Data and Project Organization:

Building your own Virtual Research Environment for Reproducible Research

Part II

Author: Georgios Kaklamanos

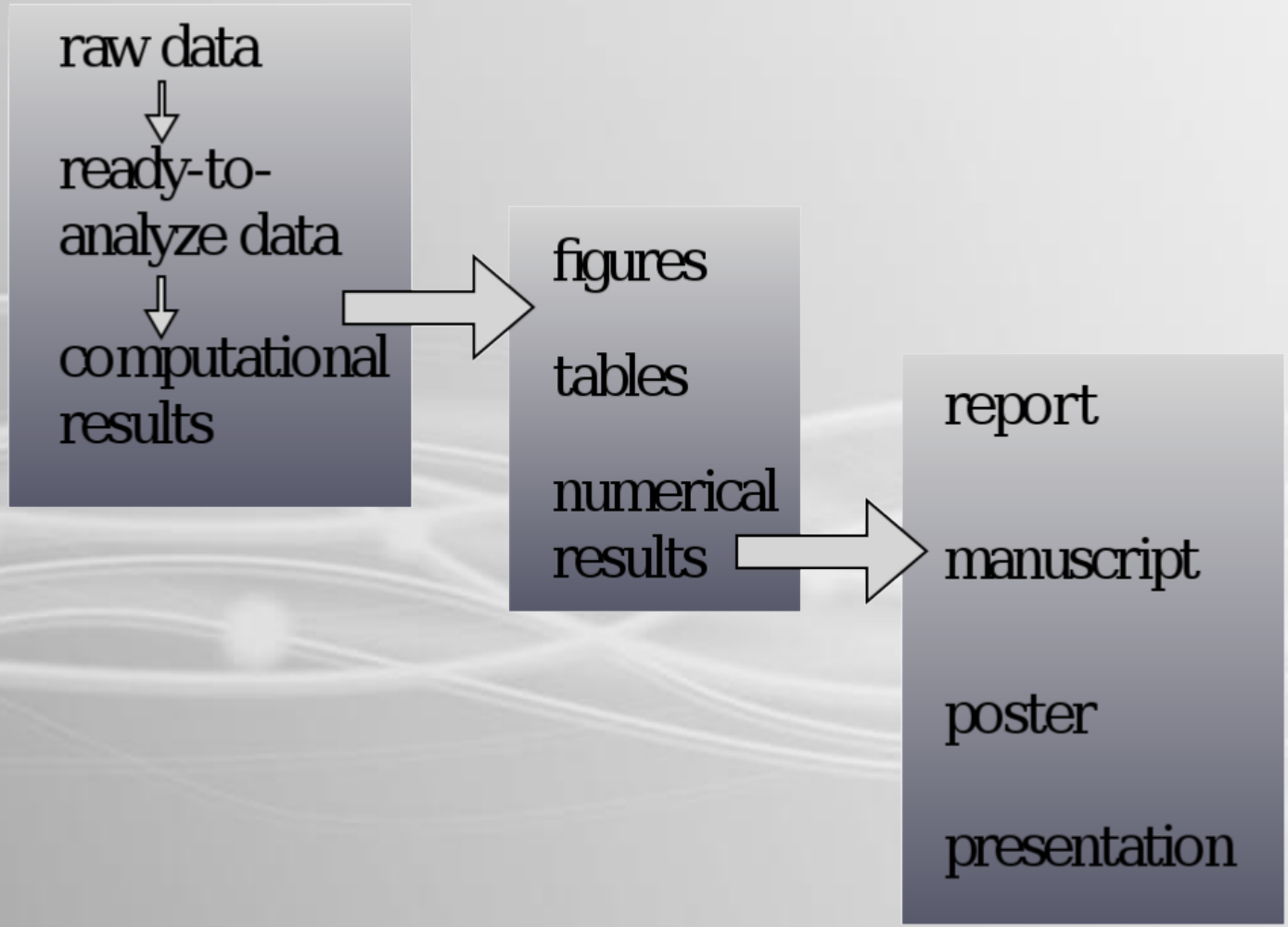
E-Science Group
Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

July 17th, 2017

Organization

Analysis Process

- There is a workflow in any research project



Prepare for change

However things will change:

- Files will change
- The analysis will change
- Results will change
- Figures will change
- ...

