LEIDEN UNIVERSITY MEDICAL CENTER

Good Research Practice for computational data analysis

(part one)

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics

Introduction

Good research practices

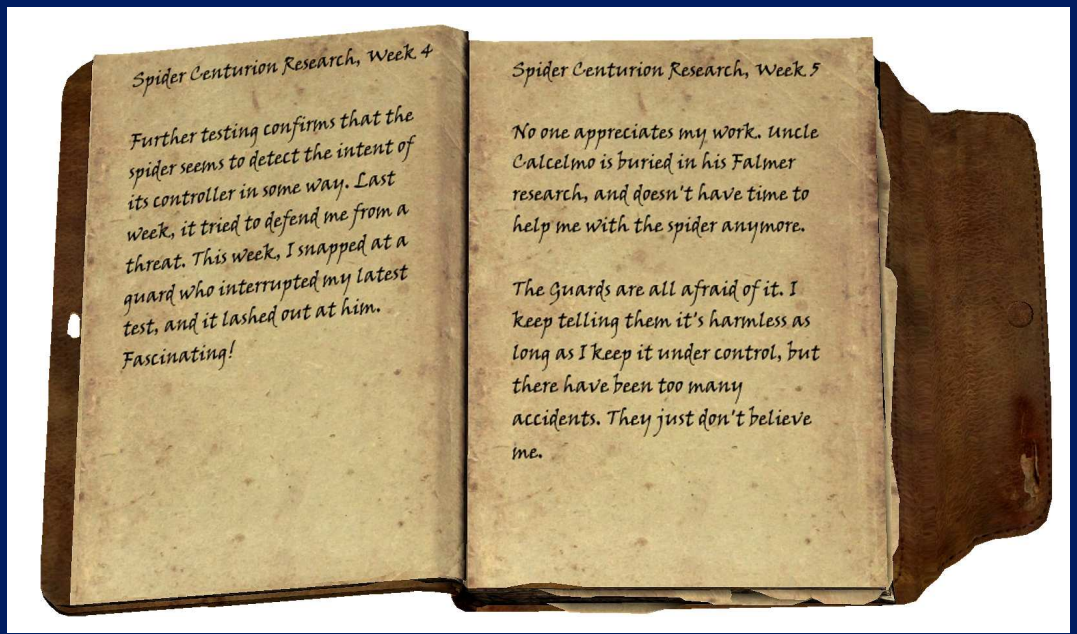


Figure 1: Lab journal.

In the lab: lab journals, standard operating procedures, etc.

Introduction

Current practice

Pipeline development and data analysis have become increasingly important.

Surprisingly, there are no standard operating procedures for data analysis.

Introduction

Current practice

Pipeline development and data analysis have become increasingly important.

Surprisingly, there are no standard operating procedures for data analysis.

Example.

- Person A creates a program.
- Person B modifies the program.

Introduction

Current practice

Pipeline development and data analysis have become increasingly important.

Surprisingly, there are no standard operating procedures for data analysis.

Example.

- Person A creates a program.
- Person B modifies the program.
- Person B received PhD.
- Person A mails the program to person C.
- Person C can not reproduce the results of person B.

Introduction

Current practice

Figure 2: Final resting place for data.

Even worse, the results disappear into a drawer.

