Organizing computational research projects

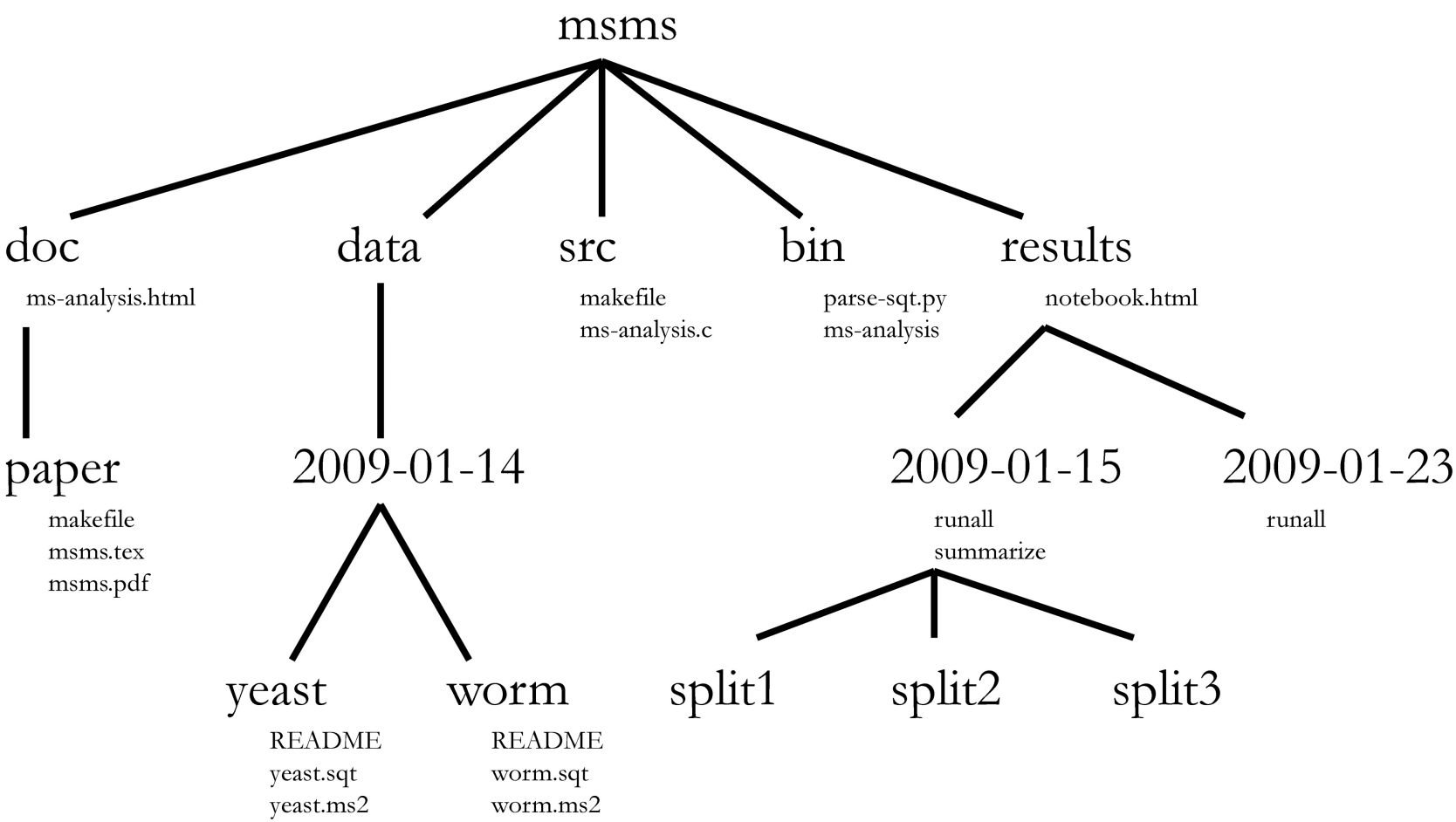
Toby Dylan Hocking
tdhock5@gmail.com

20 Feb 2014

Motivation: reproducible computations

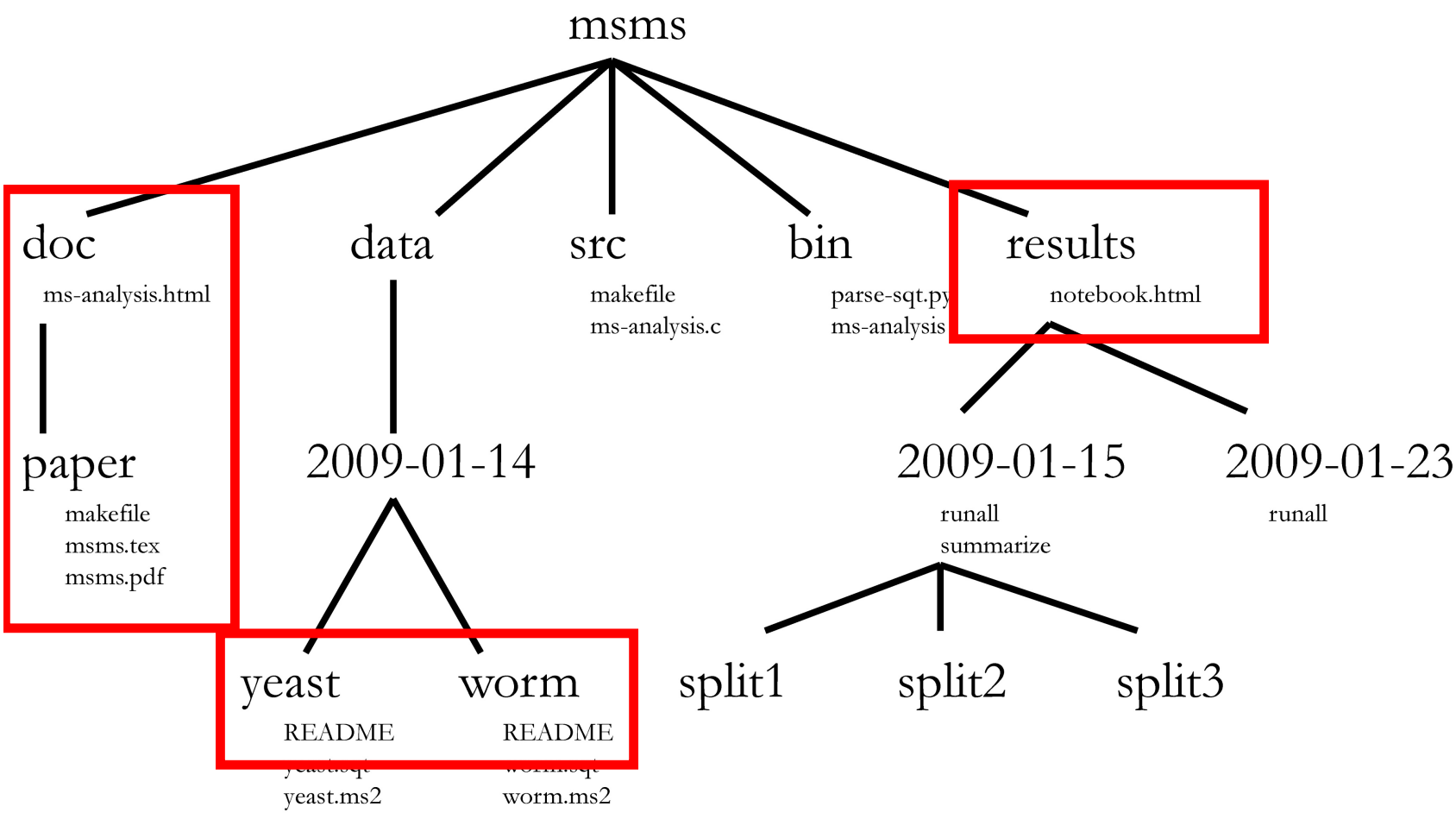
- Lab notebooks essential for reproducible experiments.
- Lab notebooks skills are taught in school.
- XXXXX essential for reproducible computations.
- XXXXX skills are NOT taught.

