Training on Web Scraping Prices for CPI

Reproducible Analytical Pipelines

Christophe Bontemps & Serge Goussev

Fundamental Principles of Official Statistics

- Clear mention of the **processes** used to produce statistics

Fundamental Principles of Official Statistics*

United Nations

For more information: unstats.un.org

The General Assembly,

Recalling recent resolutions[1] of the General Assembly and the Economic and Social Council highlighting the fundamental importance of official statistics for the national and global development agenda,

Bearing in mind the critical role of high-quality official statistical information in analysis and informed policy decision-making in support of sustainable development, peace and security, as well as for mutual knowledge and trade among the States and peoples of an increasingly connected world, demanding openness and transparency,

Bearing in mind also that the essential trust of the public in the integrity of official statistical systems and confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and respect the rights of its members, and in this context that professional independence and accountability of statistical agencies are crucial,

Stressing that, in order to be effective, the fundamental values and principles that govern statistical work have to be guaranteed by legal and institutional frameworks and be respected at all political levels and by all stakeholders in national statistical systems,

Endorses the Fundamental Principles of Official Statistics set out below, as adopted by the Statistical Commission in 1994[2] and reaffirmed in 2013, and endorsed by the Economic and Social Council in its resolution 2013/21 of 24 July 2013:

* General Assembly Resolution 68/261 adopted on 29 January 2014. The "titles" of the Principles are not part of the original text.

[1] These include General Assembly resolution 64/267 on World Statistics Day and Economic and Social Council resolutions 2005/13 on the 2010 World Population and Housing Census Programme, 2006/6 on strengthening statistical capacity and 2013/21 on the Fundamental Principles of Official Statistics.

[2] For the original preamble used on the occasion of the initial adoption of the Fundamental Principles in 1994, see chapter V of the report of the Statistical Commission on its special session (*Official Records of the Economic and Social Council, 1994, Supplement No. 9* (E/1994/29)). More information on the Fundamental Principles and their history is available from the website of the Statistics Division.

Principle 1: ***Relevance, Impartiality, and Equal Access***
Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Principle 2: ***Professional Standards, Scientific Principles, and Professional Ethics***
To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

Principle 3: ***Accountability and Transparency***
To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Principle 4: ***Prevention of Misuse***
The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Principle 5: ***Sources of Official Statistics***
Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

Principle 6: ***Confidentiality***
Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Principle 7: ***Legislation***
The laws, regulations and measures under which the statistical systems operate are to be made public.

Principle 8: ***National Coordination***
Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

Principle 9: ***Use of International Standards***
The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

Principle 10: ***International Cooperation***
Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

Fundamental Principles of Official Statistics

- Clear mention of the **processes** used to produce statistics
- *To retain trust in official statistics, the **statistical agencies need to decide** according to strictly professional considerations, including scientific principles and professional ethics, **on the methods and procedures** for the collection, processing, storage and presentation of statistical data.*

Fundamental Principles of Official Statistics*

United Nations — For more information: unstats.un.org

The General Assembly,

Recalling recent resolutions[1] of the General Assembly and the Economic and Social Council highlighting the fundamental importance of official statistics for the national and global development agenda,

Bearing in mind the critical role of high-quality official statistical information in analysis and informed policy decision-making in support of sustainable development, peace and security, as well as for mutual knowledge and trade among the States and peoples of an increasingly connected world, demanding openness and transparency,

Bearing in mind also that the essential trust of the public in the integrity of official statistical systems and confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and respect the rights of its members, and in this context that professional independence and accountability of statistical agencies are crucial,

Stressing that, in order to be effective, the fundamental values and principles that govern statistical work have to be guaranteed by legal and institutional frameworks and be respected at all political levels and by all stakeholders in national statistical systems,

Endorses the Fundamental Principles of Official Statistics set out below, as adopted by the Statistical Commission in 1994[2] and reaffirmed in 2013, and endorsed by the Economic and Social Council in its resolution 2013/21 of 24 July 2013:

* General Assembly Resolution 68/261 adopted on 29 January 2014. The "titles" of the Principles are not part of the original text.

[1] These include General Assembly resolution 64/267 on World Statistics Day and Economic and Social Council resolutions 2005/13 on the 2010 World Population and Housing Census Programme, 2006/6 on strengthening statistical capacity and 2013/21 on the Fundamental Principles of Official Statistics.

[2] For the original preamble used on the occasion of the initial adoption of the Fundamental Principles in 1994, see chapter V of the report of the Statistical Commission on its special session (*Official Records of the Economic and Social Council, 1994, Supplement No. 9* (E/1994/29)). More information on the Fundamental Principles and their history is available from the website of the Statistics Division.

Principle 1: ***Relevance, Impartiality, and Equal Access***
Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Principle 2: ***Professional Standards, Scientific Principles, and Professional Ethics***
To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

Principle 3: ***Accountability and Transparency***
To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Principle 4: ***Prevention of Misuse***
The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Principle 5: ***Sources of Official Statistics***
Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

Principle 6: ***Confidentiality***
Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Principle 7: ***Legislation***
The laws, regulations and measures under which the statistical systems operate are to be made public.

Principle 8: ***National Coordination***
Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

Principle 9: ***Use of International Standards***
The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

Principle 10: ***International Cooperation***
Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

Fundamental Principles of Official Statistics

- Clear mention of the **processes** used to produce statistics
- *To retain trust in official statistics, the **statistical agencies need to decide** according to strictly professional considerations, including scientific principles and professional ethics, **on the methods and procedures** for the collection, processing, storage and presentation of statistical data.*
- In short, **processes** are important!

Fundamental Principles of Official Statistics*

United Nations — For more information: unstats.un.org

The General Assembly,

Recalling recent resolutions[1] of the General Assembly and the Economic and Social Council highlighting the fundamental importance of official statistics for the national and global development agenda,

Bearing in mind the critical role of high-quality official statistical information in analysis and informed policy decision-making in support of sustainable development, peace and security, as well as for mutual knowledge and trade among the States and peoples of an increasingly connected world, demanding openness and transparency,

Bearing in mind also that the essential trust of the public in the integrity of official statistical systems and confidence in statistics depend to a large extent on respect for the fundamental values and principles that are the basis of any society seeking to understand itself and respect the rights of its members, and in this context that professional independence and accountability of statistical agencies are crucial,

Stressing that, in order to be effective, the fundamental values and principles that govern statistical work have to be guaranteed by legal and institutional frameworks and be respected at all political levels and by all stakeholders in national statistical systems,

Endorses the Fundamental Principles of Official Statistics set out below, as adopted by the Statistical Commission in 1994[2] and reaffirmed in 2013, and endorsed by the Economic and Social Council in its resolution 2013/21 of 24 July 2013:

* General Assembly Resolution 68/261 adopted on 29 January 2014. The "titles" of the Principles are not part of the original text.

[1] These include General Assembly resolution 64/267 on World Statistics Day and Economic and Social Council resolutions 2005/13 on the 2010 World Population and Housing Census Programme, 2006/6 on strengthening statistical capacity and 2013/21 on the Fundamental Principles of Official Statistics.

[2] For the original preamble used on the occasion of the initial adoption of the Fundamental Principles in 1994, see chapter V of the report of the Statistical Commission on its special session (*Official Records of the Economic and Social Council, 1994, Supplement No. 9* (E/1994/29)). More information on the Fundamental Principles and their history is available from the website of the Statistics Division.

Principle 1: ***Relevance, Impartiality, and Equal Access***
Official statistics provide an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation. To this end, official statistics that meet the test of practical utility are to be compiled and made available on an impartial basis by official statistical agencies to honour citizens' entitlement to public information.

Principle 2: ***Professional Standards, Scientific Principles, and Professional Ethics***
To retain trust in official statistics, the statistical agencies need to decide according to strictly professional considerations, including scientific principles and professional ethics, on the methods and procedures for the collection, processing, storage and presentation of statistical data.

Principle 3: ***Accountability and Transparency***
To facilitate a correct interpretation of the data, the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

Principle 4: ***Prevention of Misuse***
The statistical agencies are entitled to comment on erroneous interpretation and misuse of statistics.

Principle 5: ***Sources of Official Statistics***
Data for statistical purposes may be drawn from all types of sources, be they statistical surveys or administrative records. Statistical agencies are to choose the source with regard to quality, timeliness, costs and the burden on respondents.

Principle 6: ***Confidentiality***
Individual data collected by statistical agencies for statistical compilation, whether they refer to natural or legal persons, are to be strictly confidential and used exclusively for statistical purposes.

Principle 7: ***Legislation***
The laws, regulations and measures under which the statistical systems operate are to be made public.

Principle 8: ***National Coordination***
Coordination among statistical agencies within countries is essential to achieve consistency and efficiency in the statistical system.

Principle 9: ***Use of International Standards***
The use by statistical agencies in each country of international concepts, classifications and methods promotes the consistency and efficiency of statistical systems at all official levels.

Principle 10: ***International Cooperation***
Bilateral and multilateral cooperation in statistics contributes to the improvement of systems of official statistics in all countries.

