

# Structured documents

Tools for writing science

Andrés Aravena, PhD

*October 13, 2022*



# Structured documents

You probably know that using a good *data structure* can dramatically improve an algorithm

And you use *structured programs*

The same applies to structuring our documents

Maybe you have used LaTeX, or Markdown

Maybe you know HTML

# Separation of concerns

Separate style from structure

Describe the role of text, not the “looks”

The key idea is to describe *what things are*, not *how they look*

This part is based on the ideas discussed in “LaTeX: A Document Preparation System” by Leslie Lamport (1986).

# It is like a house

Structure makes the house solid and comfortable

If you only do decoration, the house looks nice but it is not solid

Structure of the walls come first

Painting the walls in a nice color is secondary

# Structural elements

- Sections, subsections, paragraphs
- Figures and Tables
- Lists
- References
- Equations
- Metadata
  - Title
  - Authors
  - Affiliations
  - Dates: submission, acceptance

# Microsoft Word

The first tool we learn today is a WYSIWYG word processor

In word processors like Word®,  
*What You See Is What You Get*

This is sometimes called **WYSIWYG**

It is easy to change fonts, sizes, colors and other visual attributes, without paying attention to *structure*

# Style is not Structure

You can follow the same philosophy:

- Separate style from structure
- Focus on content



# Structured Word documents

The screenshot shows the Microsoft Word interface. The ribbon at the top includes tabs for Home, Insert, Draw, Design, Layout, References, Mailings, Review, View, Tell me, Share, and Comments. The Home tab is active, showing font settings (Calibri Light, size 28) and paragraph alignment options. The document content is as follows:

## Ten Simple Rules for Online Learning

*PLOS Computational Biology*

David B. Searls(\*)  
Independent Consultant, Philadelphia, Pennsylvania, United States of America  
\* E-mail: David.B.Searls@gmail.com  
The author has declared that no competing interests exist.  
David B. Searls is an Associate Editor of PLOS Computational Biology.

Fran Lewitter, Editor  
Whitehead Institute, United States of America

Copyright © 2012 David B Searls  
DOI: 10.1371/journal.pcbi.1002631  
Published in print: September 2012  
Published online: 13 September 2012

The success of online courseware such as that offered by the Massachusetts Institute of Technology (MIT) (<http://ocw.mit.edu>) and now by many other institutions, together with a plethora of recent announcements of major new initiatives in this arena such as Coursera (<https://www.coursera.org>), Udacity (<http://www.udacity.com>), and the Harvard-MIT partnership edX (<http://www.edxonline.org>), have made it clear that online learning has reached a tipping point. Many signs point to the possibility in the near future of getting a quality, university-level education at a distance, and for free. As exciting as this prospect may be, it behooves online students to follow a few simple rules for getting the most out of the experience, while being realistic in their expectations, as outlined below.

### Rule 1: Make a Plan

There are many possible motivations for becoming involved in online learning, whether in bioinformatics or any other field. There's nothing wrong with taking an online course on impulse, or to fill a very specific need, or simply for fun, if that's your goal. But if you hope to

The Styles pane on the right shows the 'Current style' as 'Title' (font size 1a). Under 'Apply a style', a list of styles is shown: Clear Formatting, Normal, No Spacing, Heading 1, Heading 2, Title, and Subtitle. The 'Title' style is highlighted. Below this, a 'List' dropdown is set to 'Recommended', and there are checkboxes for 'Show styles guides' and 'Show direct formatting guides'.

Annotations: A red oval highlights the main title 'Ten Simple Rules for Online Learning'. A red line connects this title to the 'Title' style in the Styles pane. A green oval highlights the text 'while being realistic in their expectations' in the paragraph. A green line connects this text to the 'Heading 1' style in the Styles pane. Another green line connects the 'Rule 1: Make a Plan' heading to the 'Heading 1' style.

# Now the document has structure

The image shows a document editor interface with a dark sidebar on the left and a main content area on the right. The sidebar contains a table of contents with the following items:

- Rule 1: Make a Plan
- Rule 2: Be Selective
- Rule 3: Organize Your Learning
- Rule 4: Do the Readings
- Rule 5: Do the Exercises
- Rule 6: Do the Assessments
- Rule 7: Exploit the Advantages
- Rule 8: Reach Out
- Rule 9: Document Your
- Rule 10: Be Realistic
- References

The main content area displays the document titled "Ten Simple Rules for Online Learning". The document is structured as follows:

**Ten Simple Rules for Online Learning**

*PLoS Computational Biology*

David B. Searls(\*)  
Independent Consultant, Philadelphia, Pennsylvania, United States of America  
\* E-mail: David.B.Searls@gmail.com  
The author has declared that no competing interests exist.  
David B. Searls is an Associate Editor of PLOS Computational Biology.