Introduction

Current practice

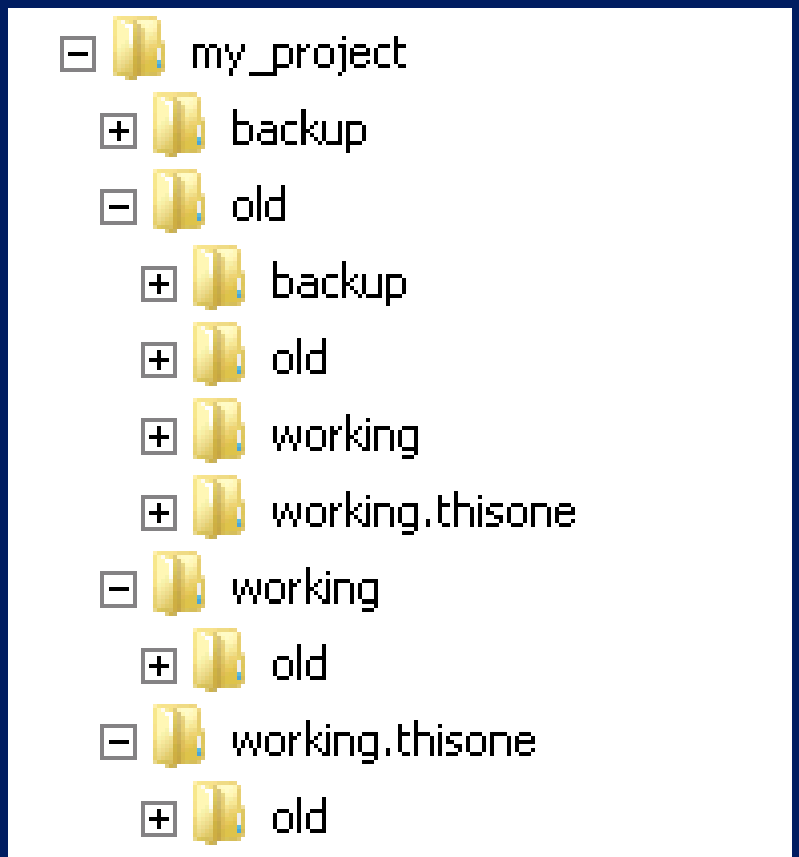


Figure 3: "I have my own system."

Introduction

Current practice in our departement

Luckily, we have tackled these problems already.

Investments in infrastructure:

- Centralised storage.
  - Backup.
- Centralised computing.

Introduction

Current practice in our departement

Luckily, we have tackled these problems already.

Investments in infrastructure:

- Centralised storage.
  - Backup.
- Centralised computing.

Adoption of *software engineering* solutions.

We want to share this knowledge with our colleagues in other departments.

Introduction

"Beleidsinitiatief" (structural funding)

Proposal for a change in policy.

- The department Human Genetics has some nice solutions available.
- We always keep the big picture in mind.
- The LUMC as a whole can benefit.

Proposal was granted.

- One FTE structural.
- One additional FTE for one year.

M. Roos, P.A.C. 't Hoen

Introduction

"Beleidsinitiatief" (structural funding)

The proposal consists of two main parts.

Work package 1.

- Version control.
- Interactive computational environment.

Work package 2.

- Data stewardship.

Introduction

"Beleidsinitiatief" (structural funding)

The proposal consists of two main parts.

Work package 1.

- Version control.
- Interactive computational environment.

Work package 2.

- Data stewardship.

Work package 1 and 2.

- Education.

Version control

Git

The management of changes to documents, computer programs, large web sites, and other collections of information.
— Wikipedia.

http://www.git-scm.com/

https://github.com/

Version control

Git

The management of changes to documents, computer programs, large web sites, and other collections of information. — Wikipedia.

General features:

- Keeping track of your files in an orderly manner.
  - Hiding old versions.
  - Recording who made changes and when.
- Enables collaboration.

http://www.git-scm.com/

https://github.com/

Version control

Why version control?

For a single user:

- Revert files to a previous state.
- Revert the entire project back to a previous state.
- Review changes made over time.
- Backup.

Version control

Why version control?

For a single user:

- Revert files to a previous state.
- Revert the entire project back to a previous state.
- Review changes made over time.
- Backup.

For multiple users:

- A reliable way to share files between people/computers.
- Allow multiple people working on the same project at the same time.
- Conflict resolution.
- See who made which changes at which time.