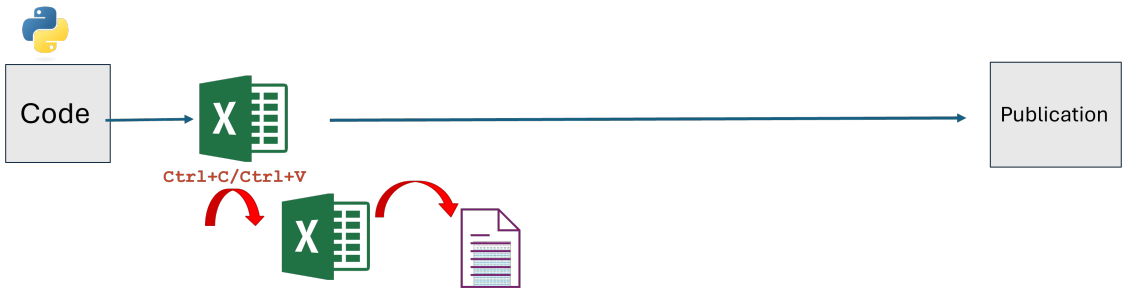
USUAL PRACTICE: THEORY VS REALITY

Comment 1

USUAL PRACTICE: THEORY VS REALITY

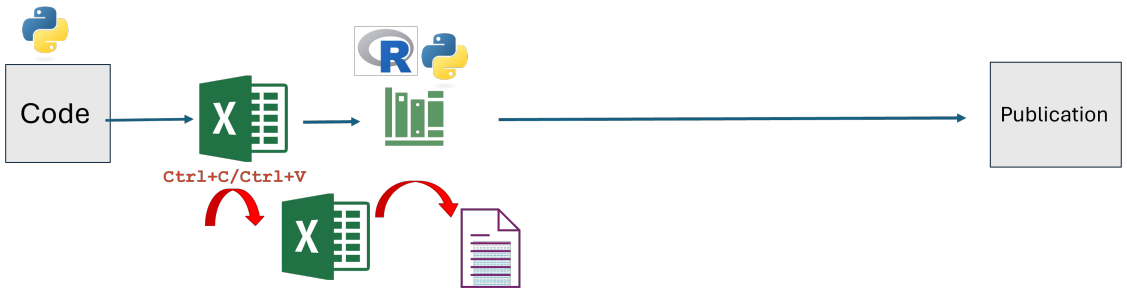
Comment 2

USUAL PRACTICE: THEORY VS REALITY



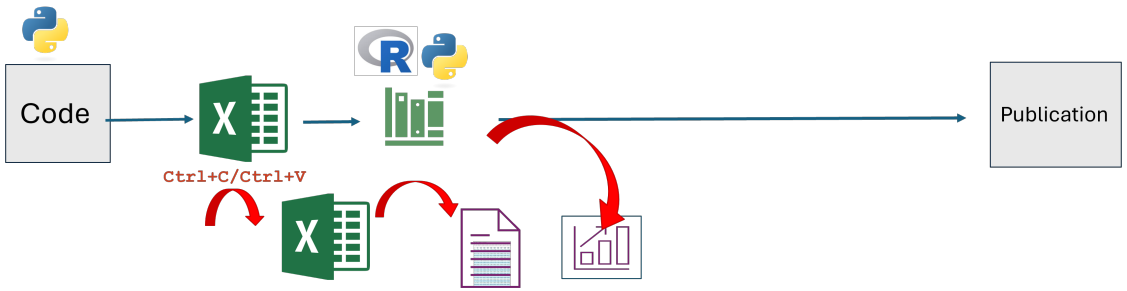
Comment 3

USUAL PRACTICE: THEORY VS REALITY



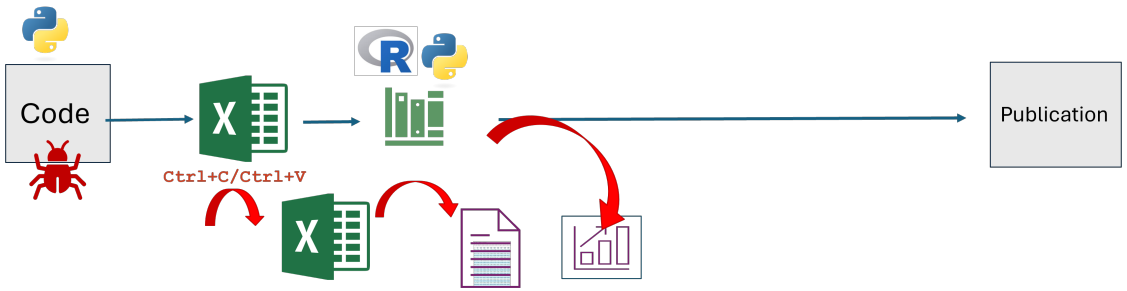
Comment 4

USUAL PRACTICE: THEORY VS REALITY



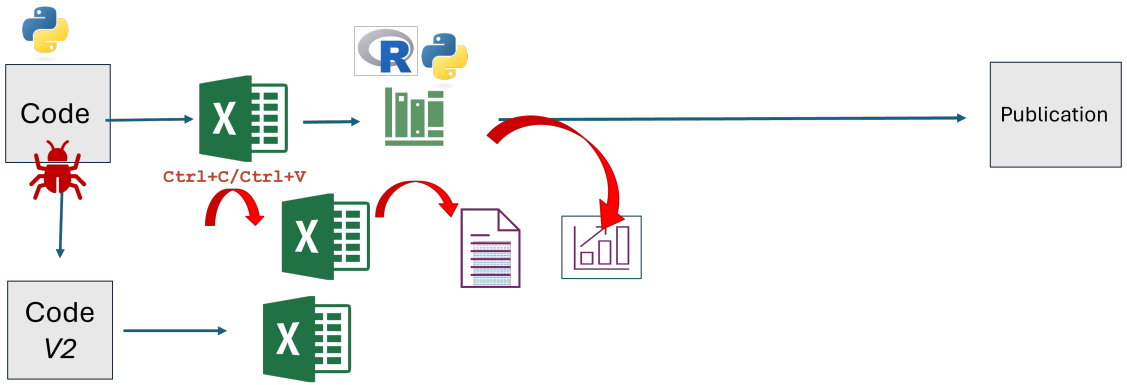
Comment 5

USUAL PRACTICE: THEORY VS REALITY



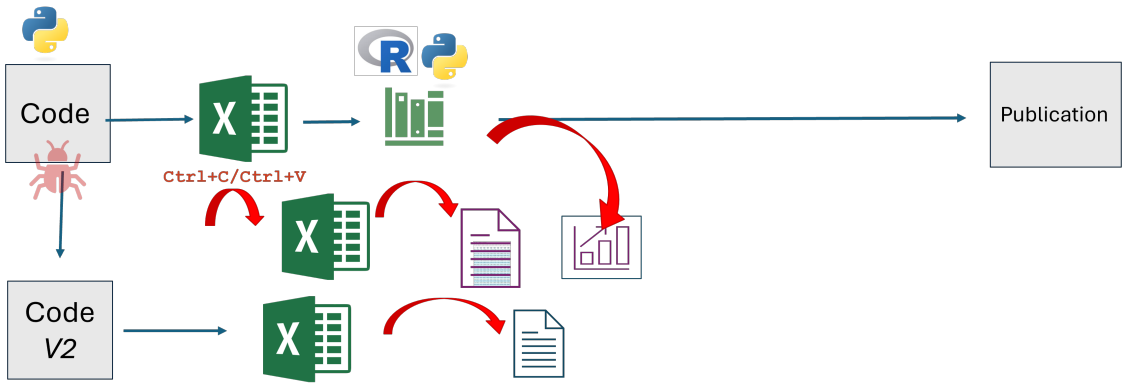
Comment 6

USUAL PRACTICE: THEORY VS REALITY



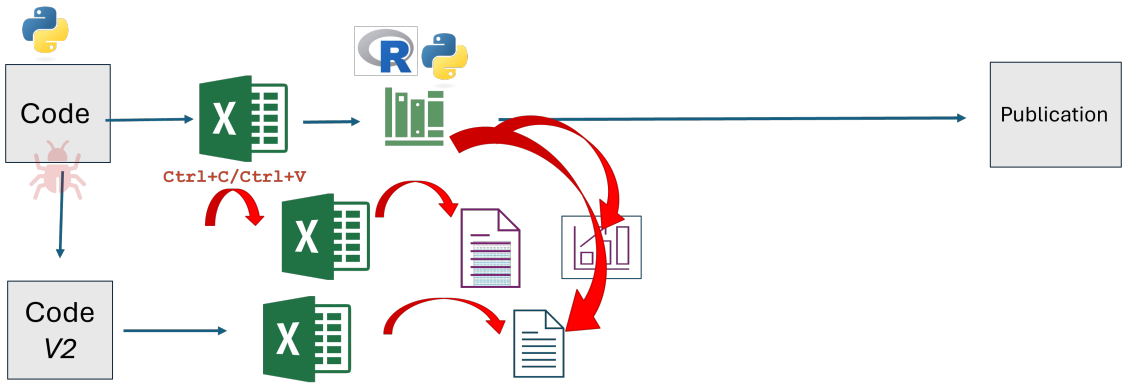
Comment 7

USUAL PRACTICE: THEORY VS REALITY



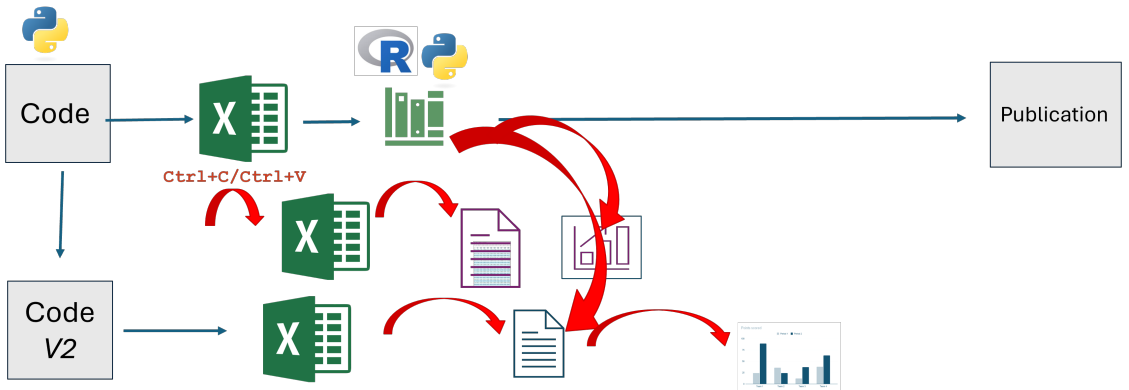
Comment 8

USUAL PRACTICE: THEORY VS REALITY



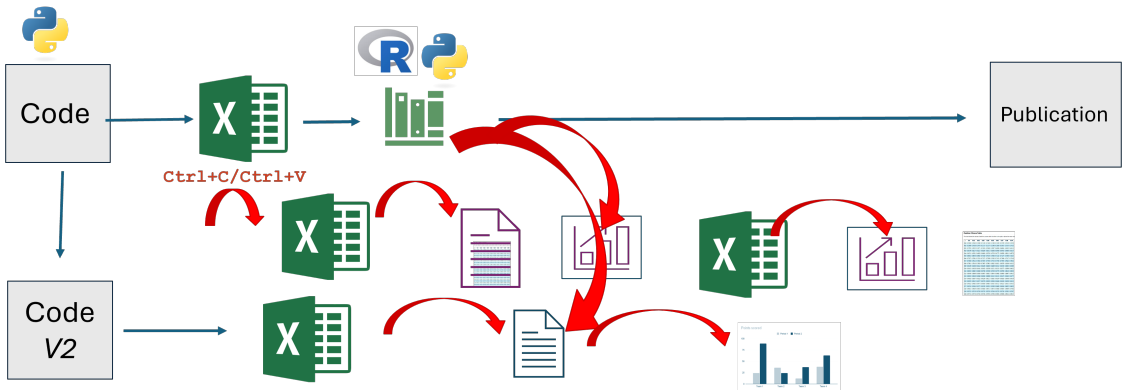
Comment 9

USUAL PRACTICE: THEORY VS REALITY



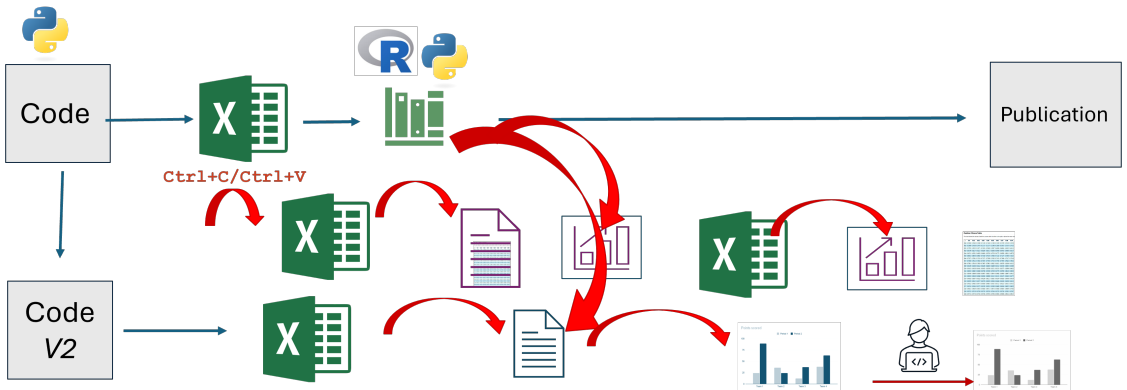
Comment 10

USUAL PRACTICE: THEORY VS REALITY



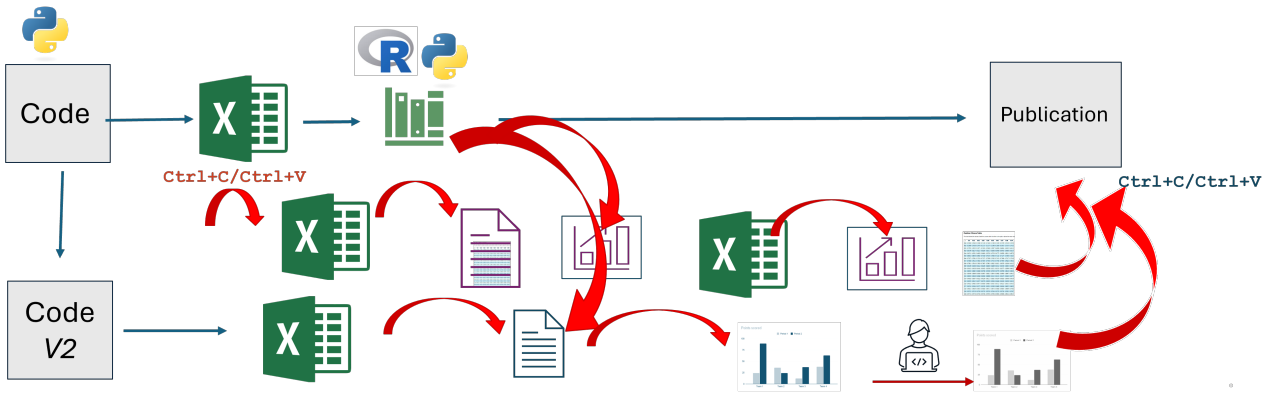
Comment 9

USUAL PRACTICE: THEORY VS REALITY



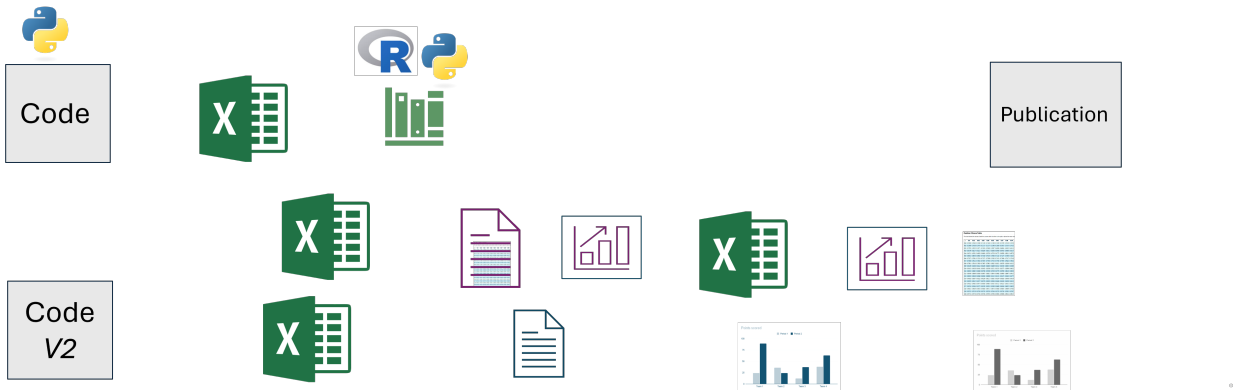
Comment 10

USUAL PRACTICE: THEORY VS REALITY



Comment 9

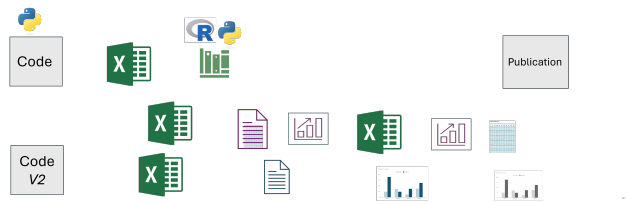
USUAL PRACTICE: THEORY VS REALITY



Comment 10

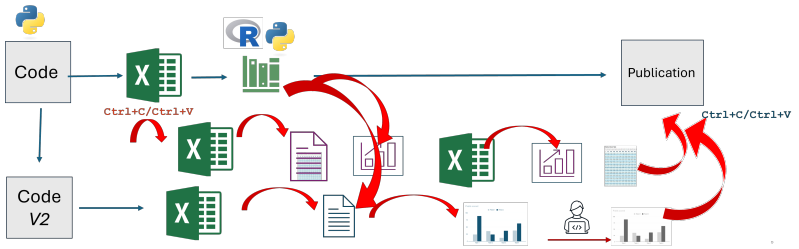
Usual practice: In the end

- Lots of files



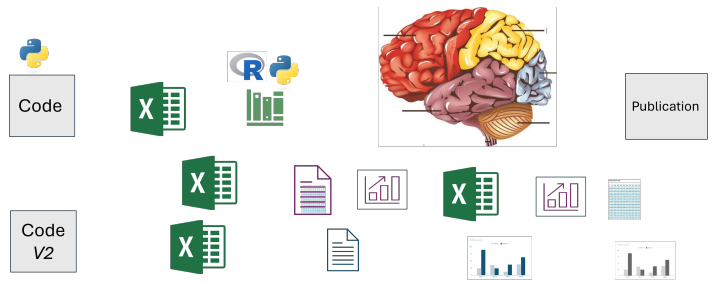
Usual practice: In the end

- Lots of files
- Cut and paste is not a reliable, reproducible approach!



Usual practice: In the end

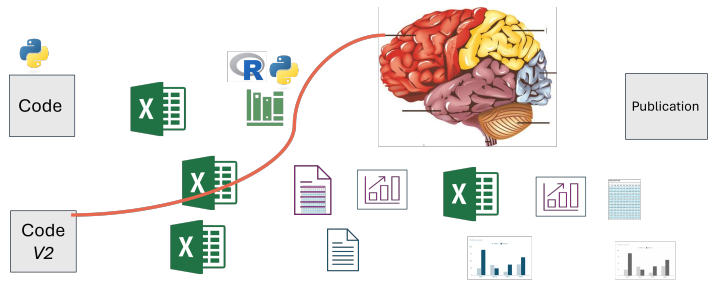
- Lots of files
- Cut and paste is not a reliable, reproducible approach!
- Your brain may remember..



Usual practice: In the end

- Lots of files
- Cut and paste is not a reliable, reproducible approach!
- Your brain may remember..

  ...all the steps...