File Traversal

Figure: Visual Traversal

Name	Date Modified	Size	Kind
.DS_Store	Today 8:09 PM	40 KB	Document
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv	2014-05-08 8:05 PM	320 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv	2014-05-08 8:05 PM	309 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv	2014-05-08 8:05 PM	336 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv	2014-05-08 8:05 PM	342 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B02.csv	2014-05-08 8:05 PM	349 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B03.csv	2014-05-08 8:05 PM	359 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C01.csv	2014-05-08 8:05 PM	328 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C02.csv	2014-05-08 8:05 PM	319 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C03.csv	2014-05-08 8:05 PM	332 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D01.csv	2014-05-08 8:05 PM	317 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D02.csv	2014-05-08 8:05 PM	330 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D03.csv	2014-05-08 8:05 PM	354 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E01.csv	2014-05-08 8:05 PM	234 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E02.csv	2014-05-08 8:05 PM	245 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E03.csv	2014-05-08 8:05 PM	268 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F01.csv	2014-05-08 8:05 PM	304 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F02.csv	2014-05-08 8:05 PM	396 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F03.csv	2014-05-08 8:05 PM	324 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G01.csv	2014-05-08 8:05 PM	331 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G02.csv	2014-05-08 8:05 PM	285 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G03.csv	2014-05-08 8:05 PM	331 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv	2014-05-08 8:05 PM	318 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv	2014-05-08 8:05 PM	206 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv	2014-05-08 8:05 PM	310 KB	comm...value
2013-06-26_BRAFWTNEGASSAY_Plas...lline-100-1MutantFraction_platefile.csv	2014-04-01 4:49 PM	9 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv	2014-05-08 8:05 PM	353 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv	2014-05-08 8:05 PM	374 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv	2014-05-08 8:05 PM	394 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv	2014-05-08 8:05 PM	361 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A05.csv	2014-05-08 8:05 PM	336 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A06.csv	2014-05-08 8:05 PM	353 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A07.csv	2014-05-08 8:05 PM	336 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A08.csv	2014-05-08 8:05 PM	371 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A09.csv	2014-05-08 8:05 PM	362 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A10.csv	2014-05-08 8:05 PM	343 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A11.csv	2014-05-08 8:05 PM	393 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_B01.csv	2014-05-08 8:05 PM	423 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_B02.csv	2014-05-08 8:05 PM	396 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_B03.csv	2014-05-08 8:05 PM	424 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_B04.csv	2014-05-08 8:05 PM	353 KB	comm...value
2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_B05.csv	2014-05-08 8:05 PM	383 KB	comm...value

Figure: Text Traversal

File Naming

- Name and Location should be informative about the file
  - What it is
  - Why it exist
  - How it relates to other things
- Three Principles
  - Human Readable
  - Machine Readable
  - Plays well with default ordering

Human Readable

- Name contains information regarding the content
- Follows the concept of semantic URLs

```
# ls -la
01_marshal-data.ipynb
02_pre-dea-filtering.ipynb
03_dea-with-limma-voom.ipynb
04_explore-dea-results.ipynb
90_limma-model-term-name-fiasco.ipynb
helper01_load-counts.py
helper02_load-exp-des.py
helper03_load-focus-statinf.py
helper04_extract-and-tidy.py
```

Machine Readable

- Example of globbing to narrow file list

```
# ls
2017-07-01_SUMMERSCHOOL_Physics_01.csv
2017-07-01_SUMMERSCHOOL_Physics_02.csv
2017-07-01_SUMMERSCHOOL_Physics_03..csv
2017-07-01_SUMMERSCHOOL_Chemistry_01.csv
2017-07-01_SUMMERSCHOOL_Chemistry_02.csv
2017-07-01_SUMMERSCHOOL_Chemistry_03.csv

# ls *Physics*
2017-07-01_SUMMERSCHOOL_Physics_01.csv
2017-07-01_SUMMERSCHOOL_Physics_02.csv
2017-07-01_SUMMERSCHOOL_Physics_03..csv
```