Version control

Collaboration

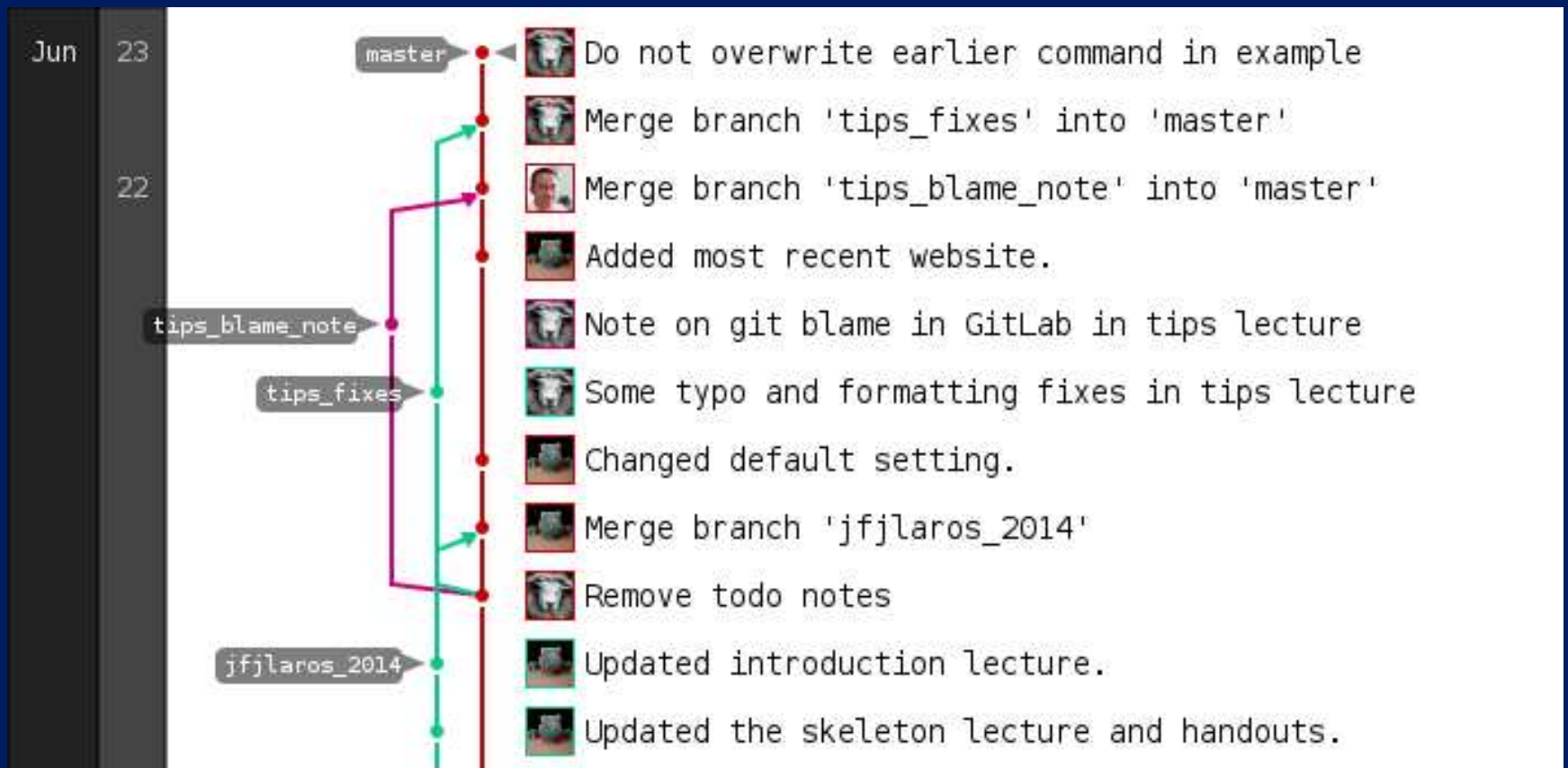


Figure 4: Collaboration with many people.

Version control

Tracking of changes

```
257  295          self._set_field(member, 1)
258  296          self._set_field(member, 1, 'FLAGS_2', _bit)
259       -       self._set_field(member, 26)
     297  +       self._set_field(member, 4)
     298  +
     299  +       self._parse_crossover(member['ID'])
     300  +
260  301          self._set_field(member, 1, 'FLAGS_3', _bit)
261  302          self._set_field(member, 205)
262  303
```

Figure 5: Compare versions.