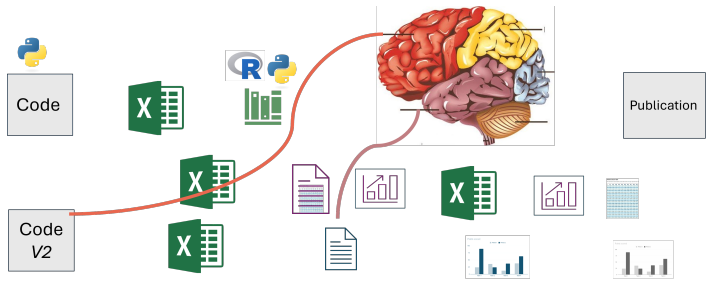


Usual practice: In the end

- Lots of files
- Cut and paste is not a reliable, reproducible approach!
- Your brain may remember..

  ...all the steps...

  .. in the right order..



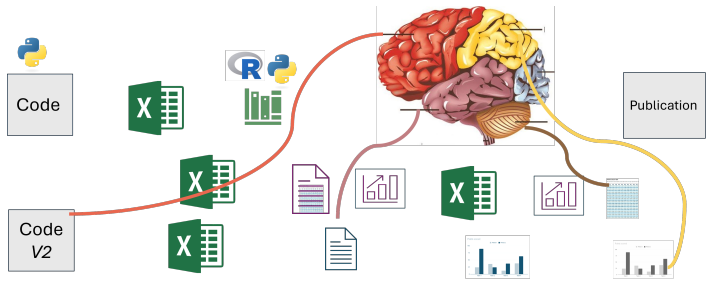
Usual practice: In the end

- Lots of files
- Cut and paste is not a reliable, reproducible approach!
- Your brain may remember..

  ...all the steps...

  .. in the right order..

  ...all of them !



Usual practice: In the end

- Lots of files
- Cut and paste is not a reliable, reproducible approach!
- Your brain may remember..

  ...all the steps...

  .. in the right order..

  ...all of them !
- Or use (bad) "*tools*"

What are the issues?

- Errors due to cut and paste

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

5 October 2020

The badly thought-out use of Microsoft's Excel software was the reason nearly 16,000 coronavirus cases went unreported in England.

And it appears that Public Health England (PHE) was to blame, rather than a third-party contractor.

What are the issues?

- Errors due to cut and paste
- Errors are difficult to track

Excel: Why using Microsoft's tool caused Covid-19 results to be lost

By Leo Kelion
Technology desk editor

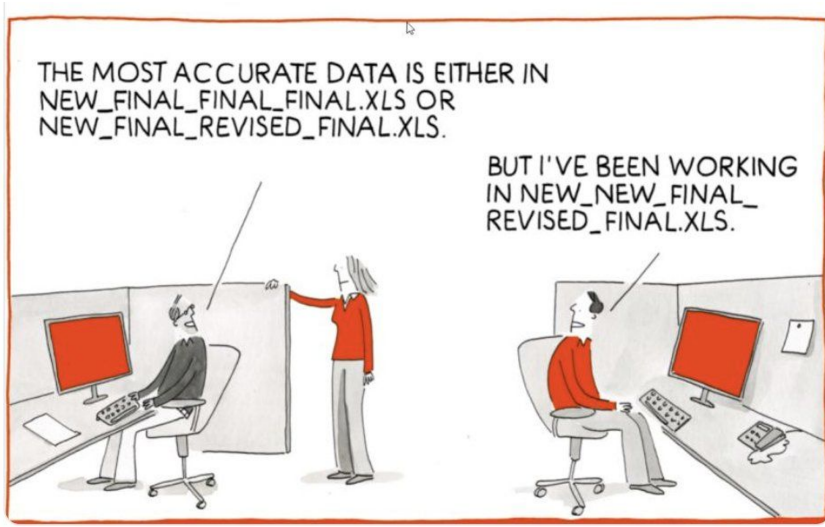
5 October 2020

The badly thought-out use of Microsoft's Excel software was the reason nearly 16,000 coronavirus cases went unreported in England.

And it appears that Public Health England (PHE) was to blame, rather than a third-party contractor.

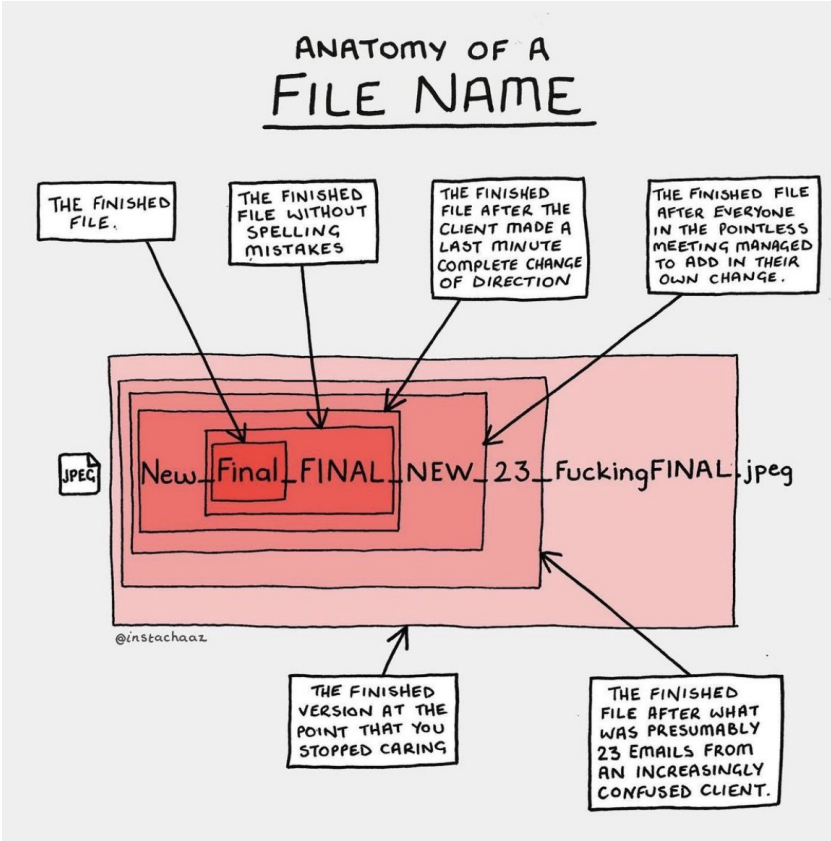
What are the issues?

- Errors due to cut and paste
- Errors are difficult to track
- Each operator has his/her own approach



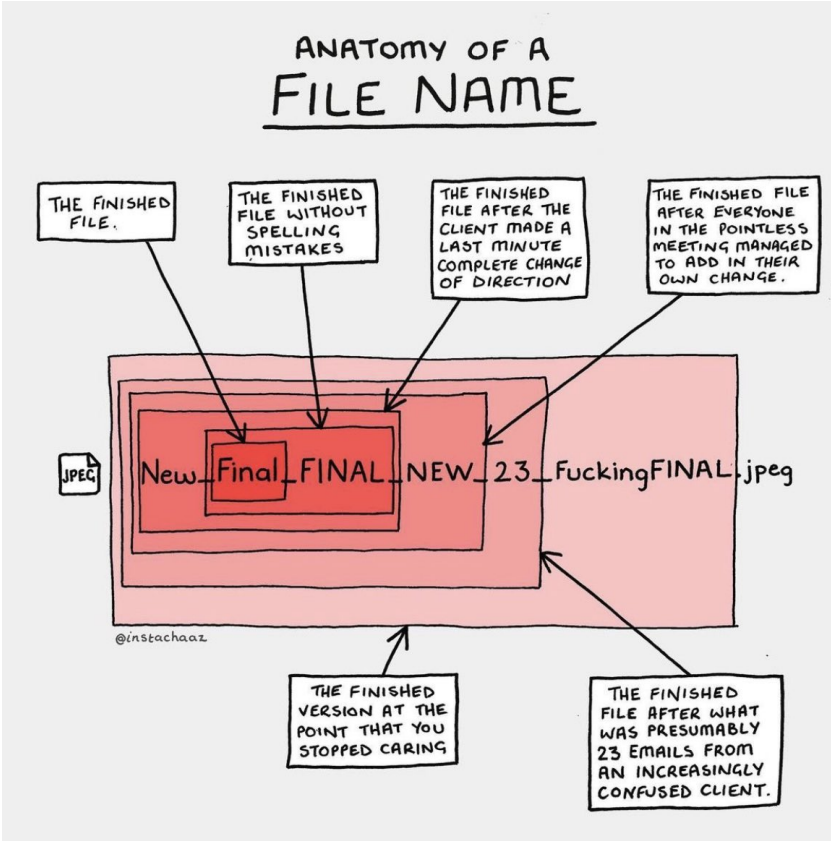
What are the issues?