Fran Lewitter, Editor  
Whitehead Institute, United States of America

Copyright © 2012 David B Searls  
DOI: 10.1371/journal.pcbi.1002631  
Published in print: September 2012  
Published online: 13 September 2012

The success of online courseware such as that offered by the Massachusetts Institute of Technology (MIT) (<http://ocw.mit.edu>) and now by many other institutions, together with a plethora of recent announcements of major new initiatives in this arena such as Coursera (<https://www.coursera.org>), Udacity (<http://www.udacity.com>), and the Harvard-MIT partnership edX (<http://www.edxonline.org>), have made it clear that online learning has reached a tipping point. Many signs point to the possibility in the near future of getting a quality, university-level education at a distance, and for free. As exciting as this prospect may be, it behooves online students to follow a few simple rules for getting the most out of the experience, while being realistic in their expectations, as outlined below.

**Rule 1: Make a Plan**

There are many possible motivations for becoming involved in online learning, whether in bioinformatics or any other field. There's nothing wrong with taking an online course on impulse, or to fill a very specific need, or simply for fun, if that's your goal. But if you hope to acquire a broader swath of knowledge for some larger purpose, you will need a directed, organized approach to be efficient and effective, especially in the absence of a formal degree program or traditional academic advisor. Don't underestimate the importance, or the difficulty, of this planning effort, particularly if you are new to the field.



# Historical note

Mechanical typewriters were  
invented in 1874

They had only one font

We still use the same keyboard



# Using UPPERCASE and underline for emphasis

## PRELIMINARY DISCUSSION OF THE LOGICAL DESIGN OF AN ELECTRONIC COMPUTING INSTRUMENT

### 1.0 Principal Components of the Machine

1.1. Inasmuch as the completed device will be a general-purpose computing machine it should contain certain main organs relating to arithmetic, memory-storage, control and connection with the human operator.

# Early computers had only *text*, no graphics



# Giving style to plain text

Since there was only one type of letter, people used some symbols as “magic”

For example `\` or `@`

If you write a “magic” symbol, you tell the computer that the next symbol shows a change of format