A quick guide to organizing computational biology projects



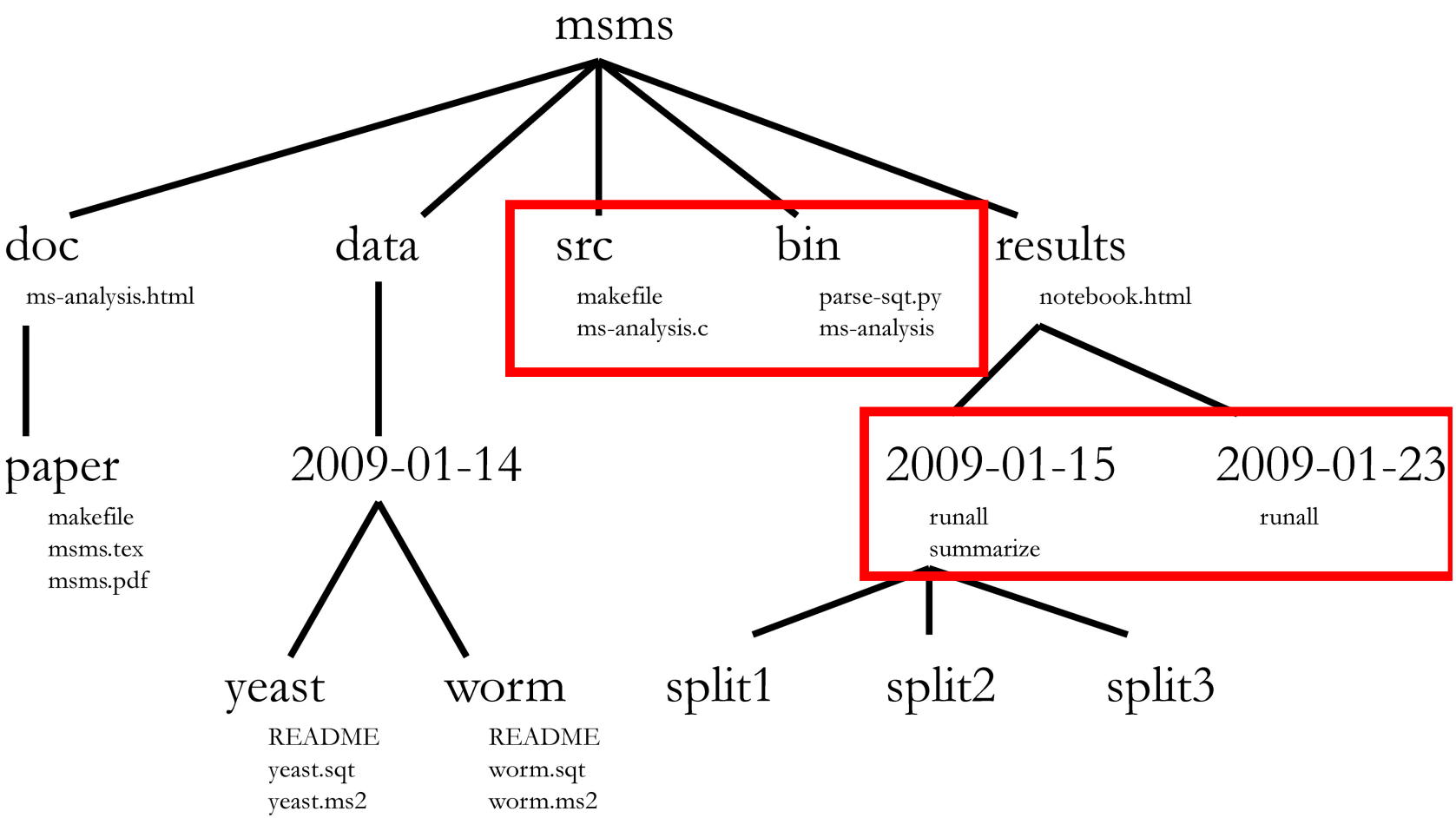
Source: William Stafford Noble, PLoS Comp Bio 2009.