- Errors due to cut and paste
- Errors are difficult to track
- Each operator has his/her own approach
- Several versions of code may coexist



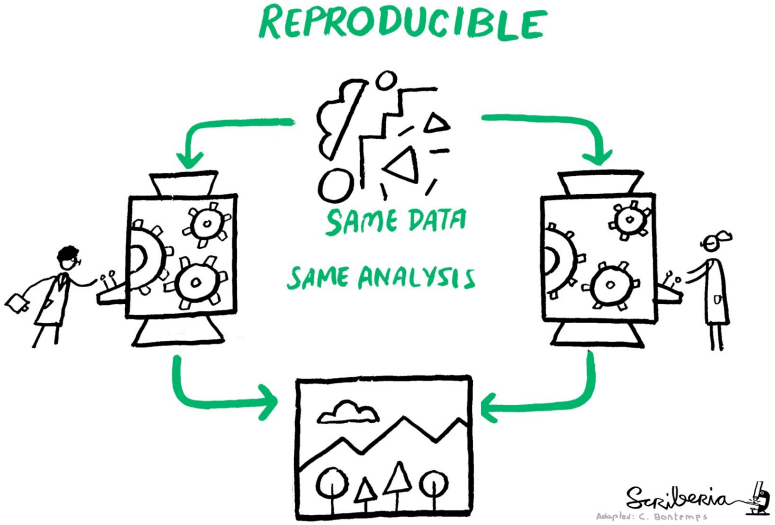
What are the issues?

- Errors due to cut and paste
- Errors are difficult to track
- Each operator has his/her own approach
- Several versions of code may coexist
- The steps aren't recorded



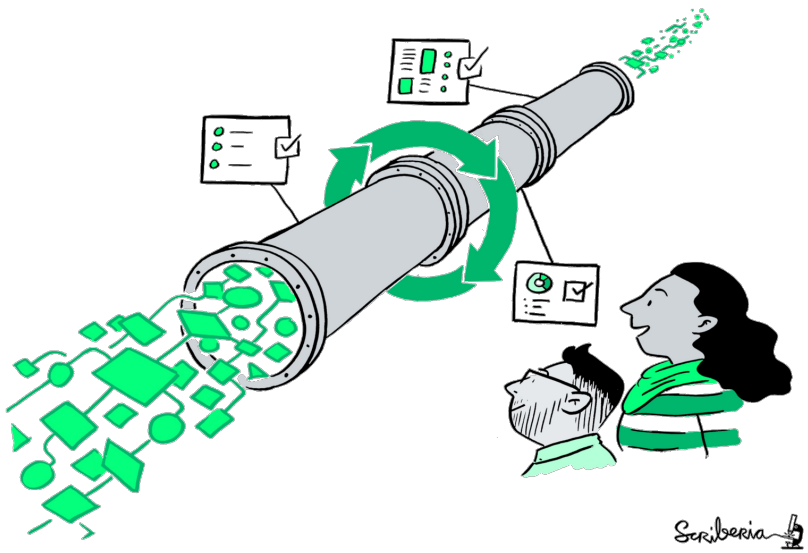
What are the issues?

- Errors due to cut and paste
- Errors are difficult to track
- Each operator has his/her own approach
- Several versions of code may coexist
- The steps aren't recorded
- Reproducibility is not granted



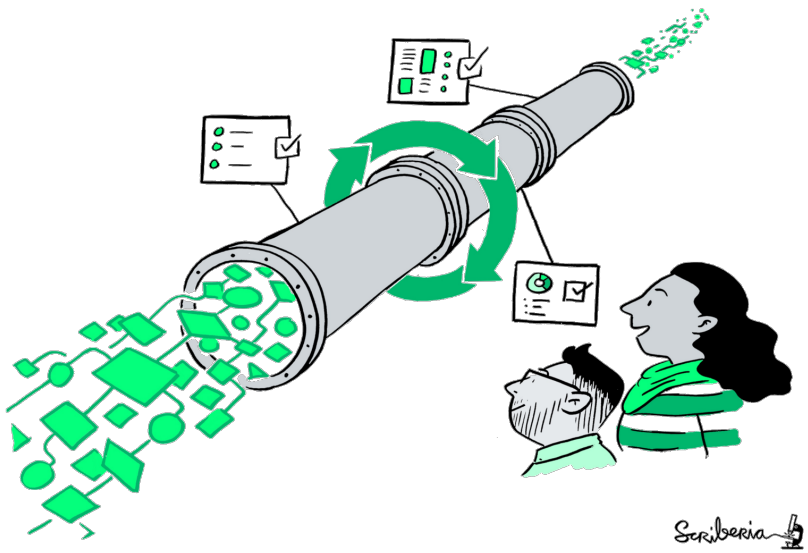
What is A Reproducible Analytical Pipeline?

- It is a process



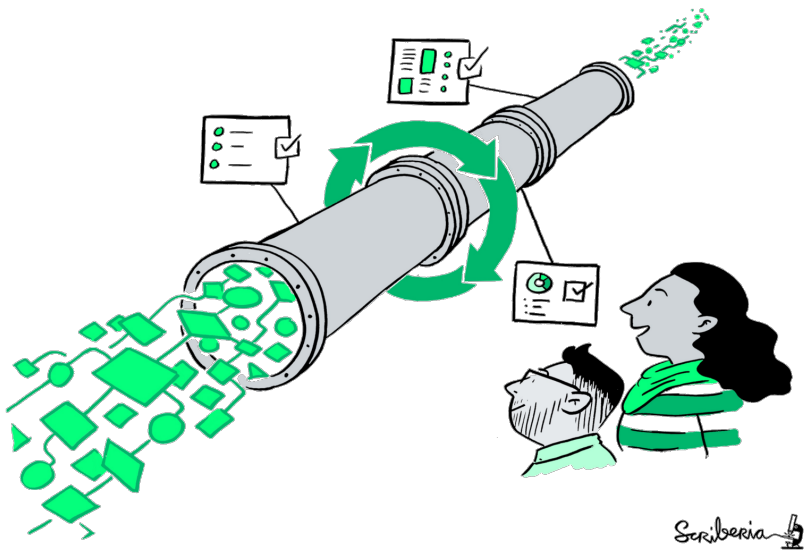
What is A Reproducible Analytical Pipeline?

- It is a process
- It is automated



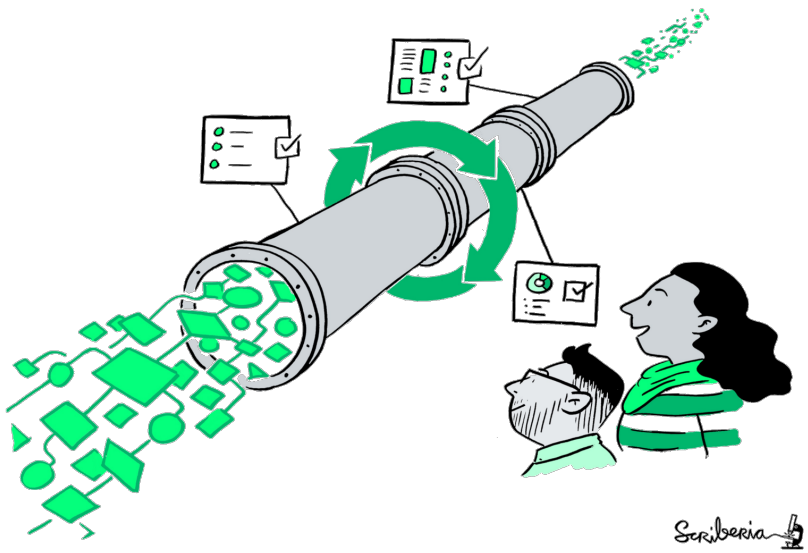
What is A Reproducible Analytical Pipeline?

- It is a process
- It is automated
- It is easily reproducible



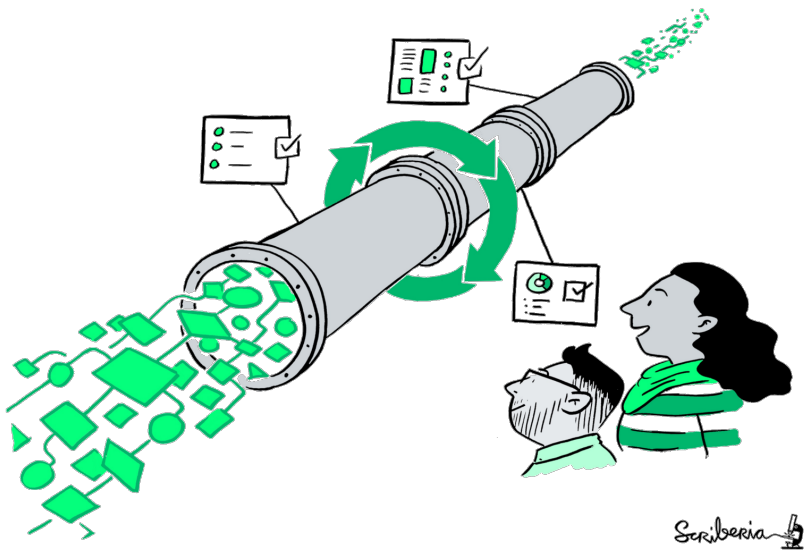
What is A Reproducible Analytical Pipeline?

- It is a process
- It is automated
- It is easily reproducible
- It minimises mistakes



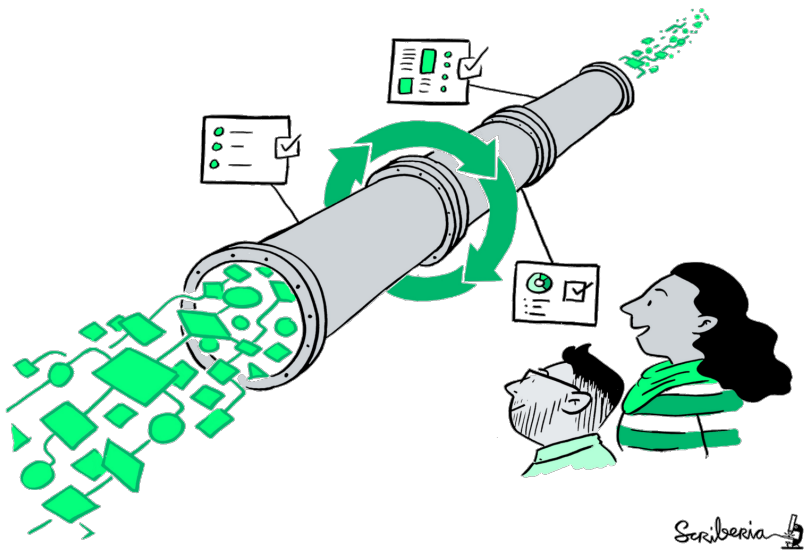
What is A Reproducible Analytical Pipeline?

- It is a process
- It is automated
- It is easily reproducible
- It minimises mistakes
- It is fast



What is A Reproducible Analytical Pipeline?

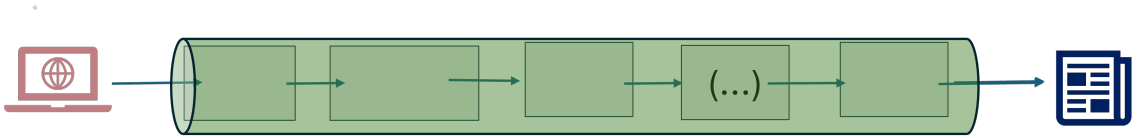
- It is a process
- It is automated
- It is easily reproducible
- It minimises mistakes
- It is fast
- It builds trust



What does a RAP look like?

Comment 1

What does a RAP look like?

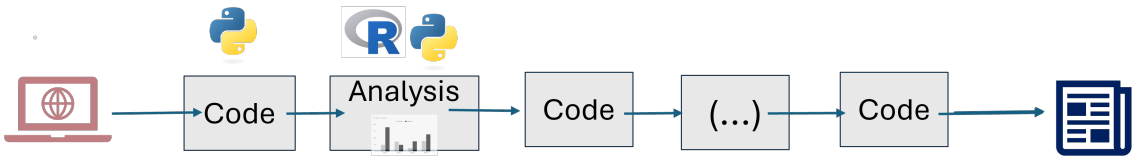


Comment 2

What does a RAP look like?