Not limited to the previous version and the latest one.

- Different authors.
- Any two versions.

Version control

Documentation

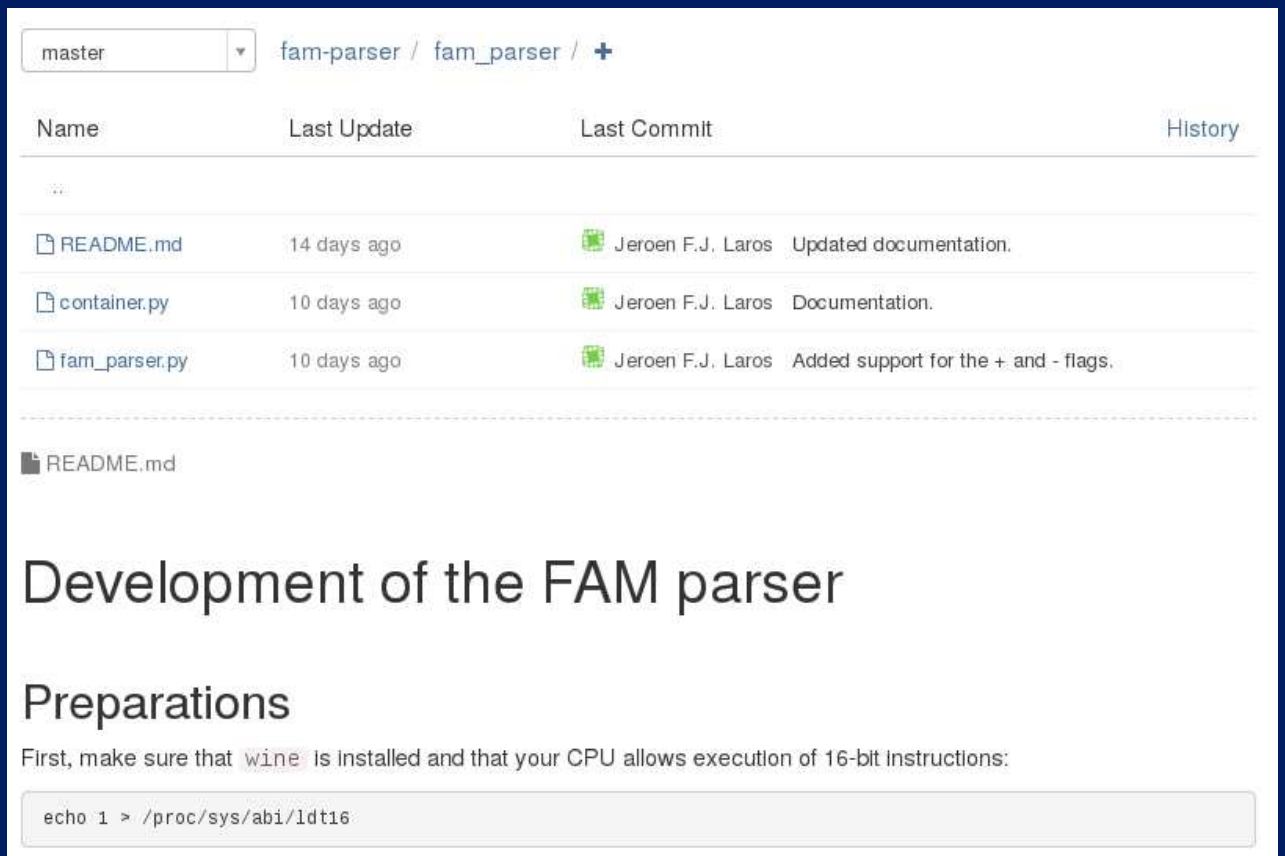


Figure 6: Documentation and programs in one place.

Version control

Projects

Most of us work on multiple projects with multiple people.

That is why is is convenient to:

- Have everything in one place.
  - Data.
  - Code.
  - Documentation.
- Have the same structure for all projects.

Version control

Projects

Most of us work on multiple projects with multiple people.