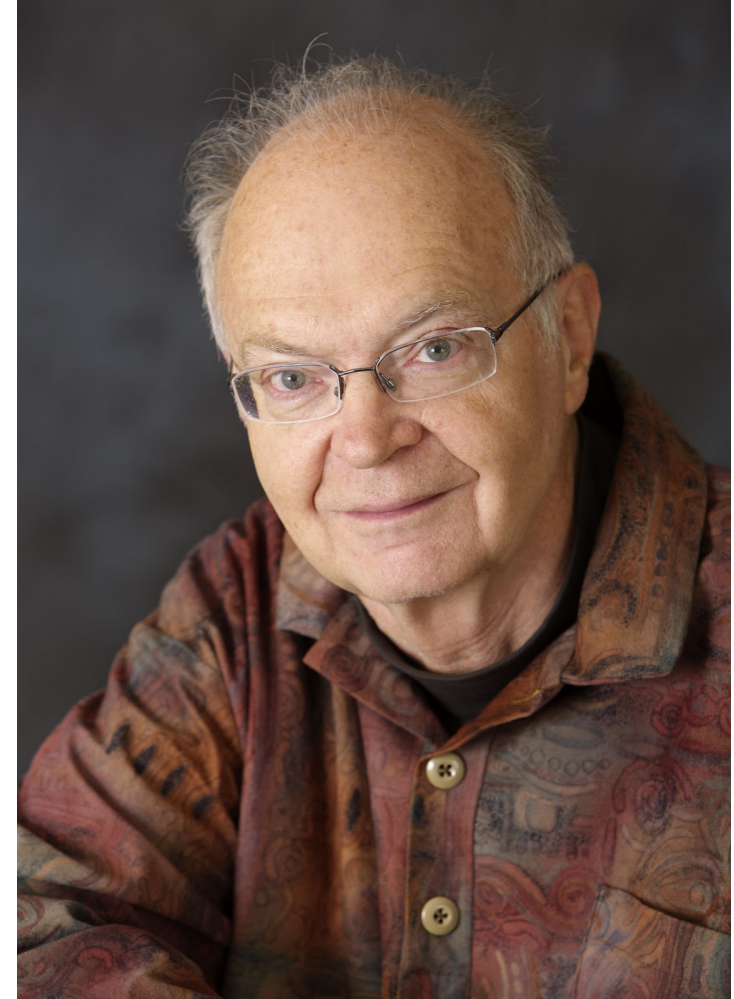
This is called **Markup Language**



# TeX

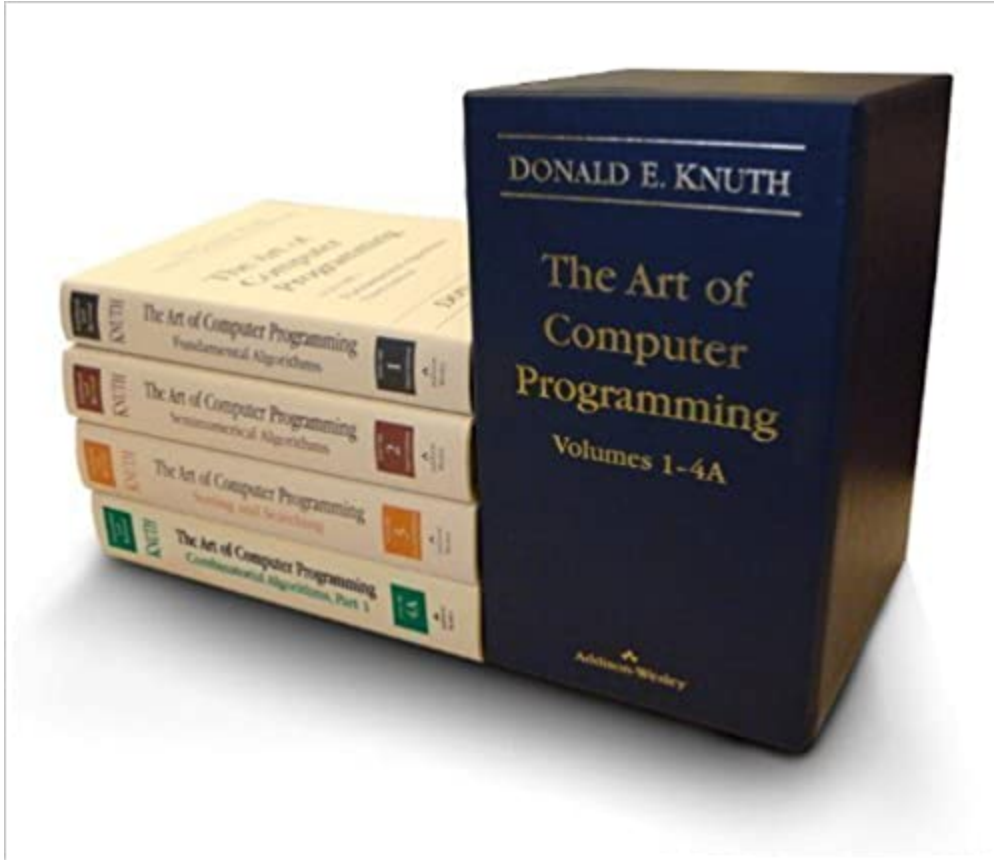
An important system to prepare documents in the computer was invented in the 70's by Donald Knuth, who is probably the most important computer scientist of the last 70 years.