Machine Readable

- Export Metadata from file name

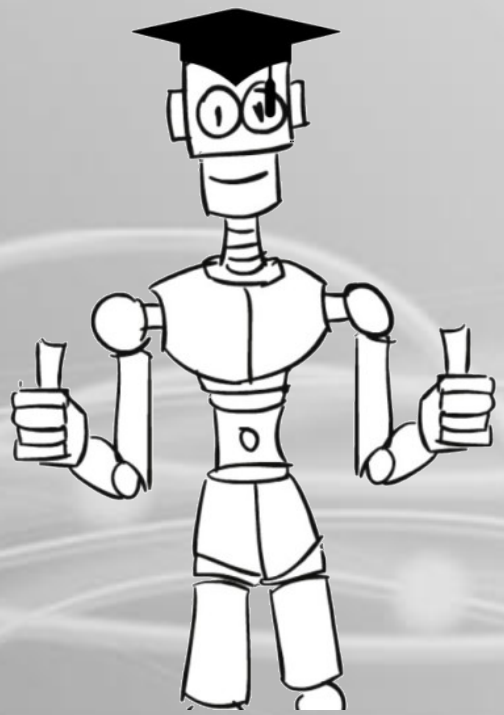
```
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
```

```
> flist <- list.files(pattern = "Plasmid") %>% head

> stringr::str_split_fixed(flist, "[_\\.]", 5)
     [,1]         [,2]              [,3]                                  [,4]  [,5]
[1,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A01" "csv"
[2,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A02" "csv"
[3,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "A03" "csv"
[4,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B01" "csv"
[5,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B02" "csv"
[6,] "2013-06-26" "BRAFWTNEGASSAY" "Plasmid-Cellline-100-1MutantFraction" "B03" "csv"
```

date assay sample set well

Machine Readable



- Regular Expression and globbing friendly
- Avoid
  - Spaces
  - Punctuation
  - Accented Characters
  - Case Sensitivity
- Easy to compute on
  - deliberate use of delimiters
- Easy to search for files later
- Easy to narrow file lists based on names
- Easy to extract info from file names

Plays Well with Default Ordering

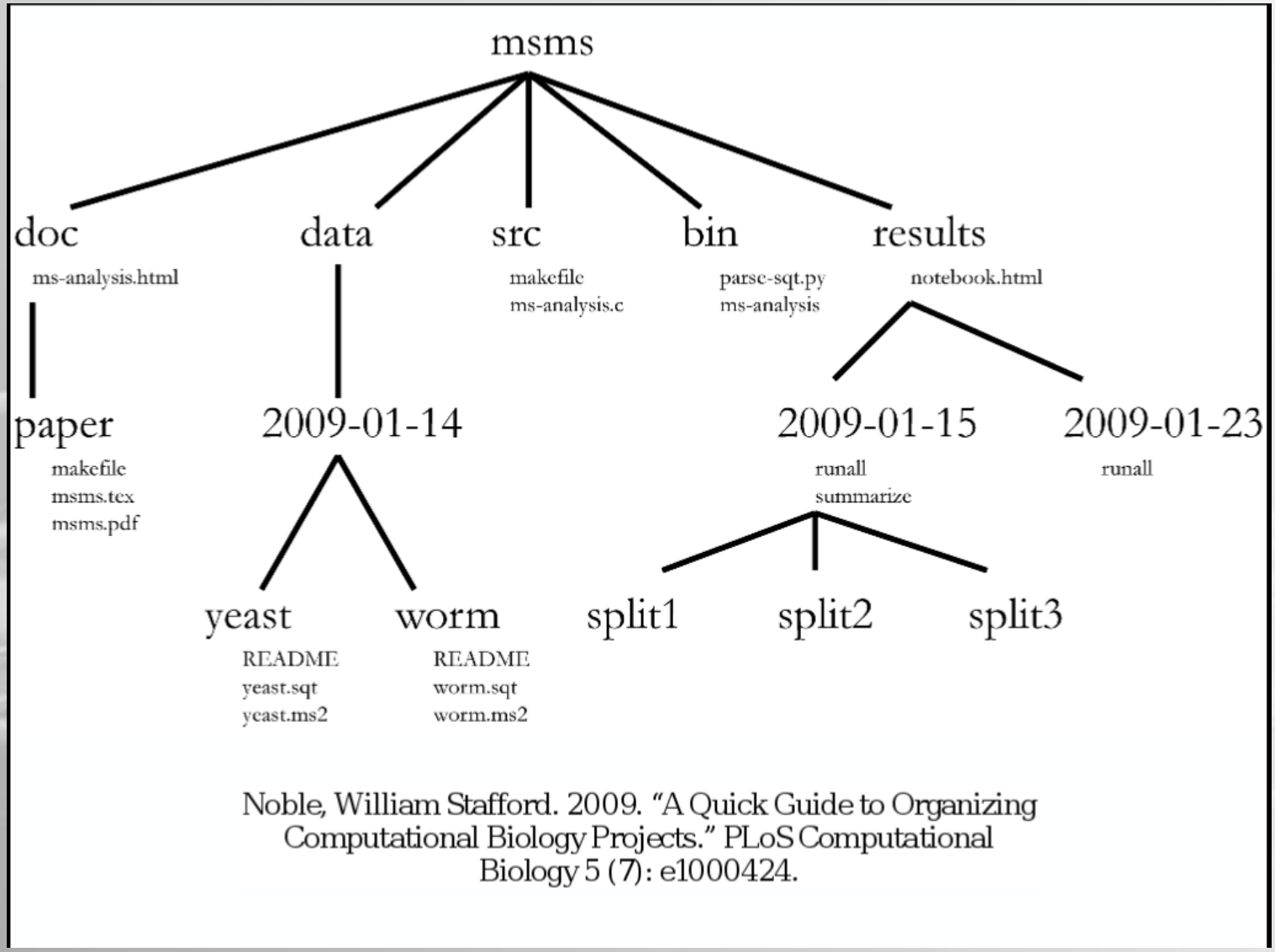

```
# Following the rules
01_marshal-data.ipynb
02_pre-dea-filtering.ipynb
03_dea-with-limma-voom.ipynb
90_limma-model-term-name-fiasco.ipynb
helper01_load-counts.py
helper02_load-exp-des.py
helper03_load-focus-statinf.py

# Without padding  Out of order
10_draft-figs-for-publication.py
1_data-cleaning.py
20_final-figs-for-publication.py
2_fit-model.py
```