Human-readable files (notes to future self)



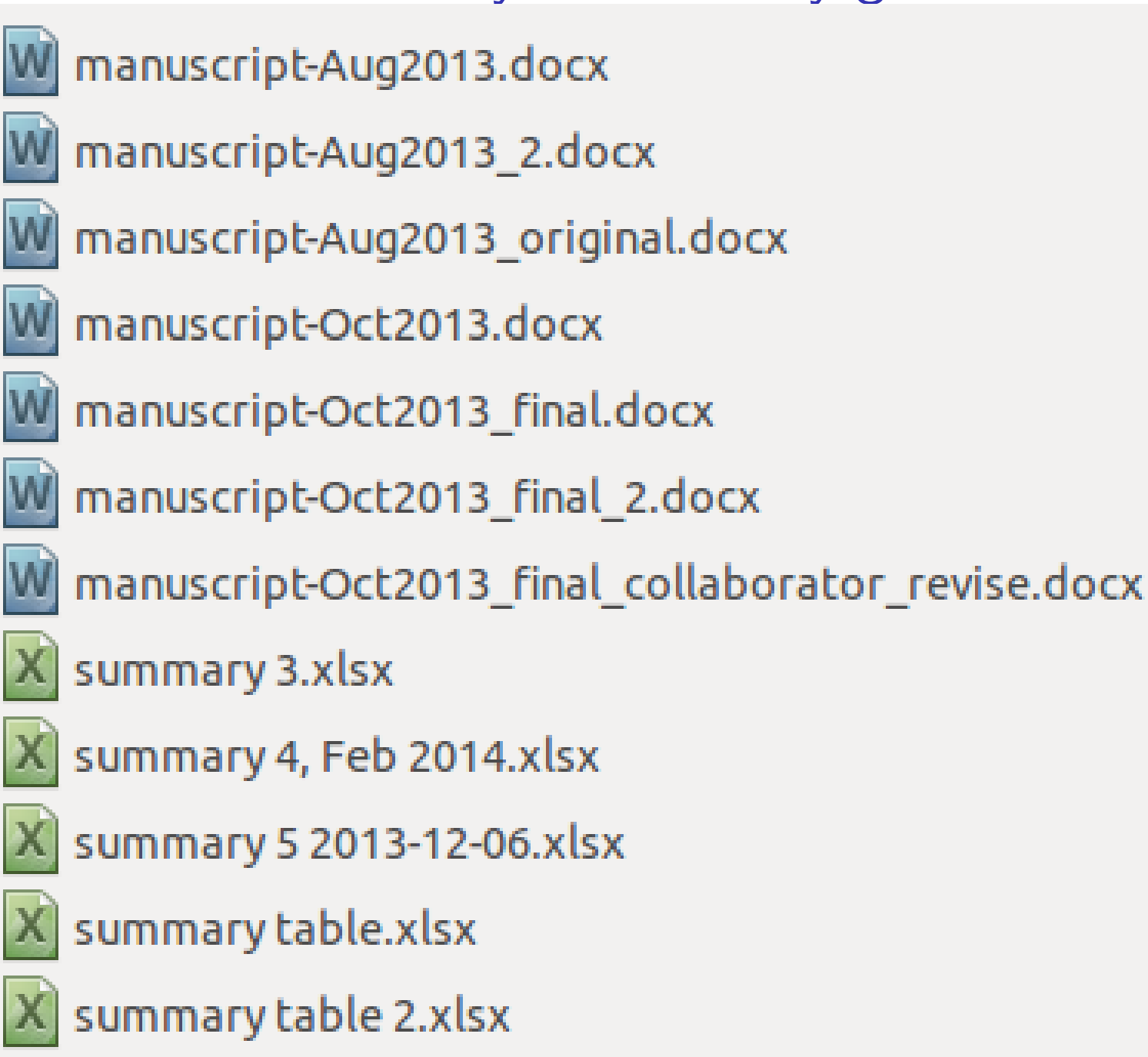
Source: William Stafford Noble, PLoS Comp Bio 2009.

Computer-readable files (reproducible computations)



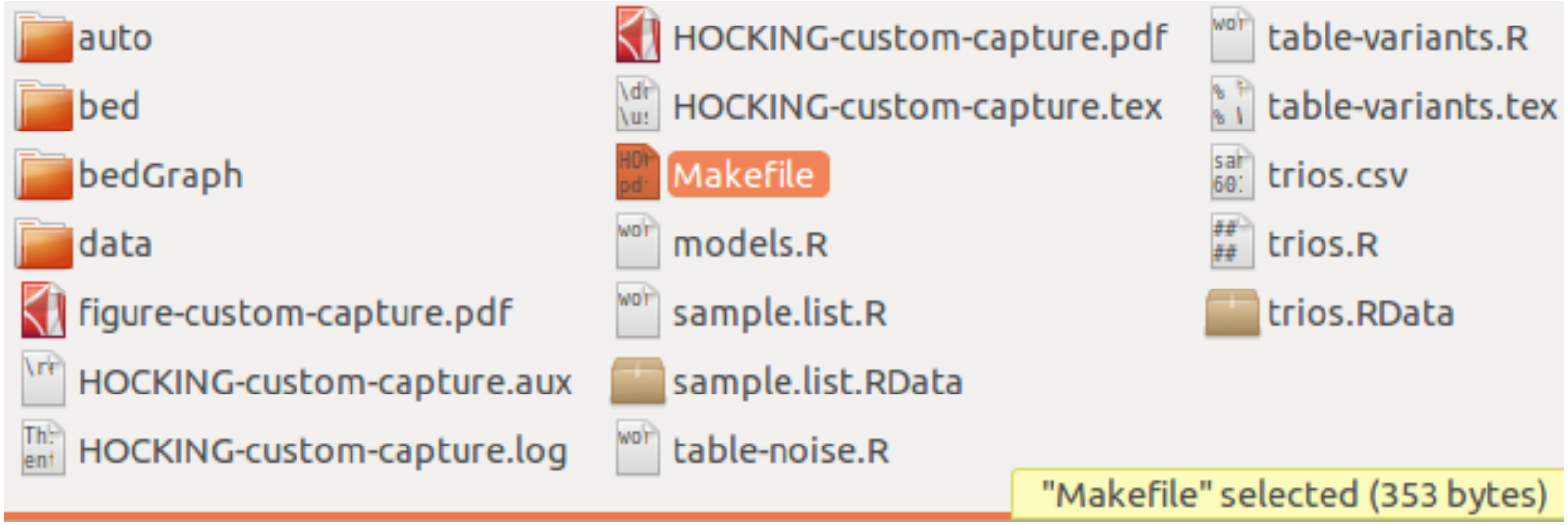
Source: William Stafford Noble, PLoS Comp Bio 2009.

Without version control you can easily get



With version control you get

- Backup.
- Go back in time.
- Collaboration.