Donald Knuth won the Turing Award in 1974





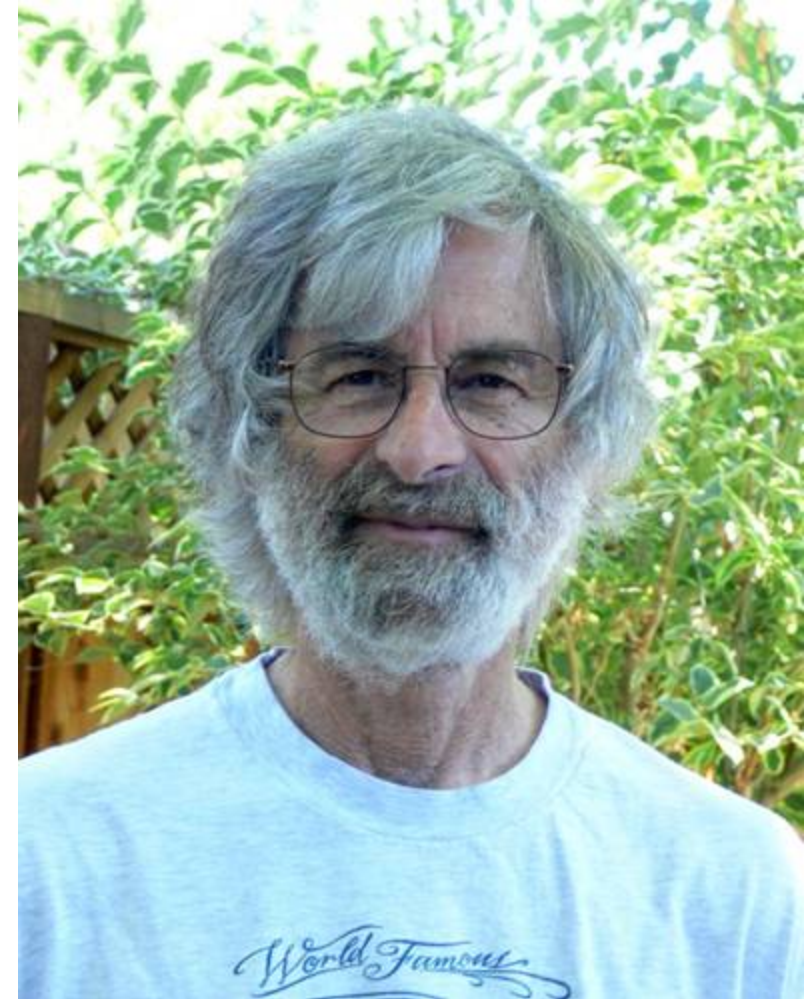
# Knuth invented TeX to write this



# LaTeX

TeX has styles but not structure. In the 80's Leslie Lamport created LaTeX as an extension of TeX

Leslie Lamport won the Turing Award in 2013



# Example: writing in LaTeX

A LaTeX document looks like this

```
\documentclass[a4paper]{article}
\title{Ten Simple Rules for Online Learning}
\author{David B. Searls}
\date{13 September 2012}
\begin{document}
\section{Rule 1: Make a Plan}
There are many possible motivations for becoming involved in online learning..
\end{document}
```

LaTeX files are text files. They will never be obsolete.

Changing the `documentclass` will change the document *look*

# Advantages of LaTeX

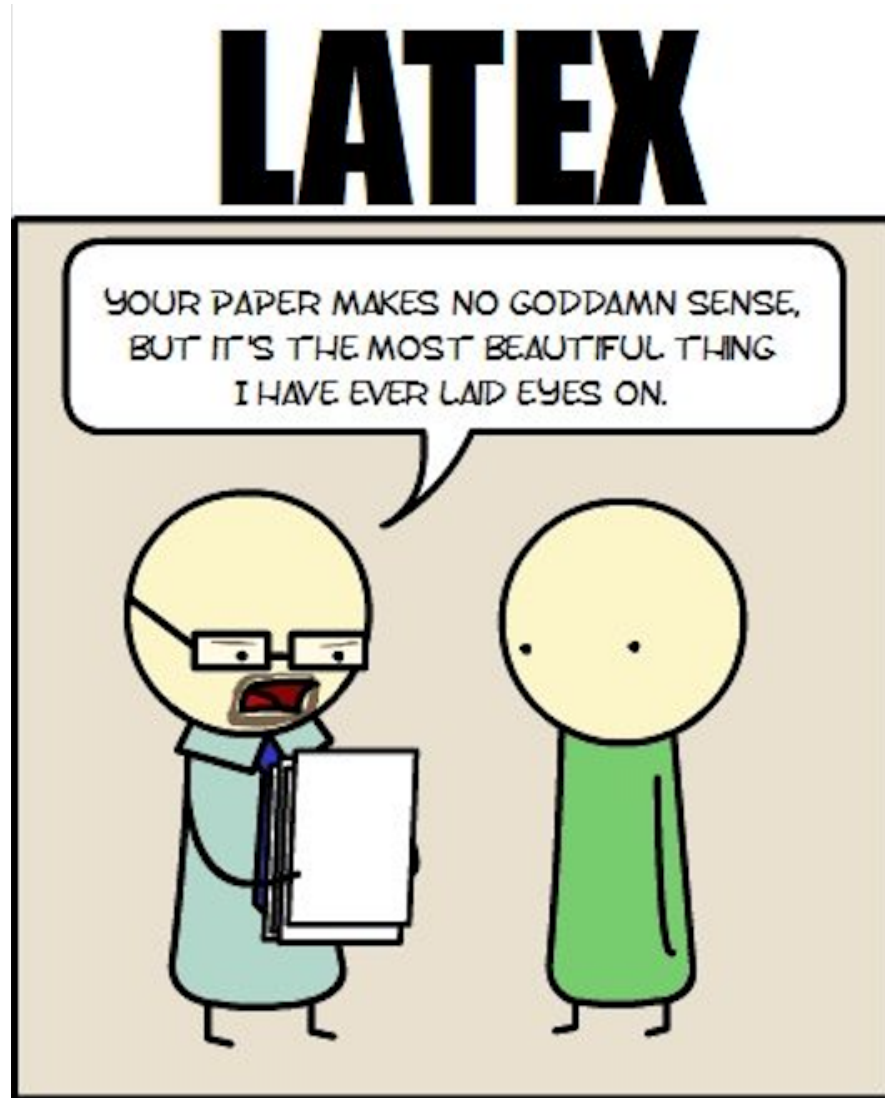
- It is free
- it *forces* you to think logically and organize your ideas
- Write first, compile later
- Do not waste time playing with fonts
- Good journals accept LaTeX submissions  
(they also accept Microsoft Word format)