Plays Well with Default Ordering

- Start with a numeric character
- Use the ISO 8601 standard for dates
  - YYYY-MM-DD
- Left pad other numbers with zeros

Project Organization



Noble, William Stafford. 2009. "A Quick Guide to Organizing Computational Biology Projects." PLoS Computational Biology 5 (7): e1000424.

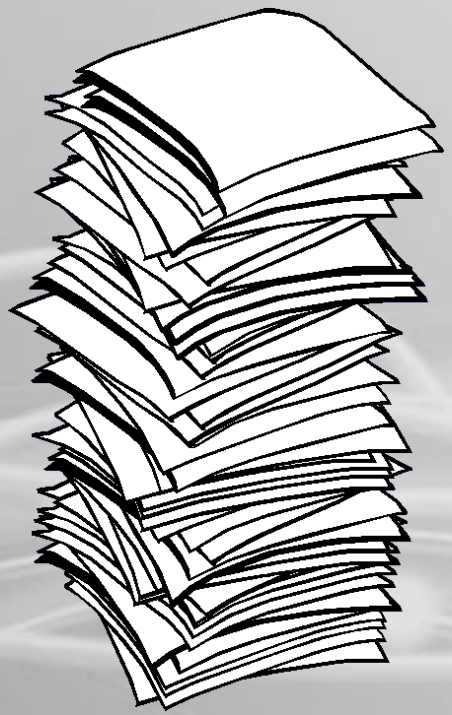
Project Organization

- Have a central **projects** directory
- Follow same folder structure for each project
- A universal strategy doesnt exist
- But its important to choose one and follow it.

General Folder Structure

```
.
|-- code
+-- data
|   +-- clean
|   +-- raw
|   +-- README.md
+-- doc
|   +-- paper
+-- README.md
+-- results
|   +-- figures
|   +-- pictures
|   +-- README.md
+-- scratch
    +-- README.md
```