I use a Makefile for each of my projects

Write relations between files as rules in a Makefile:

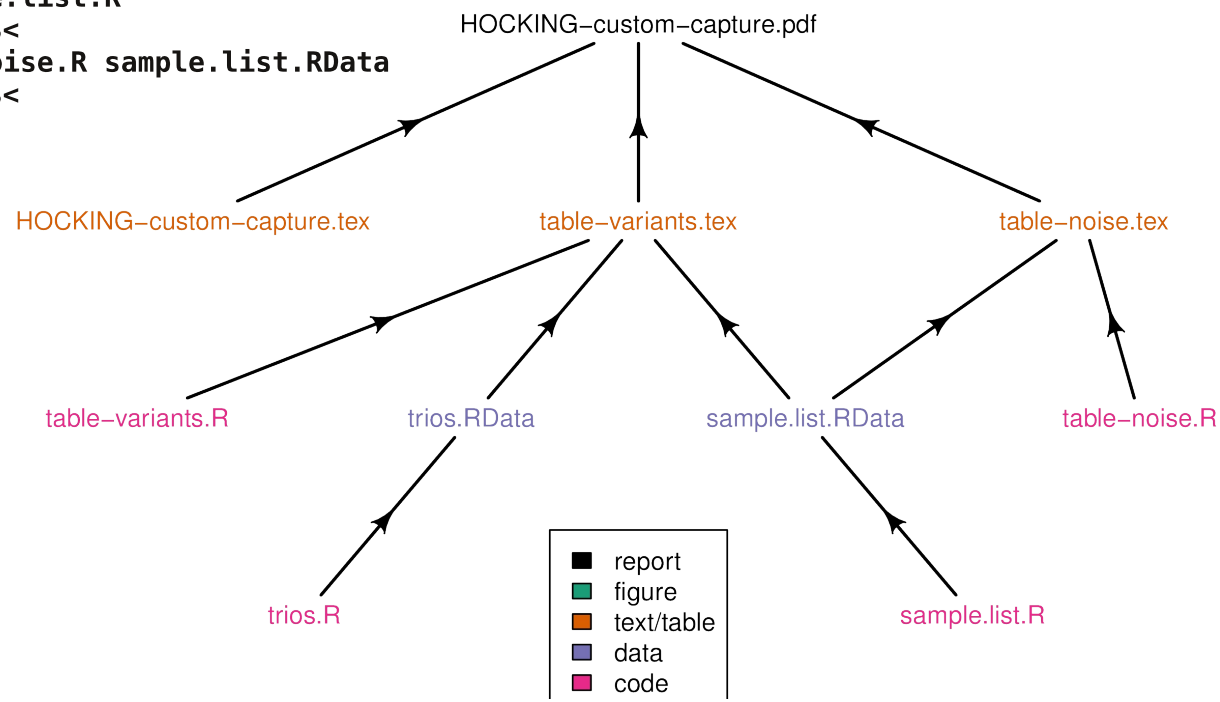
- For each computation/figure/table, save 1 file (the target).
- Write what files are required (the prerequisites).
- Write a command line to do it (the recipe).

```
thocking@silene$ make figure-3.png
thocking@silene$ make table-1.tex
thocking@silene$ make notebook.pdf
thokcing@silene$ make
```

Note: last line means make first target in the Makefile.

A Makefile records how to compute each project file

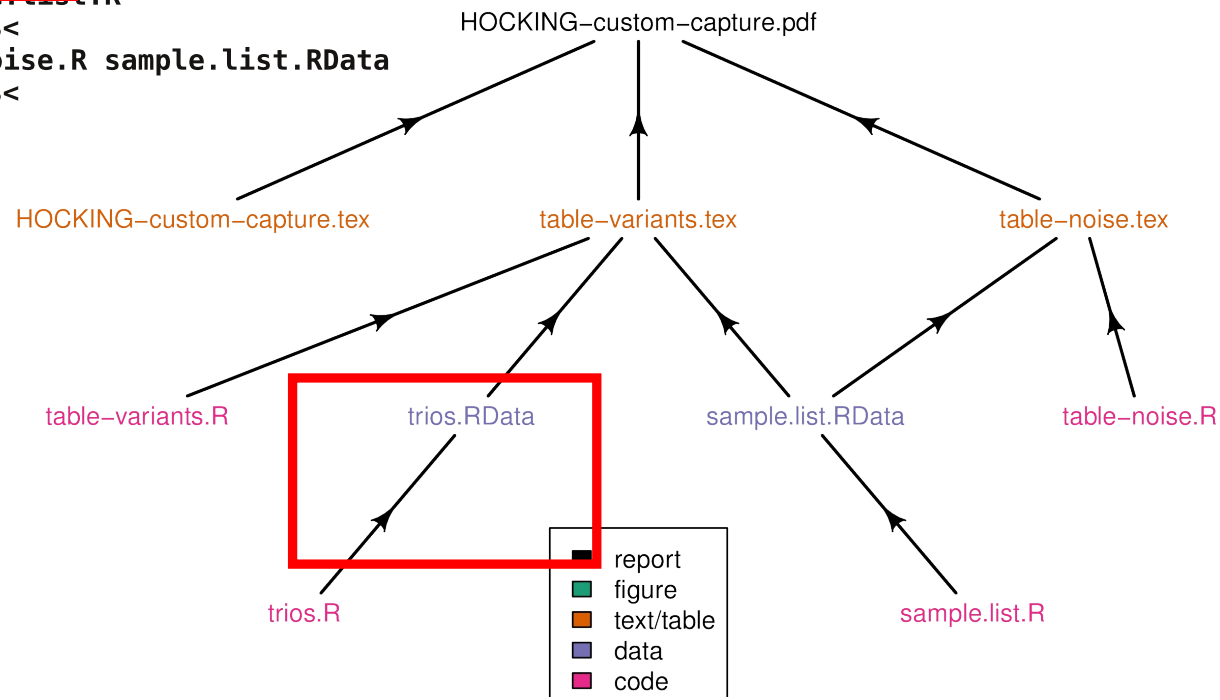
```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
	pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
	R --no-save < $<
trios.RData: trios.R
	R --no-save < $<
sample.list.RData: sample.list.R
	R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
	R --no-save < $<
```