That is why is is convenient to:

- Have everything in one place.
  - Data.
  - Code.
  - Documentation.
- Have the same structure for all projects.

This also makes transferring projects easier.

Version control

Projects

Ideally, every directory in the project has a README file.

directory	description
data	Raw immutable data.
doc	Sample sheets, papers, etc.
src	Programs specific for this project.
analysis	Results.

Table 1: Project layout.

This structure is used by the LGTC, later adopted by SASC.

https://git.lumc.nl/lgtc-bioinformatics/project-skeleton/

Exploratory data analysis

Interactive computational environments

Combine code execution, text, mathematics, plots and rich media into a single document.

Ideal for exploration of data.

- Documentation and code are interwoven.
- Results are displayed inline.
- Web based.
- Versions.

http://ipython.org/notebook.html

Exploratory data analysis

Interactive computational environments

Combine code execution, text, mathematics, plots and rich media into a single document.

Ideal for exploration of data.

- Documentation and code are interwoven.
- Results are displayed inline.
- Web based.
- Versions.

Integration with GitLab.

http://ipython.org/notebook.html

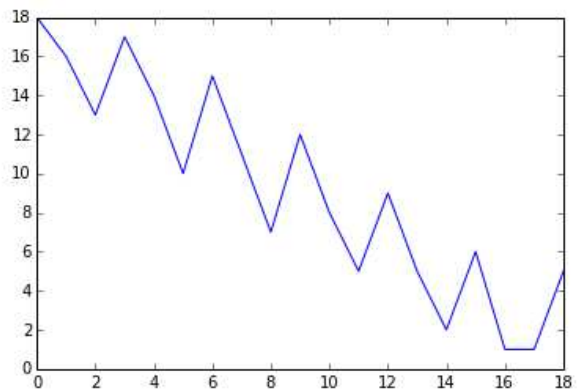
Exploratory data analysis

iPython notebook

Ranked data

In [7]:
```
def ranked(data):
    rank = {v: k for k, v in enumerate(sorted(data))}
    return map(lambda x: rank[x], data)

plot(ranked(data));
```



By using ranks insead of values, we circumvent a lot of scaling isues. Notice that this plot looks a lot like the logarithmic plot above, except that the "dampening effect" is gone.

Figure 7: iPython notebook.

Education

Overview

We offer the following courses:

- Introduction to clusters.
- Introduction to Git.
- Python programming.

Apart from this, we:

- Teach people running our pipelines (SASC).
- Answer questions about the infrastructure.
- Keep documentation up to date.

https://humgenprojects.lumc.nl/

Education

Introduction course clusters

Half day course.

An overview of the available infrastructure.

In particular:

- Connecting to a cluster.
- Using the Sun Grid Engine.
- Do's and don'ts.
- Makefiles.

M. Villerius

Education

Git

Everyone in the Bioinformatics field:

- Software development.
- Project management.
- Collaboration.

M. Vermaat, W. Arindrarto

https://humgenprojects.lumc.nl/trac/humgenprojects/wiki/git

Education

Git

Everyone in the Bioinformatics field:

- Software development.
- Project management.
- Collaboration.

Topics:

- Git Basics
- Branching
- Remotes
- Project skeleton / git annex

M. Vermaat, W. Arindrarto

https://humgenprojects.lumc.nl/trac/humgenprojects/wiki/git

Education

Programming in Python

Four day course.

- Python basics.
- Standard data structures.
- Working with NumPy arrays.
- Plotting with matplotlib.
- Object-oriented programming.
- The Biopython library.

M. Vermaat, W. Arindrato, Z. Tatum, W.Y. Leung

https://humgenprojects.lumc.nl/trac/programming-course

Part two

Data stewardship

This topic will be covered by Marco Roos at a later time.

Questions?

Acknowledgements:

Marco Roos
Peter-Bram 't Hoen
Martijn Vermaat
Zuotian Tatum
Wibowo Arindrarto
Wai Yi Leung
Michel Villerius
Madeleine Nivard
Silvère van der Maarel