Project Documentation

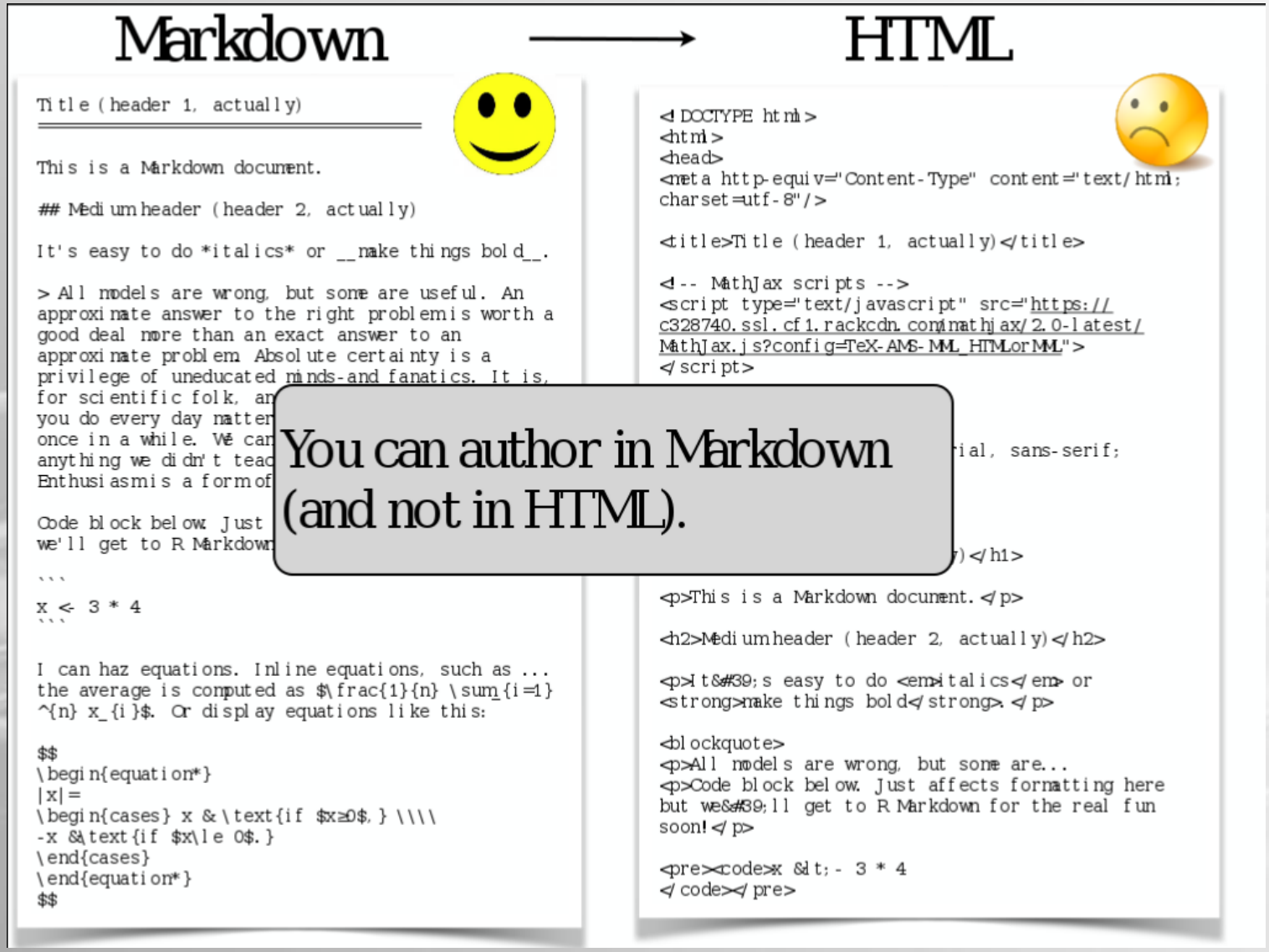


- The most neglected part
  - Especially for personal projects
- Types of Documentation
  - Manuals
  - Notes regarding the analysis procedure
  - Publication Paper
  - Presentation Slides
  - ...
- Of various formats