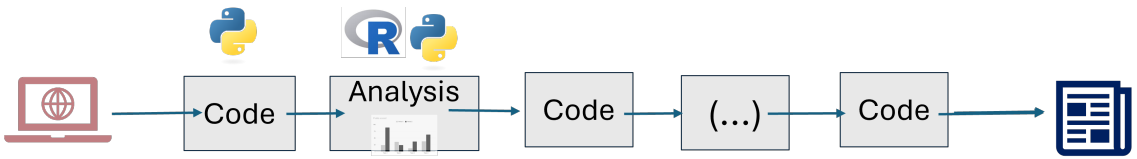
Comment 3

What does a RAP look like?



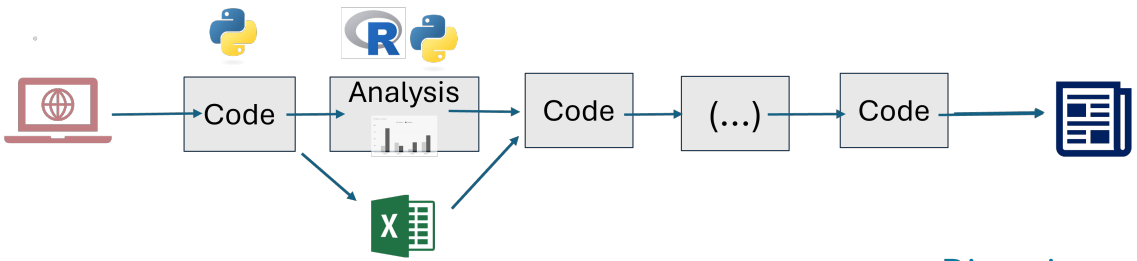
Comment 4

What does a RAP look like?



Comment 5

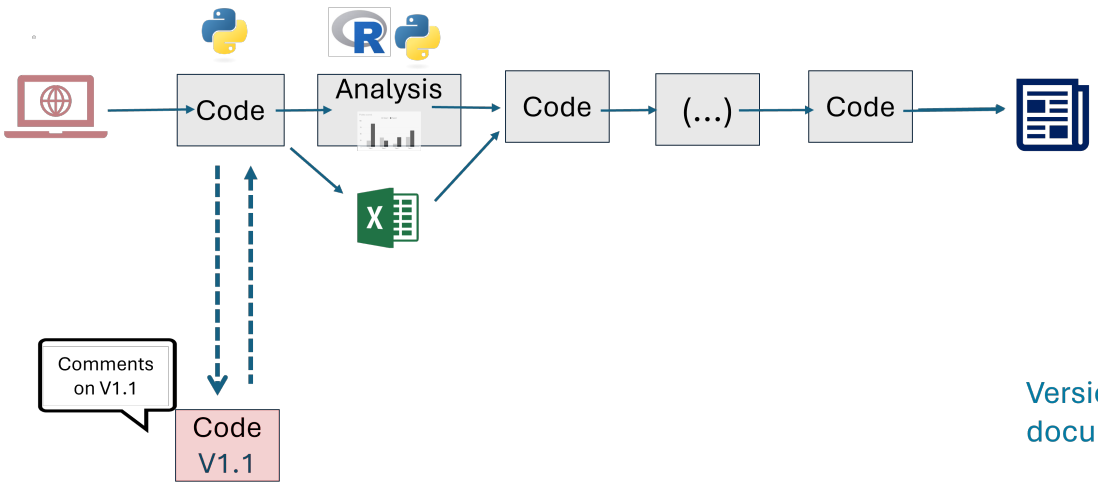
What does a RAP look like?



Direct input-output of data

Comment 6

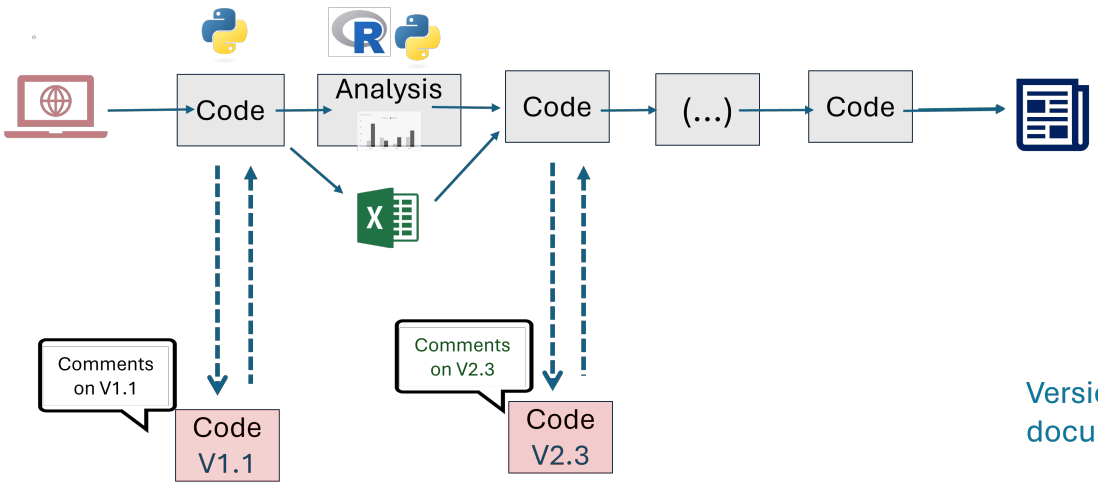
What does a RAP look like?



Versioning and documenting

Comment 7

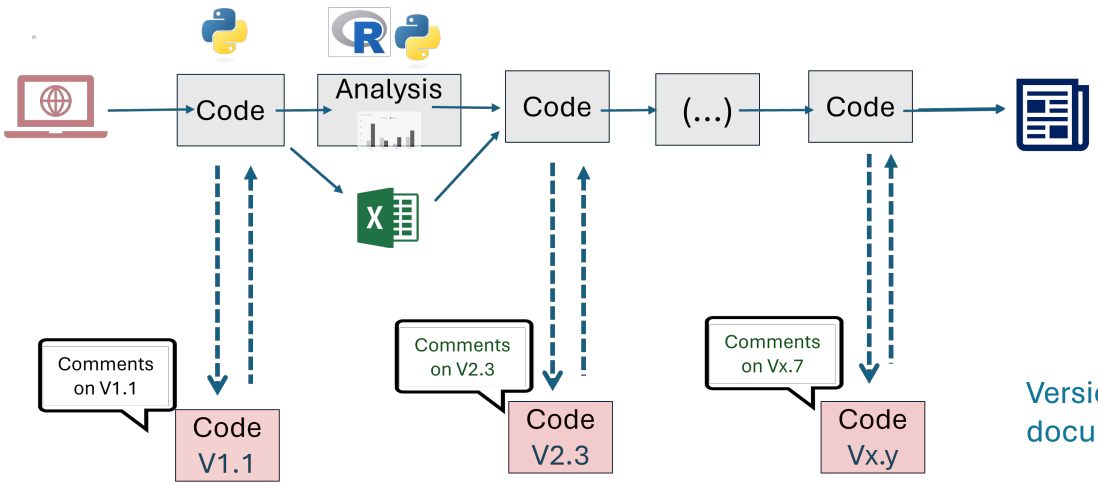
What does a RAP look like?



Versioning and documenting

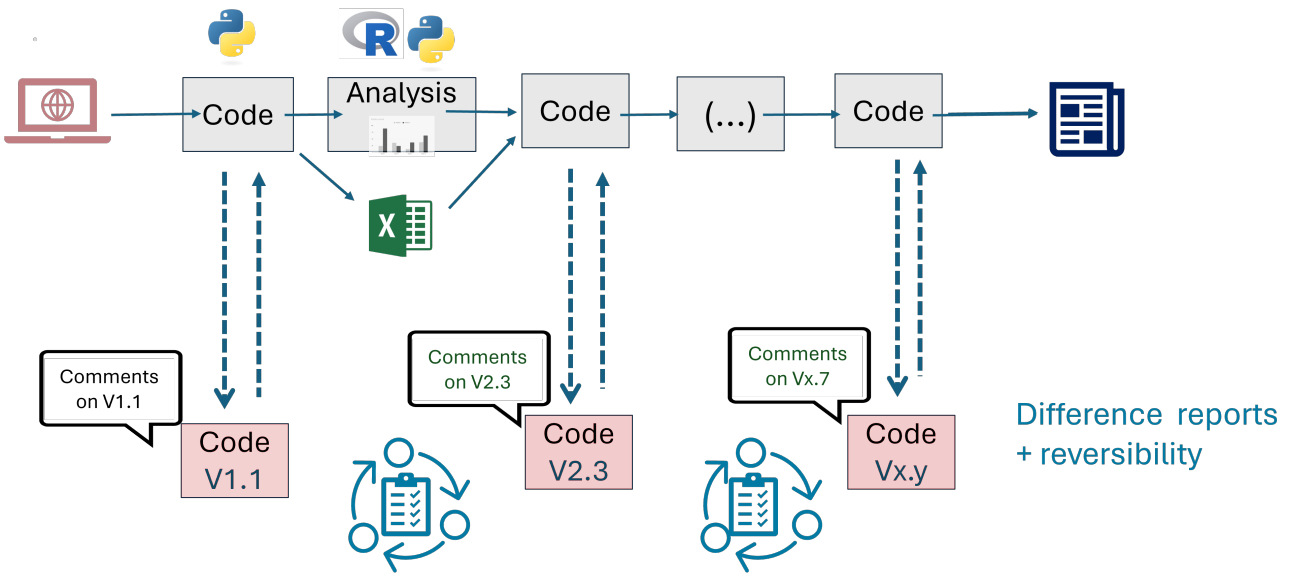
Comment 8

What does a RAP look like?



Comment 9

What does a RAP look like?



Comment 10

3 MAIN PRINCIPLES:

1. Organize your work

3 MAIN PRINCIPLES:

1. Organize your work
2. Code for others (including your future self)

3 MAIN PRINCIPLES:

1. Organize your work
2. Code for others (including your future self)
3. DRY: **D**o not **R**epeat **Y**ourself

3 MAIN PRINCIPLES:

1. Organize your work
2. Code for others (including your future self)
3. DRY: **D**o not **R**epeat **Y**ourself

3 MAIN PRINCIPLES:

1. Organize your work
2. Code for others (including your future self)
3. DRY: **D**o not **R**epeat **Y**ourself

Apply this in context (colleagues, code, software,...)

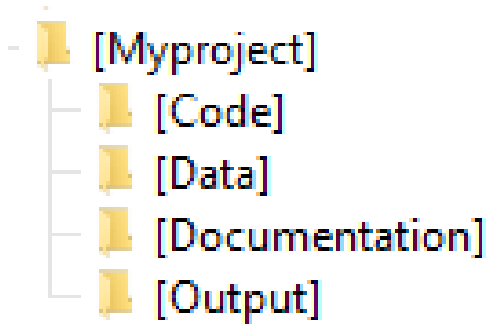
Organize your work

Have a clear directory structure

Organize your work

Have a clear directory structure

- Separate files into data, code, docs, etc.

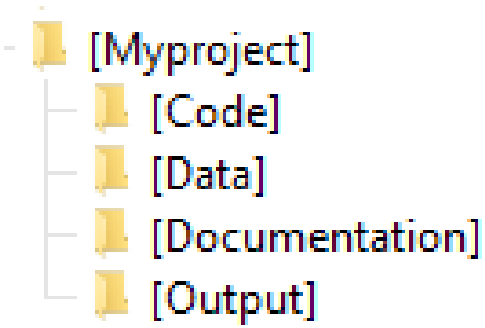


Example of a well-organized directory structure.

Organize your work

Have a clear directory structure

- Separate files into data, code, docs, etc.
- Make directories portable (relative path)



Example of a well-organized directory structure.

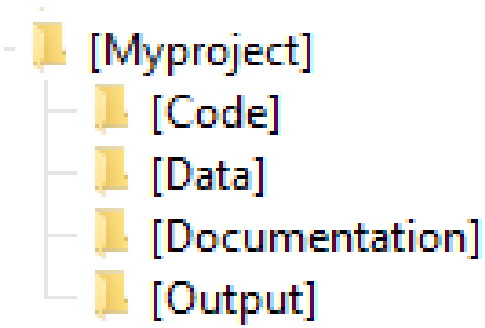
Usual

```
mydata =
pd.read_csv("c://ESCAP/Webscraping/Data/WebData.csv")
```

ORGANIZE YOUR WORK

Have a clear directory structure

- Separate files into data, code, docs, etc.
- Make directories portable (relative path)



Example of a well-organized directory structure.