A Makefile contains a rule for each computed file

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```

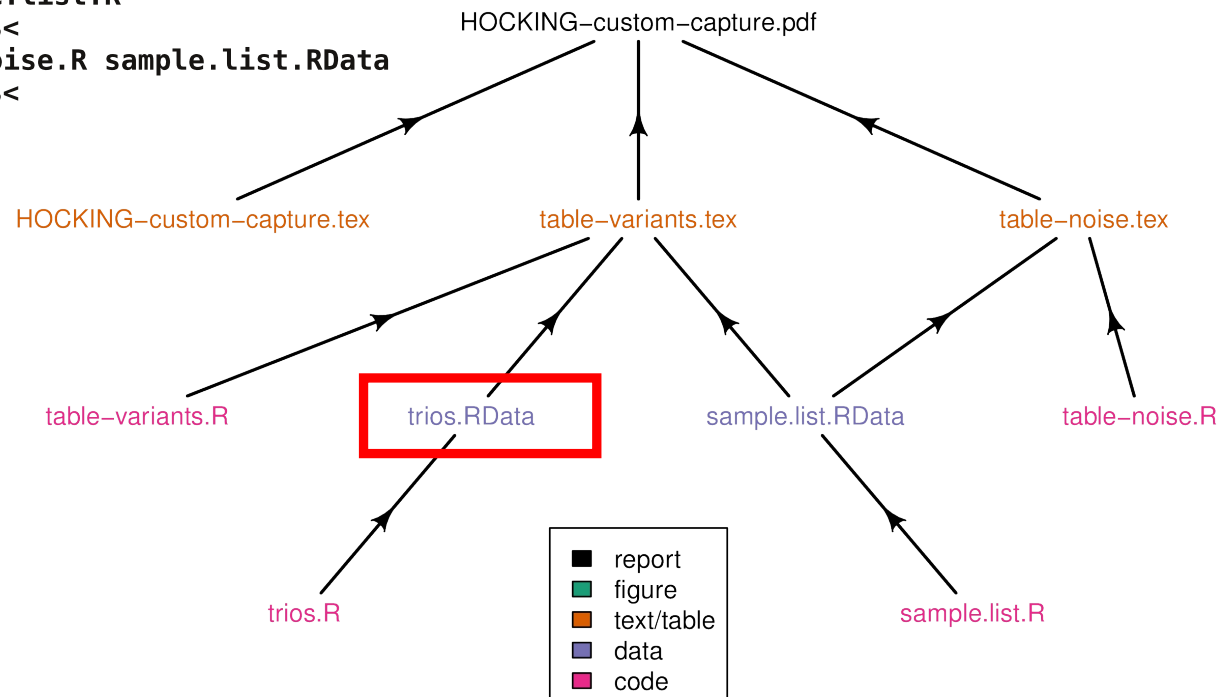
Rule



The computed file is called the target

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```

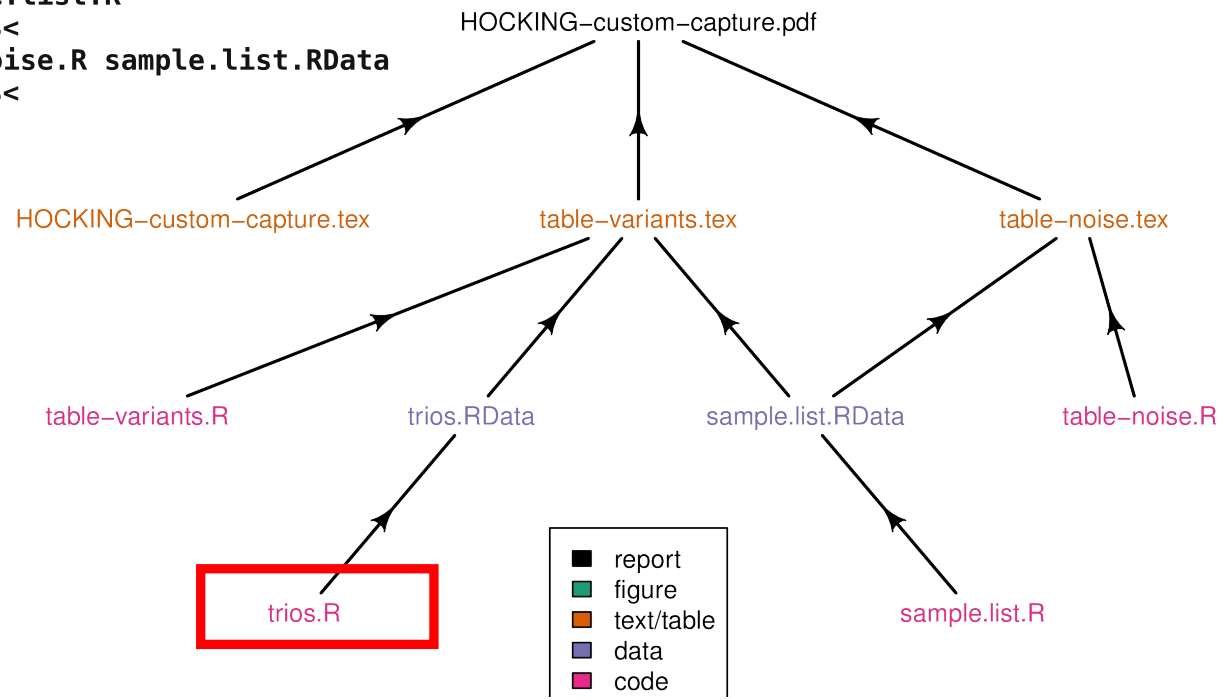
Target



The input files are called prerequisities

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```

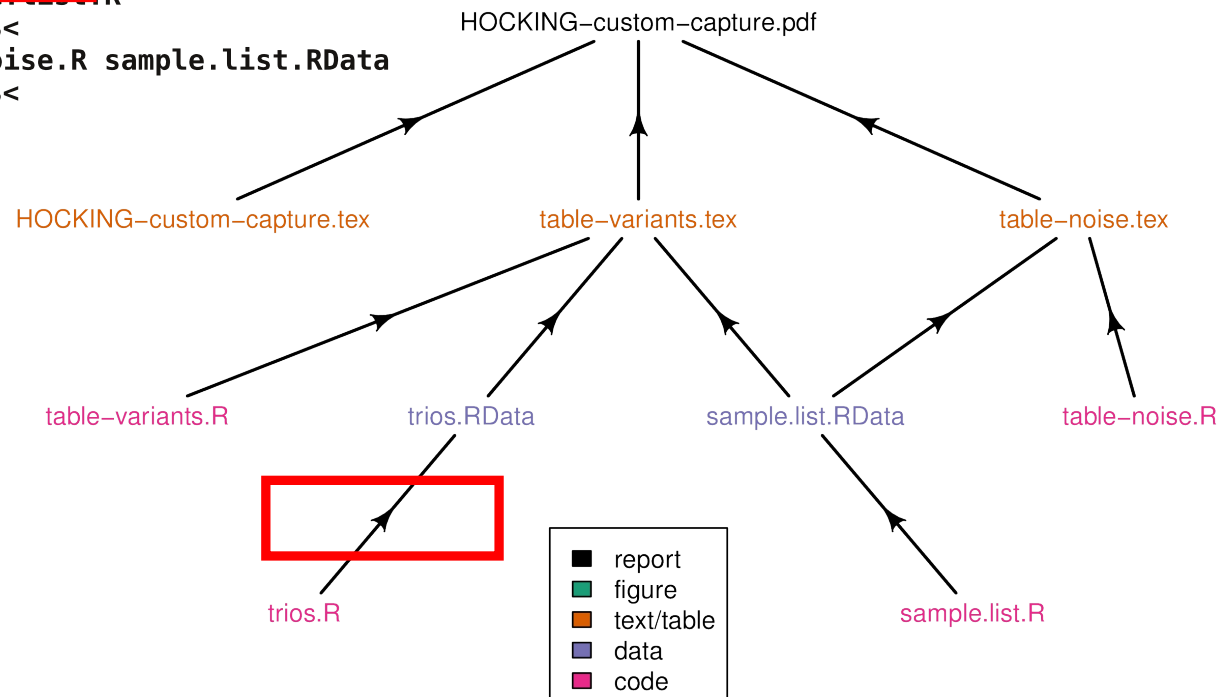
Prerequisites



The recipe is the command line used to compute

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```

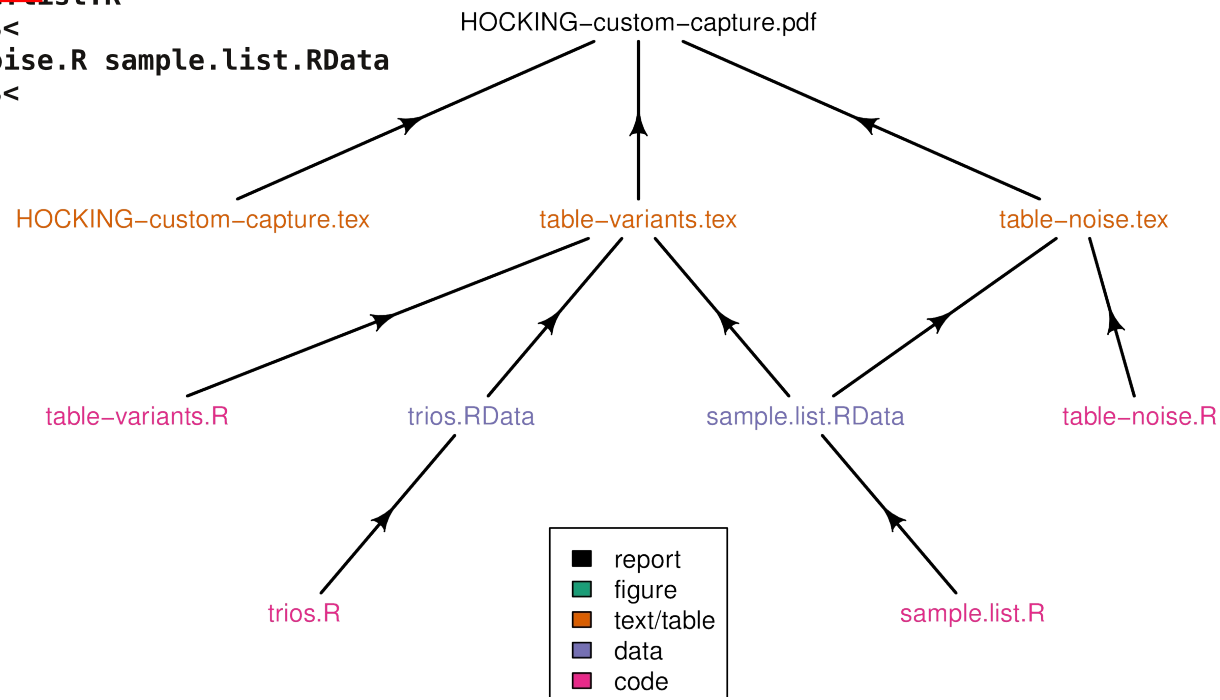
Recipe



Make provides abbreviations to avoid repetition

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```