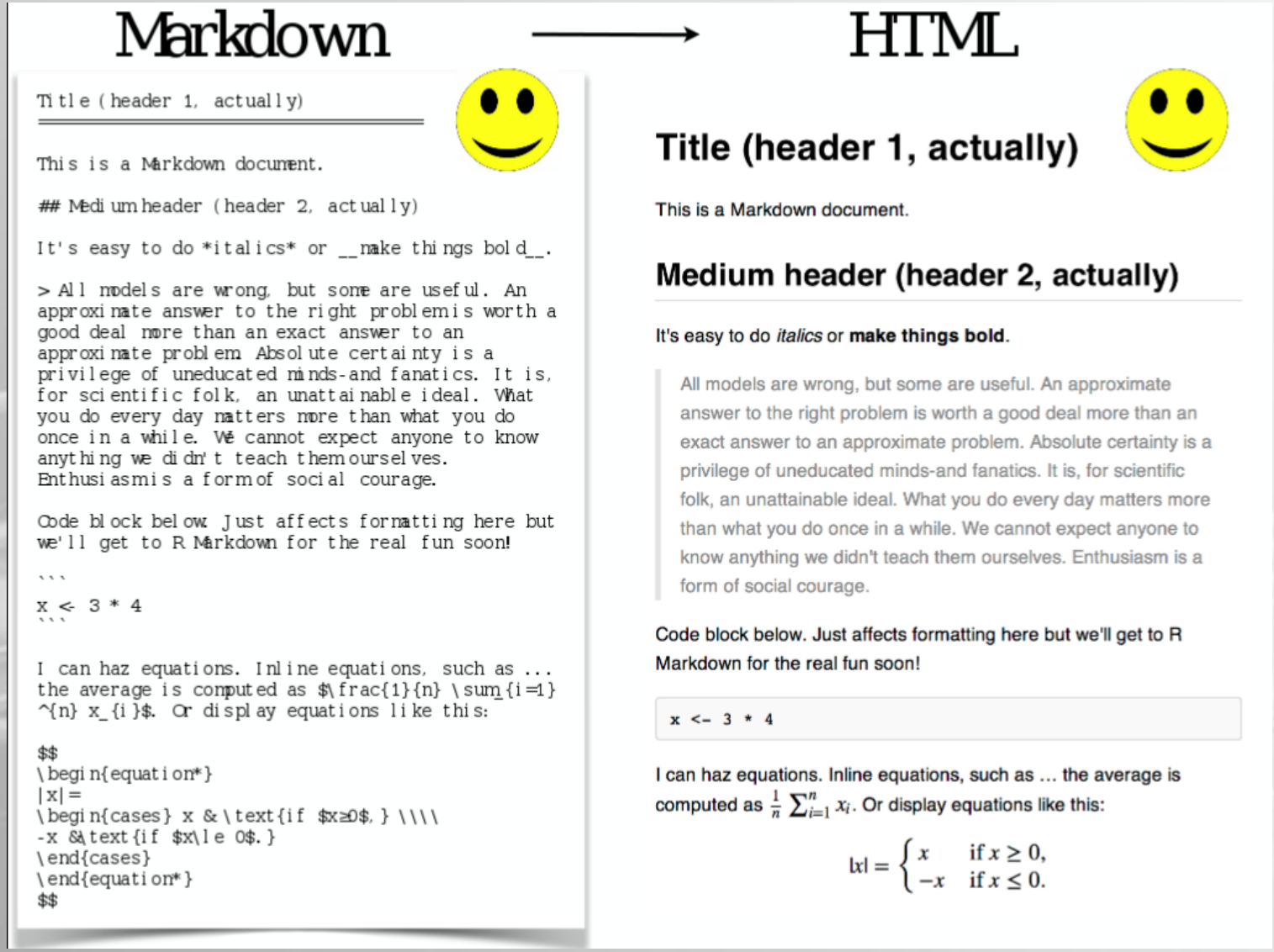
Literate Programming

- Knuth, 1984
- **Weaving**
  - Generating comprehensive document about program and its maintenance.
- **Tangling**
  - Generating machine executable code.