Usual
mydata = pd.read_csv("c://ESCAP/Webscraping/Data/WebData.csv")

Better
Assuming your code is in c://ESCAP/Webscraping/Code/
mydata = pd.read_csv("../Data/WebData.csv")

Organize your work

Use naming conventions:
For **files/code**

- Avoid lazy names

Usual

```
prog1.ipynb
prog2.ipynb
Stat.ipynb
progC.ipynb
progP.ipynb
```

Organize your work

Use naming conventions:
For **files/code**

- Avoid lazy names
- Meaningful files names

Usual	Better
`prog1.ipynb`	`Scraping_Data.ipynb`
`prog2.ipynb`	`Cleaning_Data.ipynb`
`Stat.ipynb`	`Stats_Tables.ipynb`
`progC.ipynb`	`Classification.ipynb`
`progP.ipynb`	`Price_CPI.ipynb`

Organize your work

Use naming conventions:
For **files/code**

- Avoid lazy names
- Meaningful files names
- Order of execution

Usual	Even better
`prog1.ipynb`	`01_Scraping_data.ipynb`
`prog2.ipynb`	`02_Cleaning_data.ipynb`
`Stat.ipynb`	`03_Classification.ipynb`
`progC.ipynb`	`04_Stats_Tables.ipynb`
`progP.ipynb`	`04_Price_CPI.ipynb`

Organize your work

Use naming conventions:

For **outputs**

- Avoid numbering

Usual

```
Table1.pdf
Table2.pdf
Graph.jpg
Model.csv
```

Organize your work

Use naming conventions:
For **outputs**

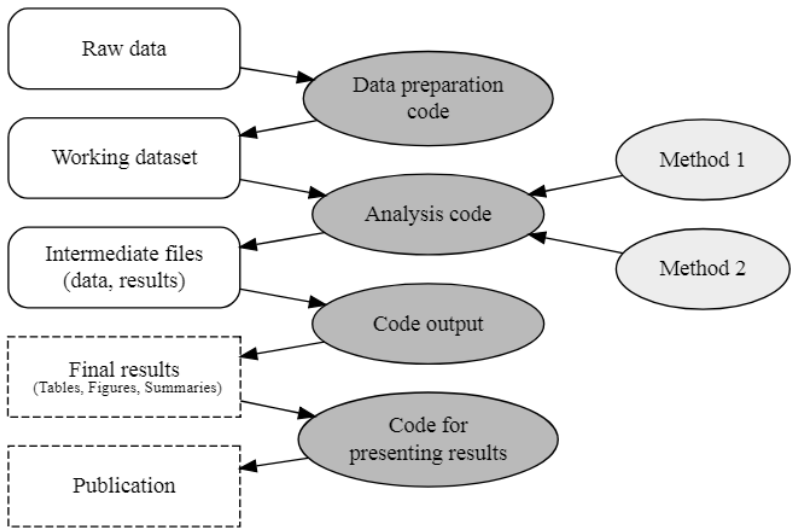
- Avoid numbering
- Explicit type of output

Usual	Better
`Table1.pdf`	`Stat_Desc_Table.pdf`
`Table2.pdf`	`Price_Stat_Table.pdf`
`Graph.jpg`	`Dress_Prices_Graphic.jpg`
`Model.csv`	`All_prices_Results.csv`

ORGANIZE YOUR WORK

Keep track of the workflow:

- Cut and paste should be avoided

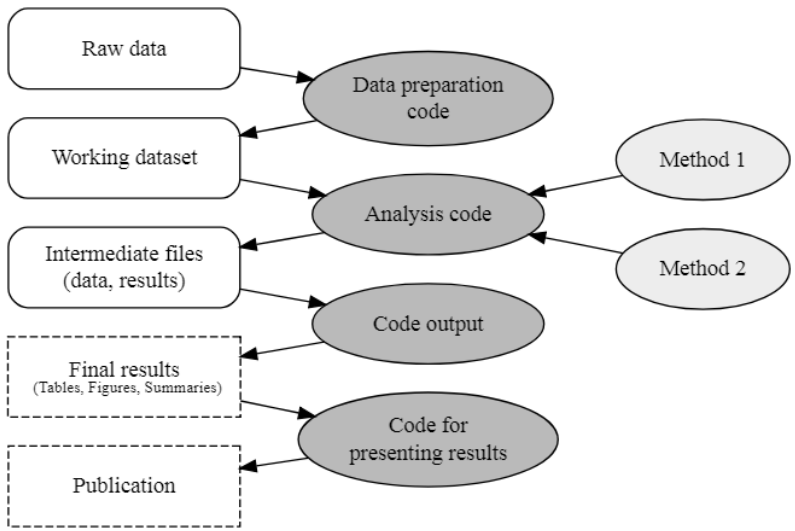


Example of a simple workflow.

ORGANIZE YOUR WORK

Keep track of the workflow:

- Cut and paste should be avoided
- Every step of the process is coded

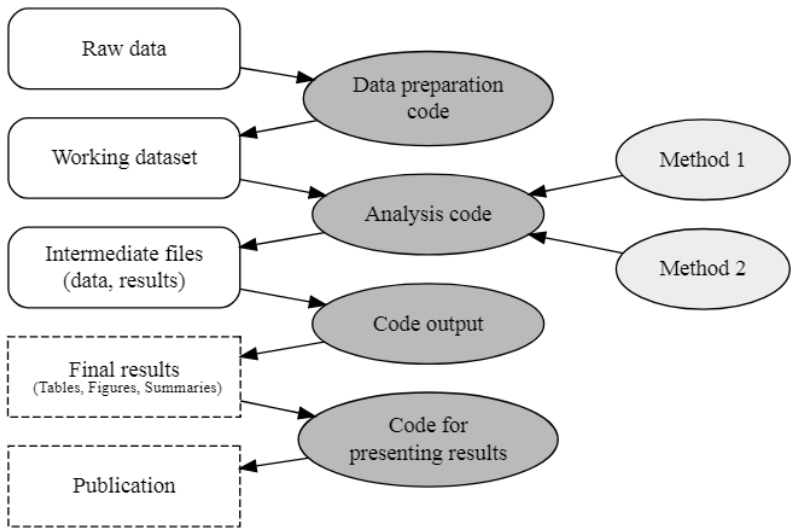


Example of a simple workflow.

Organize your work

Keep track of the workflow:

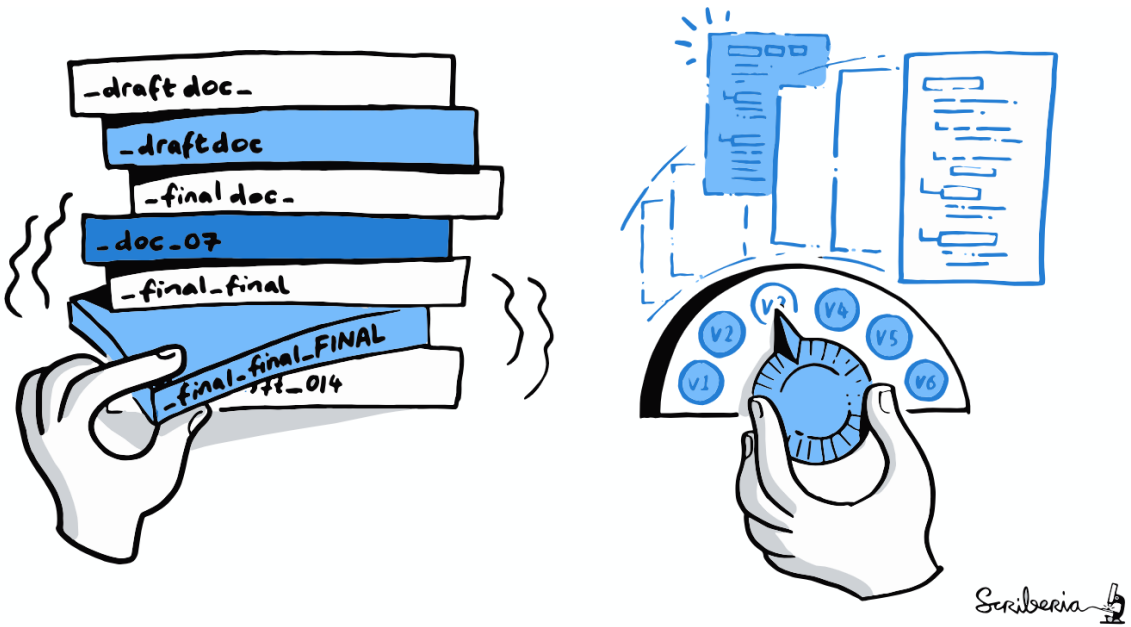
- Cut and paste should be avoided
- Every step of the process is coded
- Manage (and draw) the workflow



Example of a simple workflow.

ORGANIZE YOUR WORK

Use a version control system (Git/GitHub)



More on Version Control later

CODE FOR OTHERS (INCLUDING YOUR "*future self*")

Program with style:

Use *literate programming*

> *"Let us concentrate rather on explaining to humans what we want the computer to do"*
>
> D. Knuth (1984)

CODE FOR OTHERS (INCLUDING YOUR "*future self*")

Program with style:

"code is read much more often than it is written"

Guido van Rossum (2013 -PEP8)

PEP stands for Python Enhancement Proposals

CODE FOR OTHERS (INCLUDING YOUR "*future self*")

Program with style:

Use conventions on layout (Comments, indentation,...)

Contents

PEP 8 – Style Guide for Python Code

Author: Guido van Rossum <guido at python.org>, Barry Warsaw <barry at python.org>, Alyssa Coghlan <ncoghlan at gmail.com>

Status: Active
Type: Process
Created: 05-Jul-2001
Post-History: 05-Jul-2001, 01-Aug-2013

▶ **Table of Contents**

Introduction

This document gives coding conventions for the Python code comprising the standard library in the main Python distribution. Please see the companion informational PEP describing style guidelines for the C code in the C implementation of Python.

This document and PEP 257 (Docstring Conventions) were adapted from Guido's original Python Style Guide essay, with some additions from Barry's style guide [2].

This style guide evolves over time as additional conventions are identified and past conventions are rendered obsolete by changes in the language itself.

Many projects have their own coding style guidelines. In the event of any conflicts, such project-specific guides take precedence for that project.

A Foolish Consistency is the Hobgoblin of Little Minds

One of Guido's key insights is that code is read much more often than it is written. The guidelines provided here are intended to improve the readability of code and make it consistent across the wide spectrum of Python code. As PEP 20 says, "Readability counts".

A style guide is about consistency. Consistency with this style guide is important. Consistency within a project is more important. Consistency within one module or function is the most important.

However, know when to be inconsistent – sometimes style guide recommendations just aren't applicable. When in doubt, use your best judgment. Look at other examples and decide what looks best. And don't hesitate to ask!