$< means the first prerequisite



Read slow CSV data and save in fast RData format

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
	pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
	R --no-save < $<
trios.RData: trios.R
	R --no-save < $<
sample.list.RData: sample.list.R
	R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
	R --no-save < $<
```

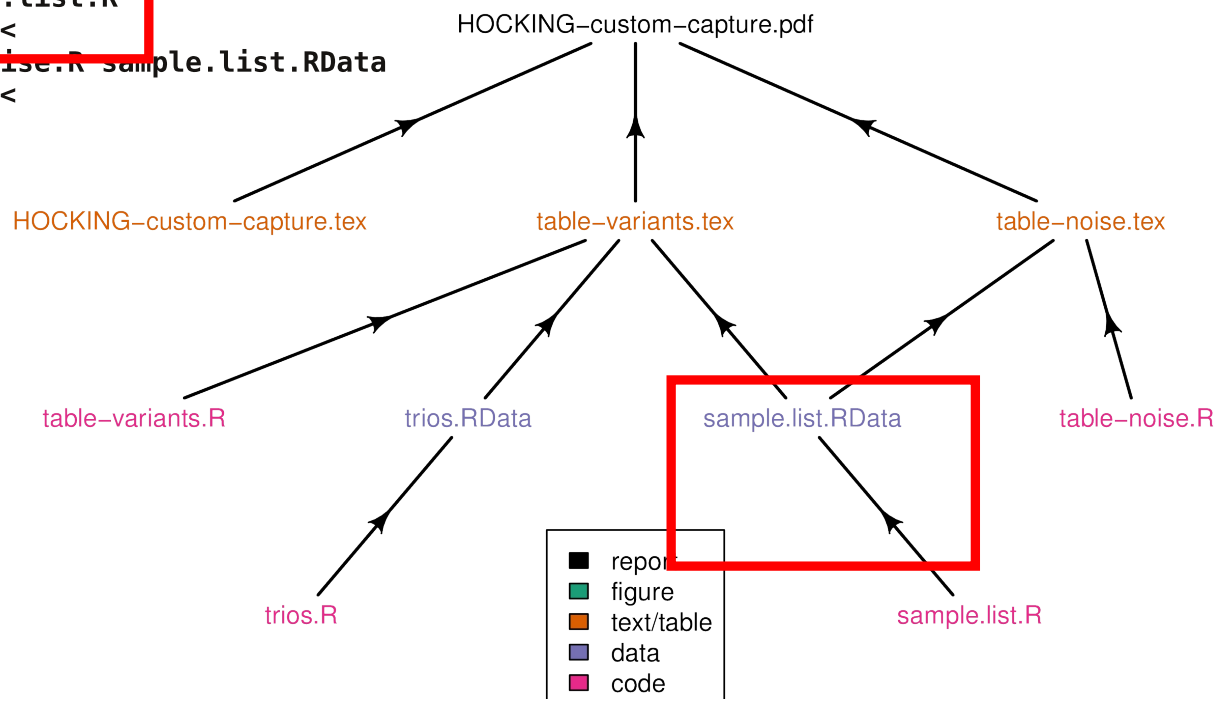


Table of genes that were excluded as noise

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
	pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
	R --no-save < $<
trios.RData: trios.R
	R --no-save < $<
sample.list.RData: sample.list.R
	R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
	R --no-save < $<
```

