightforward style.
* Try to avoid long or complex sentence structures.
* Use common and basic vocabulary (e.g., use the word “unusual” rather than the word “arcane”).
* Briefly define or explain all technical terms that may be unfamiliar to readers.
* Explain all acronyms the first time they are used in your text—e.g., “Digital Signal Processing (DSP)”.
* Explain local references (e.g., not everyone knows all city names in a particular country).
* Explain “insider” comments. Ensure that your whole audience understands any reference whose meaning you do not describe (e.g., do not assume that everyone has used an Android phone, or a particular application).
* Explain colloquial language and puns. Understanding phrases like “red herring” may require a local knowledge of English. Humor and irony are difficult to translate.
* Use unambiguous forms for culturally localized concepts, such as times, dates, currencies, and numbers (e.g., “1-5- 97” or “5/1/97” may mean 5 January or 1 May, and “seven o’clock” may mean 7:00 am or 19:00). For currencies, indicate equivalences: “Participants were paid ₩22, or roughly US$29.”
* Be careful with the use of gender-specific pronouns (*he*, *she*) and other gendered words (*chairman*, *manpower*, *man-months*). Use inclusive language that is gender-neutral (e.g., *she* *or* *he*, *they*, *s/he*, *chair*, *staff*, *staff-hours*, *person-years*). See the *Guidelines for Bias-Free Writing* for further advice and examples regarding gender and other personal attributes [9]. Be particularly aware of considerations around writing about people with disabilities.
* If possible, use the full (extended) alphabetic character set for names of persons, institutions, and places (e.g., Grønbæk, Lafreniére, Sánchez, Nguyễn, Universität, Weißenbach, Züllighoven, Århus, etc.). These characters are already included in most versions and variants of Times, Helvetica, and Arial fonts.

# Accessibility

The Executive Council of SIGCHI has committed to making SIGCHI conferences more inclusive for researchers, practitioners, and educators with disabilities. As a part of this goal, the all authors are asked to work on improving the accessibility of their submissions. Specifically, we encourage authors to carry out the following five steps:

1. Add alternative text to all figures
2. Mark table headings
3. Generate a tagged PDF
4. Verify the default language
5. Set the tab order to “Use Document Structure”

For more information and links to instructions and resources, please see: <http://chi2016.acm.org/accessibility>.

# Page Numbering, Headers, and Footers

Your final submission should not contain footer or header information at the top or bottom of each page. Specifically, your final submission should not include page numbers. Initial submissions may include page numbers, but these *must* be removed for camera-ready. Page numbers will be added to the PDF when the proceedings are assembled.

# Producing and testing PDF files

We recommend that you produce a PDF version of your submission well before the final deadline. Your PDF file must be ACM DL Compliant. The requirements for an ACM Compliant PDF are available at:

<http://sheridanprinting.com/typedept/ACM-distilling-settings.htm>

When creating your PDF from Word, ensure that you generate a tagged PDF from improved accessibility. This can be done by using the Adobe PDF add-in, also called PDFMaker. Select Acrobat | Preferences from the ribbon and ensure that “Enable Accessibility and Reflow with tagged Adobe PDF” is selected. You can then generate a tagged PDF by selecting “Create PDF” from the Acrobat ribbon. Test your PDF file by viewing or printing it with the same software the publisher will use, Adobe Acrobat Reader Version 10, which is widely available at no cost. Note that most reviewers will use a North American/European version of Acrobat Reader, so please check your PDF accordingly.

# Conclusion

It is important that you write for the SIGCHI audience. Please read previous years’ proceedings to understand the writing style and conventions that successful authors have used. State clearly what you have done, not merely what you plan to do, and explain how your work is different from previously published work, i.e., *the unique contribution that your work makes to the field*. Please consider what the reader will learn from your submission, and how they will find your work useful. If you write with these questions in mind, your work is more likely to be successful, both in being accepted into the conference, and in influencing the work of our field.

# ACKNOWLEDGMENTS

Sample text: We thank all the volunteers, and all publications support and staff, who wrote and provided helpful comments on previous versions of this document. Authors 1, 2, and 3 gratefully acknowledge the grant from NSF (#1234-2012-ABC). This is just an example.

# References format

References must be the same font size as other body text. References should be in alphabetical order by last name of first author. Example reference formatting for individual journal articles [3], articles in conference proceedings [7], books [9], theses [10], book chapters [11], an entire journal issue [6], websites [1,4], tweets [1], patents [5], and online videos [8] is given here. This formatting is a slightly edited version of the format automatically generated by the ACM Digital Library (http://dl.acm.org) as “ACM Ref”. More details of reference formatting are available at:

<http://www.acm.org/publications/submissions/latex_style>

Note that the Hyperlink style used throughout this document uses blue links; however, URLs that appear in the references section may appear in black.

# REFERENCES

1. @\_CHINOSAUR. 2014. VENUE IS TOO COLD. #BINGO #CHI2016. Tweet. (1 May, 2014). Retrieved February 2, 2014 from https://twitter.com/\_CHINOSAUR/status/461864317415989248
2. ACM. How to Classify Works Using ACM’s Computing Classification System. 2014. Retrieved August 22, 2014 from [http://www.acm.org/class/how\_to\_use.html](http://www.acm.org/class/how_to_use.html%20)
3. Ronald E. Anderson. 1992. Social impacts of computing: Codes of professional ethics. *Soc Sci Comput Rev* 10, 2: 453-469.
4. Anna Cavender, Shari Trewin, Vicki Hanson. 2014. Accessible Writing Guide. Retrieved August 22, 2014 from <http://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/>
5. Morton L. Heilig. 1962. Sensorama Simulator, U.S. Patent 3,050,870, Filed January 10, 1961, issued August 28, 1962.
6. Jofish Kaye and Paul Dourish. 2014. Special issue on science fiction and ubiquitous computing. *Personal Ubiquitous Comput*. 18, 4 (April 2014), 765-766. <http://dx.doi.org/10.1007/s00779-014-0773-4>
7. Scott R. Klemmer, Michael Thomsen, Ethan Phelps-Goodman, Robert Lee, and James A. Landay. 2002. Where do web sites come from?: capturing and interacting with design history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '02), 1-8. <http://doi.acm.org/10.1145/503376.503378>
8. Psy. 2012. Gangnam Style. Video. (15 July 2012.). Retrieved August 22, 2014 from <https://www.youtube.com/watch?v=9bZkp7q19f0>
9. Marilyn Schwartz. 1995. *Guidelines for Bias-Free Writing.* Indiana University Press.
10. Ivan E. Sutherland. 1963. *Sketchpad, a Man-Machine Graphical Communication System*. Ph.D Dissertation. Massachusetts Institute of Technology, Cambridge, MA.
11. Langdon Winner. 1999. Do artifacts have politics? In *The Social Shaping of Technology* (2nd. ed.), Donald MacKenzie and Judy Wajcman (eds.). Open University Press, Buckingham, UK, 28-40.