# LaTeX files are *text files*

- Independent of any provider
- Use your favorite text editor (VScode?)
- Version control friendly (GitHub?)
- Can probably still be read 20 years from now

We cannot say the same about Microsoft Word

The real advantage: it looks correct



# According to the author of LaTeX

The main mistake that people should stop making is

*Worrying too much about formatting and not enough about content.*

“How (La)TeX changed the face of Mathematics”. An E-interview with Leslie Lamport. <http://lamport.azurewebsites.net/pubs/lamport-latex-interview.pdf>

# Bonus: Slides for presentations

After writing your paper, you will probably present it  
(or maybe before finishing it)

Using structured document makes it easy to *recycle* your  
material to presentation slides

In LaTeX you can do that using the `beamer` document class



# Good ideas in LaTeX

- Chapters, sections, subsections
- Automatic creation of Table of Contents
- Automatic numbering of sections, figures, tables
- Cross referencing sections, figures, tables
- Floating figures
- Math formulas
- Bibliographic references

# Writing Math Expressions

LaTeX is favored by people who writes mathematical formulas

```
$$ (a+b)^n = \sum_{k=0}^n \frac{n!}{k!(n-k)!} a^k b^{n-k} $$
```

$$(a + b)^n = \sum_{k=0}^n \frac{n!}{k!(n-k)!} a^k b^{n-k}$$

You can use this syntax in Microsoft Word's Equation Editor

Learning how to write math is a good investment

# Bibliographic References

There are *hundreds* of citation styles

Life is too short to sort references *manually*

LaTeX also provides a convenient way to handle references

References are stored in a separate text file, in BiBTeX format

Many tools can create BiBTeX files for you

- Zotero
- Mendeley

# LaTeX disadvantages

- LaTeX is hard to learn
  - This discourages many people
  - Your collaborators may not use it
  - You need to have the Reference Manual at hand
- It is oriented to producing printed material
  - It produces PDF files or equivalents
  - Not suitable for Web or eBook
- Writing tables is hard



# Web Pages

In the 90's most computers had good graphic capabilities and Internet access

Researchers at CERN invented the web, using “hyper-text”  
(That is, text with links to other text)

Web pages are written in *Hyper Text Markup Language*

# HTML

These are also *text files*. It looks like this:

```
<head>
<title>Ten Simple Rules for Online Learning</title>
</head>
<body>
<h1>Rule 1: Make a Plan</h1>
There are many possible motivations for becoming involved in online learning..
</body>
```

# Good ideas from HTML

- Works well on the screen: adapts to screen size
- Links to other pages
- Structural elements
  - `<h1>...</h1>` marks Header level 1
  - There are also `<h2>...<h6>`
- Comments: `<!-- this part is not shown -->`
- Structure separated from Style
  - Style is defined in CSS files



# Disadvantages of HTML

- It does not work well for paper
- It is hard to write manually
- There are editors, but they often focus on style, not structure

# Alternative: Markdown

It is a light markup system that can be easily converted into nice presentations

```
---  
title: Ten Simple Rules for Online Learning  
author: David B. Searls  
date: 13 September 2012  
...  
  
# Rule 1: Make a Plan  
  
There are many possible motivations for becoming involved in online learning...
```