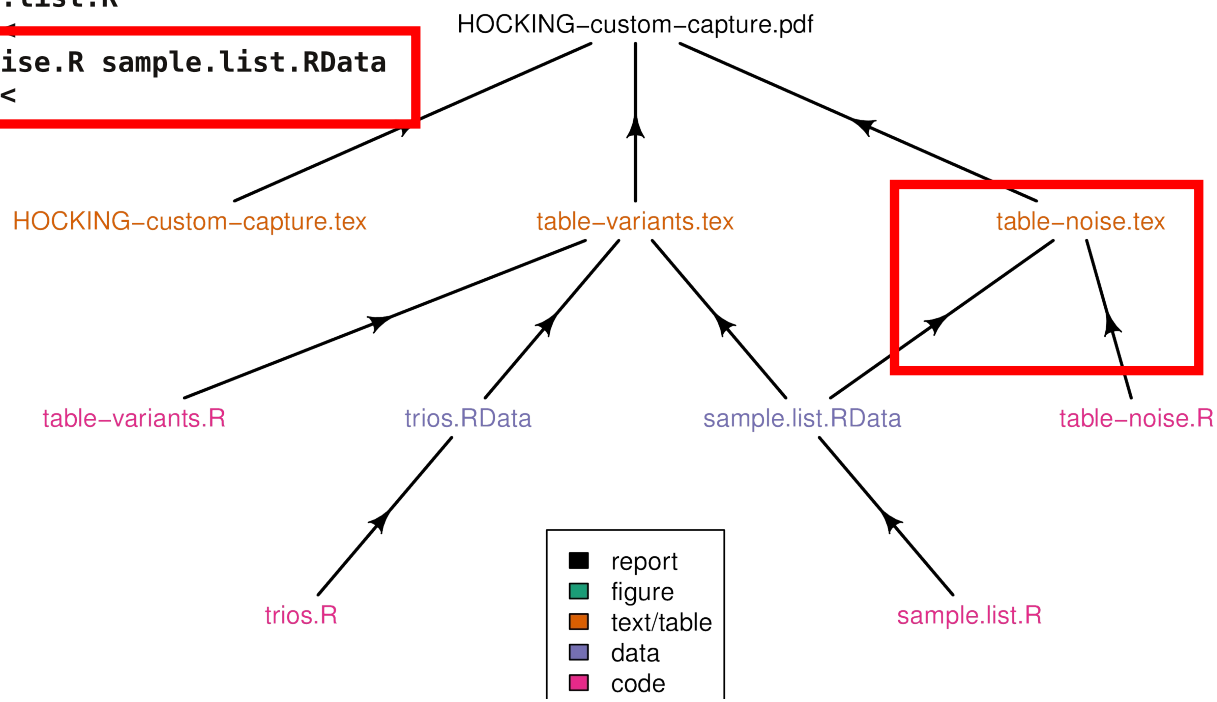
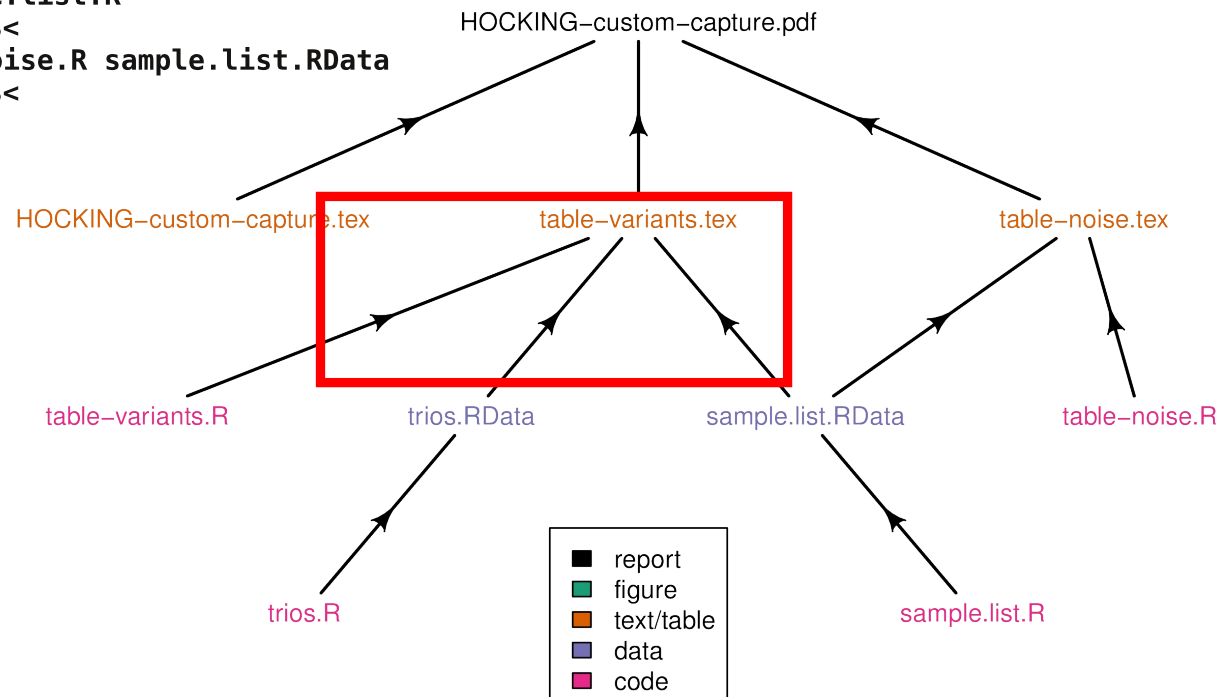