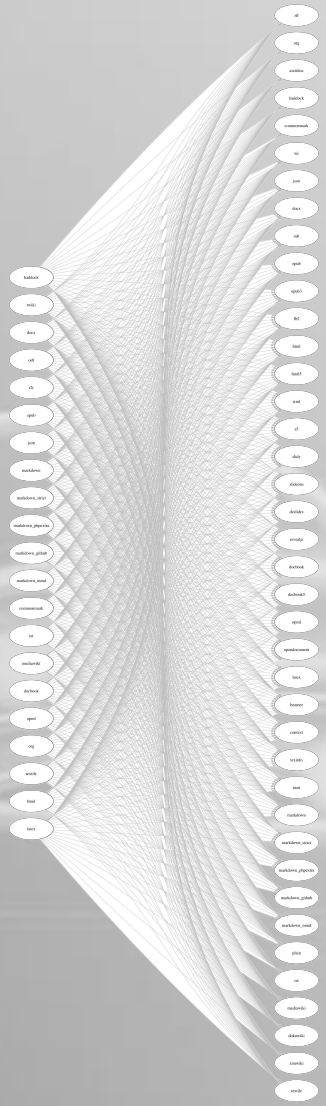
Markdown


Markdown
HTML
Title (header 1, actually)
This is a Markdown document.
Medium header (header 2, actually)
It's easy to do *italics* or __make things bold__.
> All models are wrong, but some are useful. An
approximate answer to the right problem is worth a
good deal more than an exact answer to an
approximate problem. Absolute certainty is a
privilege of uneducated minds-and fanatics. It is,
for scientific folk, an
you do every day matter
once in a while. We can
anything we didn't teac
Enthusiasm is a form of
Code block below. Just
we'll get to R Markdown
```
x <- 3 * 4
```
I can haz equations. Inline equations, such as ...
the average is computed as $\frac{1}{n} \sum_{i=1}
^{n} x_{i}$. Or display equations like this:
$$
\begin{equation*}
|x| =
\begin{cases} x &\text{if $x\ge 0$,} \\\\
-x &\text{if $x\le 0$.}
\end{cases}
\end{equation*}
$$
<!DOCTYPE html>
<html>
<head>
<meta http-equiv="Content-Type" content="text/html;
charset=utf-8" />
<title>Title (header 1, actually)</title>
<!-- MathJax scripts -->
<script type="text/javascript" src="https://
c328740.ssl.cf1.rackcdn.com/mathjax/2.0-latest/
MathJax.js?config=TeX-AMS-MML_HTMLorMML">
</script>
rial, sans-serif;
)</h1>
This is a Markdown document.
<h2>Medium header (header 2, actually)</h2>
It's easy to do italics or
make things bold.
<blockquote>
All models are wrong, but some are...
Code block below. Just affects formatting here
but we'll get to R Markdown for the real fun
soon!
<pre><code>x <- 3 * 4
</code></pre>
You can author in Markdown
(and not in HTML).

Markdown


Markdown
HTML
Title (header 1, actually)
This is a Markdown document.
Medium header (header 2, actually)
It's easy to do *italics* or __make things bold__.
> All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. What you do every day matters more than what you do once in a while. We cannot expect anyone to know anything we didn't teach them ourselves. Enthusiasm is a form of social courage.
Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!
```
x <- 3 * 4
```
I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^{n} x_{i}$. Or display equations like this:
$$
\begin{equation*}
|x| =
\begin{cases} x & \text{if $x \geq 0$,} \\
-x & \text{if $x \le 0$.}
\end{cases}
\end{equation*}
$$
Title (header 1, actually)
This is a Markdown document.
Medium header (header 2, actually)
It's easy to do italics or make things bold.
All models are wrong, but some are useful. An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem. Absolute certainty is a privilege of uneducated minds-and fanatics. It is, for scientific folk, an unattainable ideal. What you do every day matters more than what you do once in a while. We cannot expect anyone to know anything we didn't teach them ourselves. Enthusiasm is a form of social courage.
Code block below. Just affects formatting here but we'll get to R Markdown for the real fun soon!
x <- 3 * 4
I can haz equations. Inline equations, such as ... the average is computed as $\frac{1}{n} \sum_{i=1}^{n} x_{i}$. Or display equations like this:
$$|x| = \begin{cases} x & \text{if } x \geq 0, \\ -x & \text{if } x \leq 0. \end{cases}$$

Pandoc



- Universal Markup Converter
- Word processor formats
  - Microsoft Word docx
  - OpenOffice/LibreOffice ODT
- Ebooks
- TeX formats
  - LaTeX
  - LaTeX Beamer slides
- PDF via LaTeX

README

- Include README files
- The top level should contain
  - Project name
  - Date
  - Maintainer's contact info
  - Data Origin
  - 3-4 sentences about the goal of the project

Conclusion

Organization & Documentation

- Often neglected
- Easy to implement
- Benefits accumulate over time
- Consistency is Important
- Jupyter Notebooks is an ideal environment for Literate Programming

Questions