CODE FOR OTHERS <– XXXX TO REVISE THIS

Program with style

- Avoid ambiguities

Usual

```
sex <- ifelse(gender == "1001", 1, 2)
```

Better

```
female <- ifelse(gender == "1001",1,0)
 male <- ifelse(gender != "1001",1,0)
```

CODE FOR OTHERS <– XXXX TO REVISE THIS

Program with style

- Avoid ambiguities
- Avoid changing units

Usual

```
gdp <- gdp/118.722
```

CODE FOR OTHERS <– XXXX TO REVISE THIS

Program with style

- Avoid ambiguities
- Avoid changing units

Usual

```
gdp <- gdp/118.722
```

Better

```
gdp_US <- gdp / 118.722
```

CODE FOR OTHERS <– XXXX TO REVISE THIS

Program with style

- Avoid ambiguities
- Avoid changing units

Usual

```
gdp <- gdp/118.722
```

Even better

```
US_Vanu_exch_rate <- 118.722
gdp_US <- gdp / US_Vanu_exch_rate
```

DO NOT REPEAT YOURSELF <–XXXX TO REVISE THIS

Create reusable objects

- Store values

Usual

```
Current_Data <- subset(Mydata, year ==2023)
```

DO NOT REPEAT YOURSELF <–XXXX TO REVISE THIS

Create reusable objects

- Store values

Usual

```
Current_Data <- subset(Mydata, year ==2023)
```

Better

```
Current_year <- 2023
Current_Data <- subset(Mydata,
                       year == Current_year)
```

DO NOT REPEAT YOURSELF <–XXXX TO REVISE THIS

Create reusable objects

- Store values
- Avoid repetitions

Usual

```
data <- Mydata[Mydata$export == "Beef", ]
plot(data$Year, data$Value,
                          main = "Export for Beef")

data <- Mydata[Mydata$export == "Kava", ]
plot(data$Year, data$Value,
                          main = "Export for Kava")
 ...
```

DO NOT REPEAT YOURSELF <–XXXX TO REVISE THIS

Create reusable objects

- Store values
- Avoid repetitions

Better

```
type <- "Beef"

Mydata %>%
filter(exports == type) %>%
ggplot() +
aes(x = Year, y = Value) +
geom_point() +
ggtitle(paste("Export for ", type))
```

DO NOT REPEAT YOURSELF <–XXXX TO REVISE THIS

Create reusable objects

- Store values
- Avoid repetitions
- Use functions

Even better

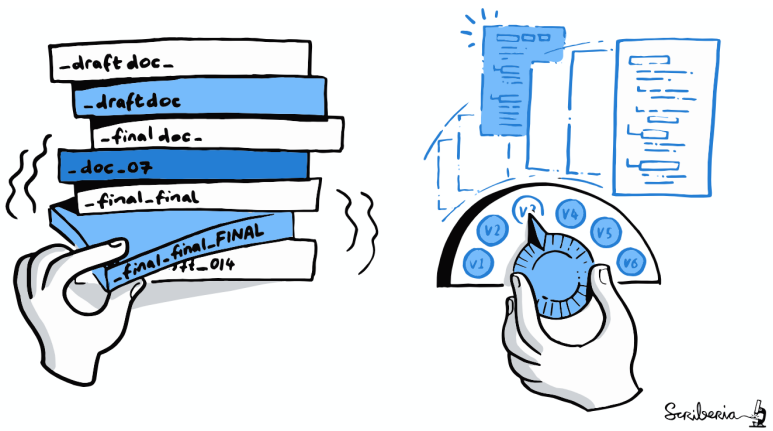
```
Exports_graphic <- function(type) {
 Mydata %>%
 filter(exports == type) %>%
 ggplot() +
 aes(x = Year, y = Value) +
 geom_point() +
 ggtitle(paste("Export for ", type))
}
```

```
Exports_graphic("Beef")
Exports_graphic("Kava")
```

VERSION CONTROL KEEPS TRACKS OF YOUR WORK

Tracking three **W** questions:

What changes?



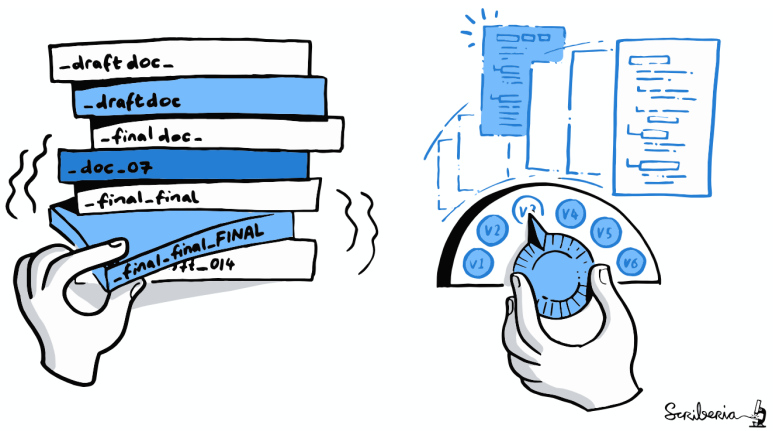
Source: The Turing Way project

VERSION CONTROL KEEPS TRACKS OF YOUR WORK

Tracking three **W** questions:

What changes?

Who made the changes?

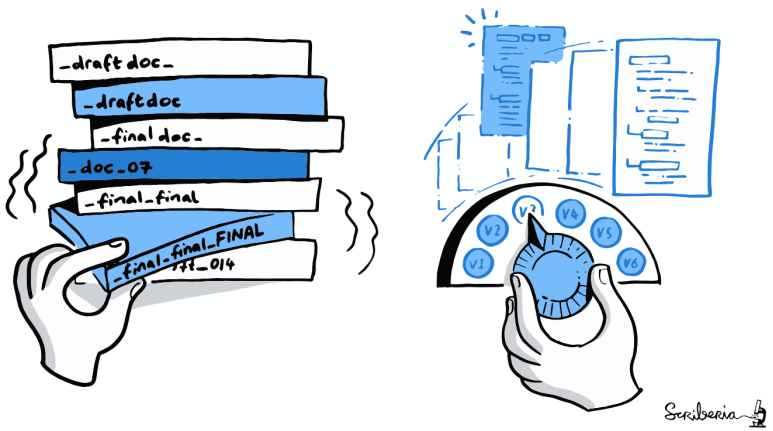


Source: The Turing Way project

VERSION CONTROL KEEPS TRACKS OF YOUR WORK

Tracking three **W** questions:

- **W**hat changes?
- **W**ho made the changes?
- **W**hen were the changes made?

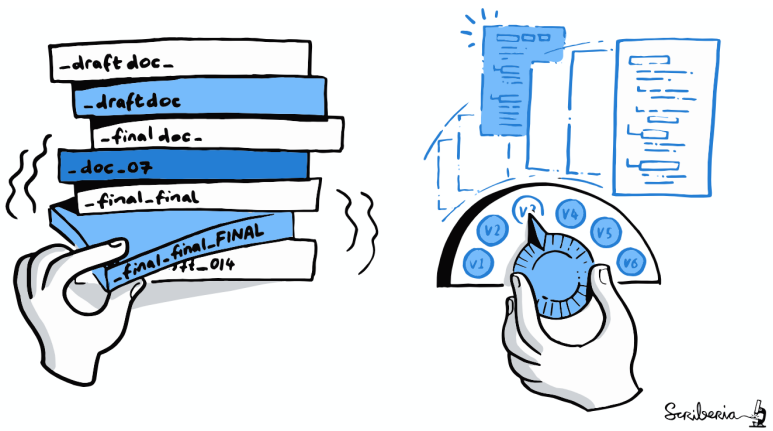


Source: The Turing Way project

VERSION CONTROL KEEPS TRACKS OF YOUR WORK

Tracking three **W** questions:

- **W**hat changes?
- **W**ho made the changes?
- **W**hen were the changes made?



Source: The Turing Way project

Transparency, Accountability & Reproducibility

- Version control provides a detailed history of changes

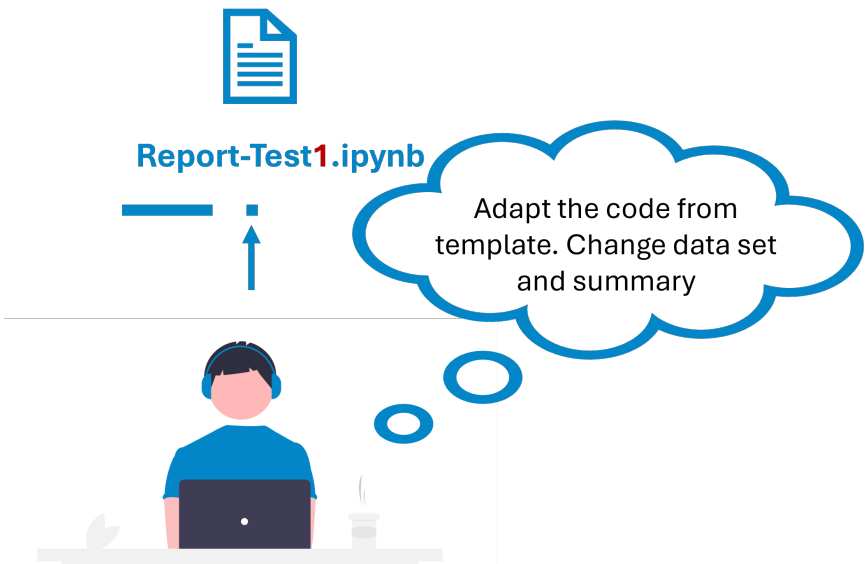
Transparency, Accountability & Reproducibility

- Version control provides a detailed history of changes
- Each modification is attributed to a specific user

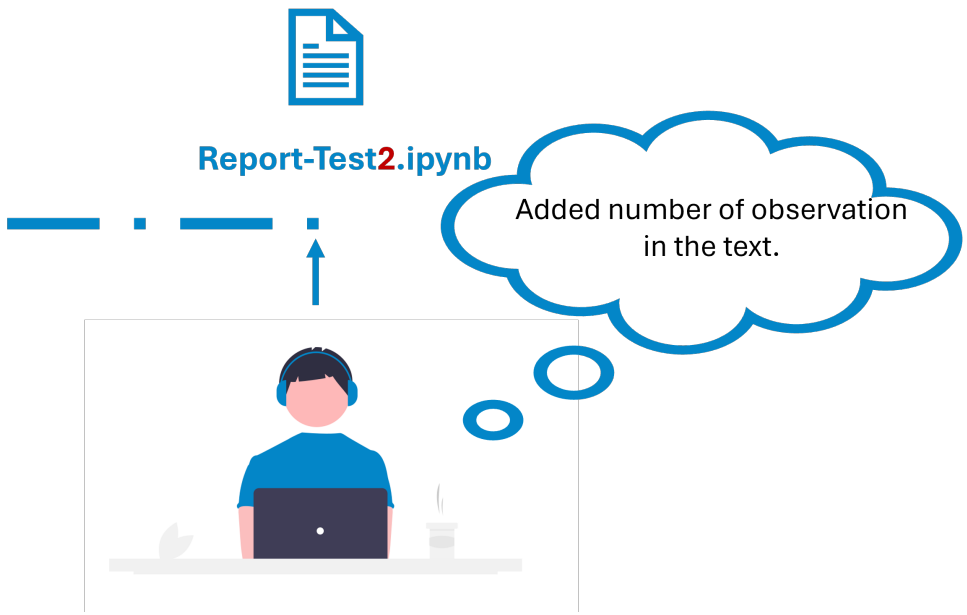
Transparency, Accountability & Reproducibility

- Version control provides a detailed history of changes
- Each modification is attributed to a specific user
- Promotes accountability, transparency & reproducibility

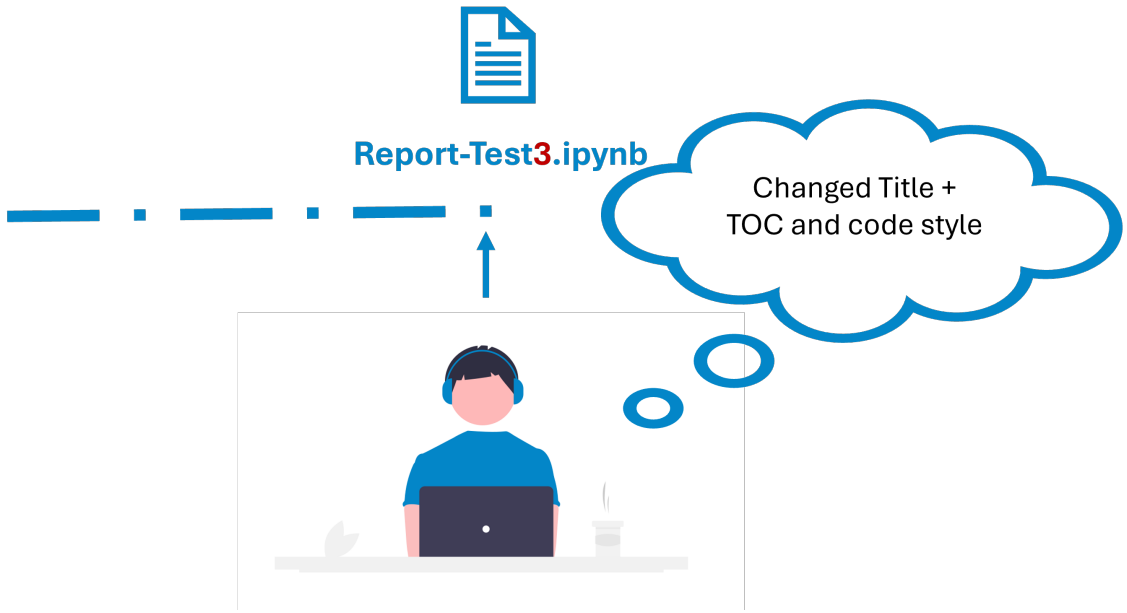
THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