Table of detected variants for every sample

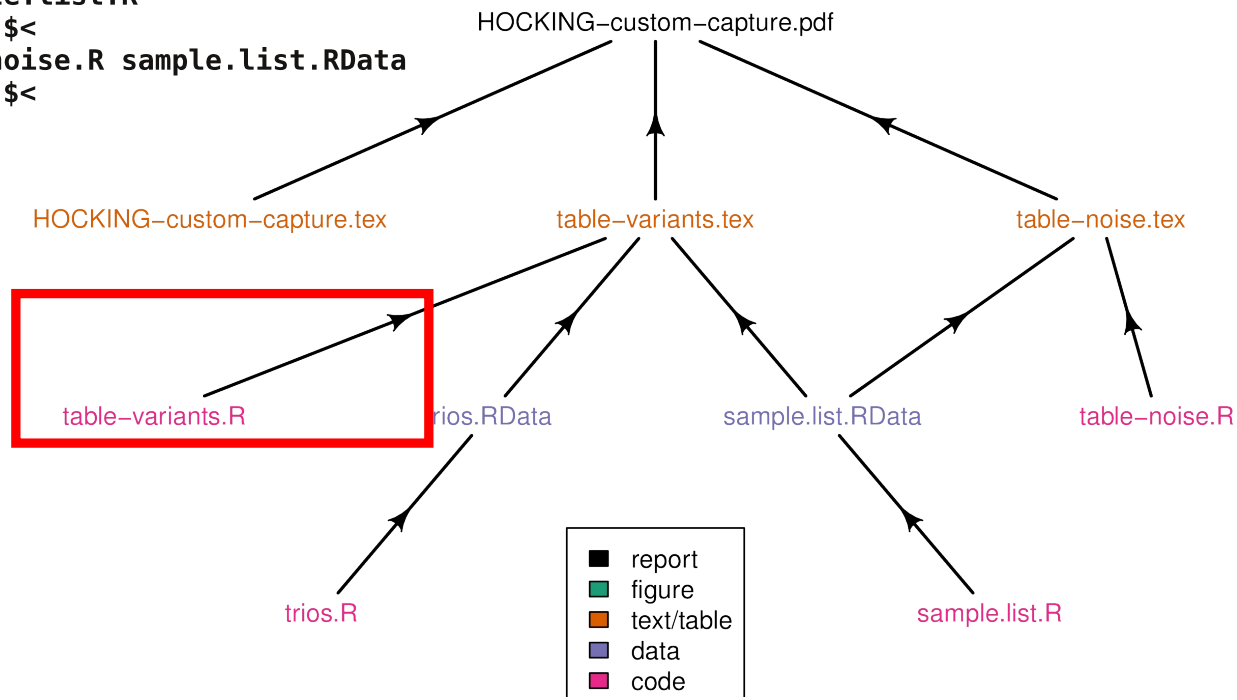
```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```



A recipe can produce more files if necessary

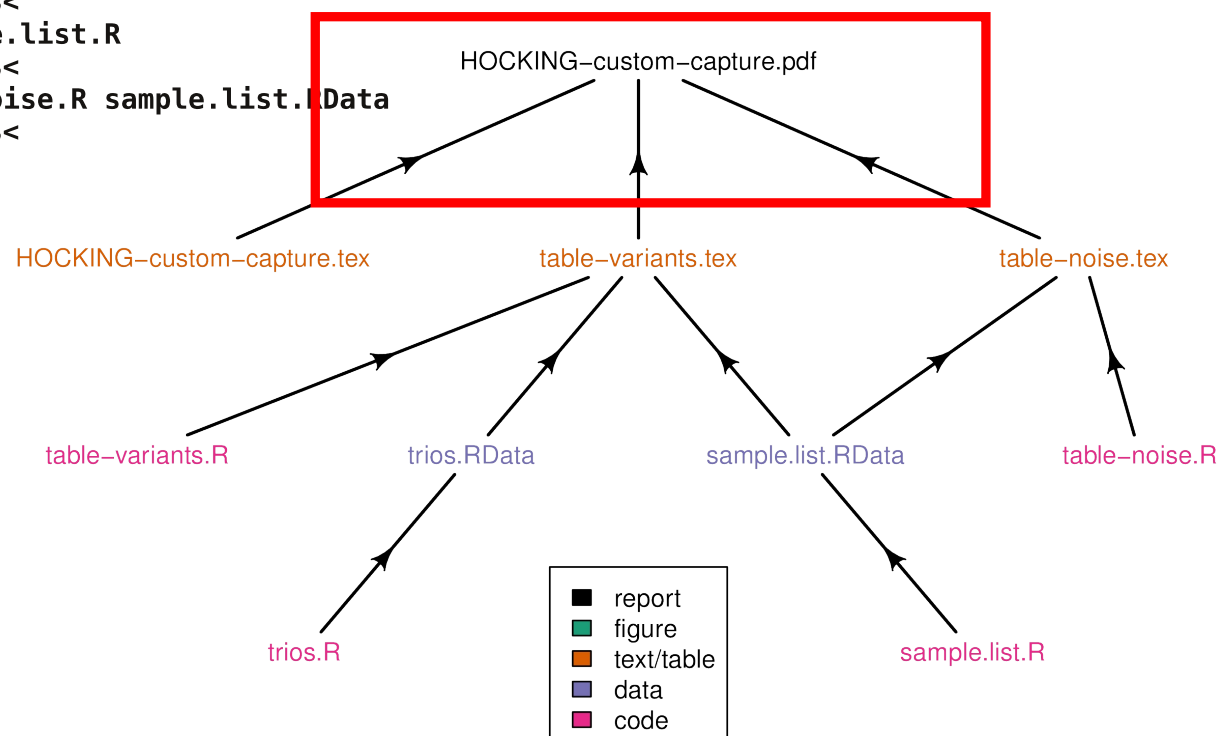
```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
        pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
        R --no-save < $<
trios.RData: trios.R
        R --no-save < $<
sample.list.RData: sample.list.R
        R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
        R --no-save < $<
```

Side effect: save a bed.gz file for every sample.



The final report PDF with computed data tables

```
HOCKING-custom-capture.pdf: HOCKING-custom-capture.tex table-variants.tex table-noise.tex
	pdflatex HOCKING-custom-capture
table-variants.tex: table-variants.R sample.list.RData trios.RData
	R --no-save < $<
trios.RData: trios.R
	R --no-save < $<
sample.list.RData: sample.list.R
	R --no-save < $<
table-noise.tex: table-noise.R sample.list.RData
	R --no-save < $<
```



Limitations

- You need to write R/LaTeX code (complicated).
- Really big files/calculations?
- Where to record software versions?
  `install.packages("changepoint")` vs
  `works_with_R("3.0.2", changepoint="1.1.1")`