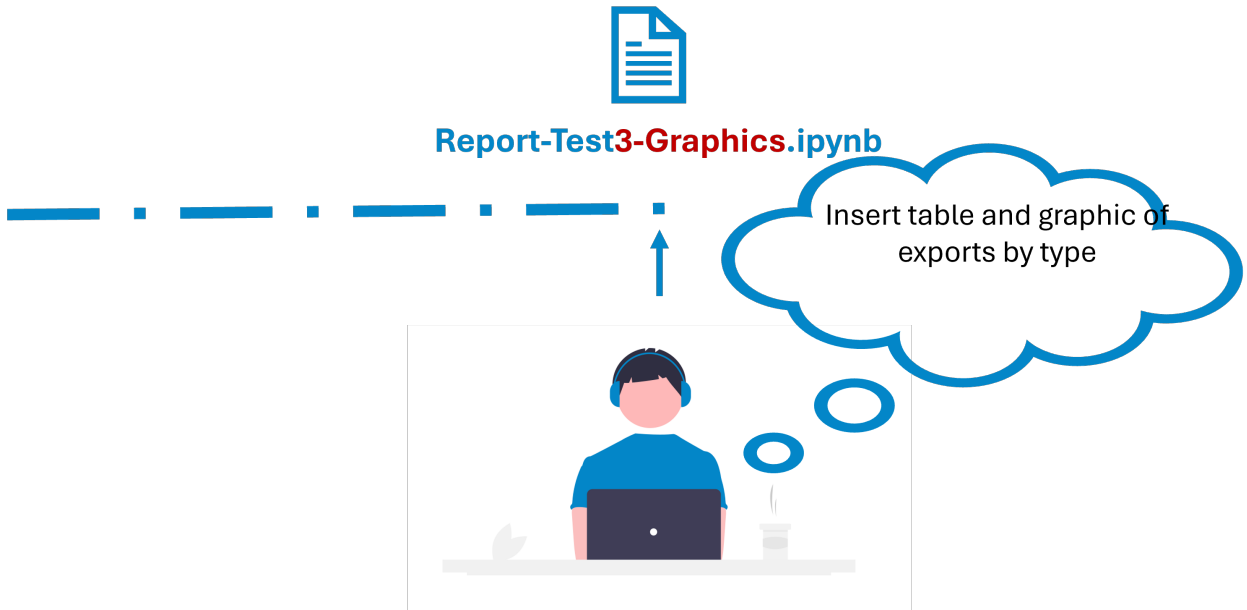
THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



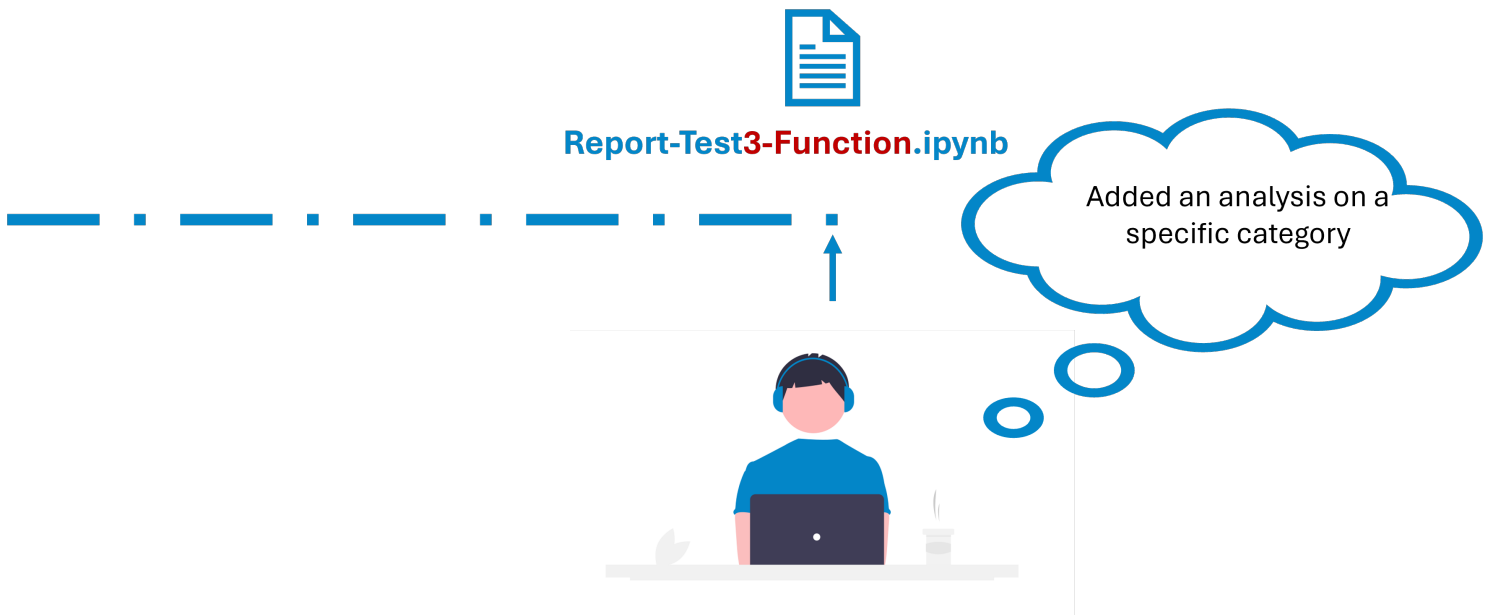
THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



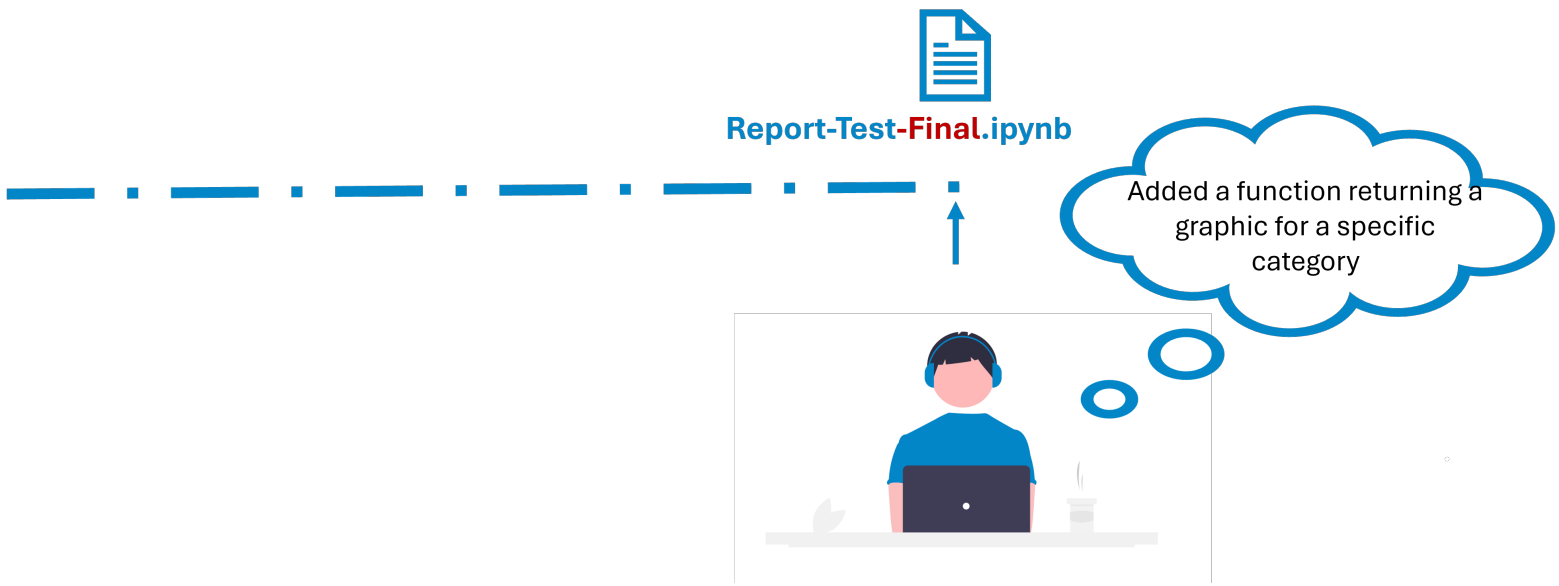
THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



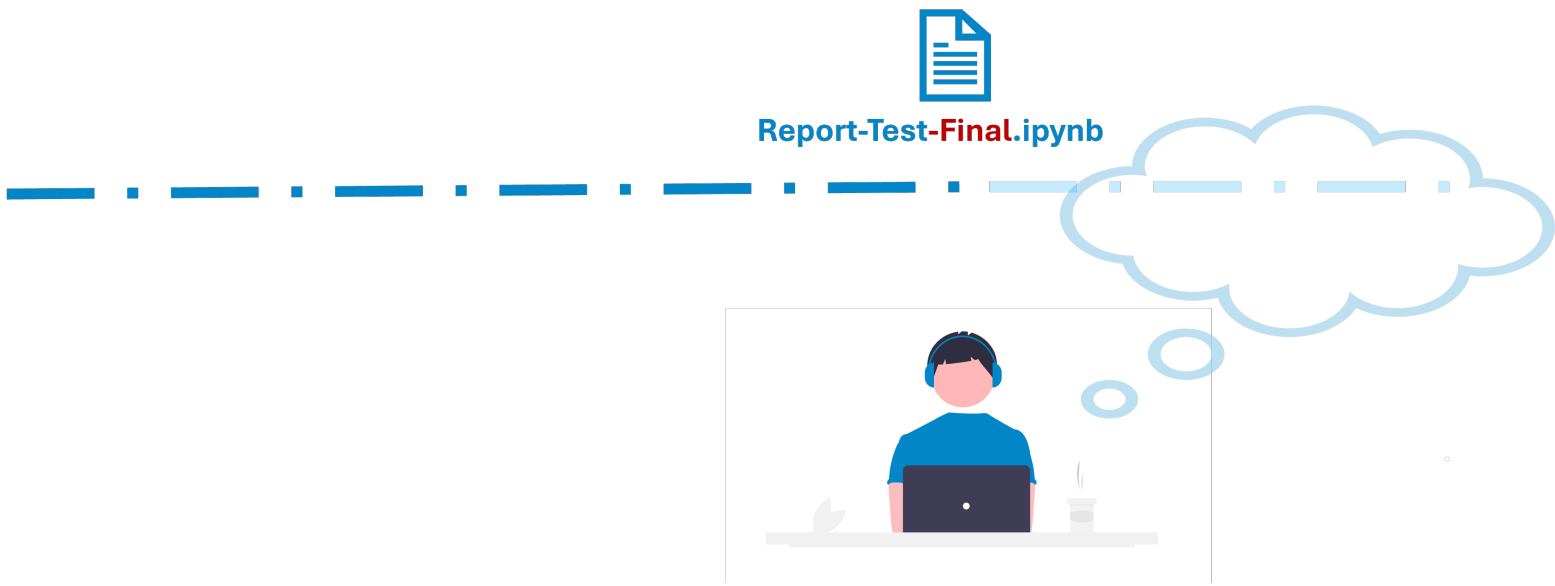
THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS



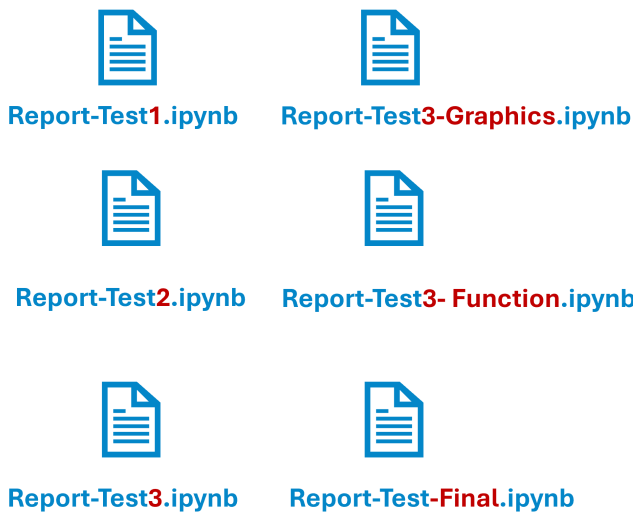
THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS

Usual ways to keep track of changes:

THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS

Usual ways to keep track of changes:

- New file after each change

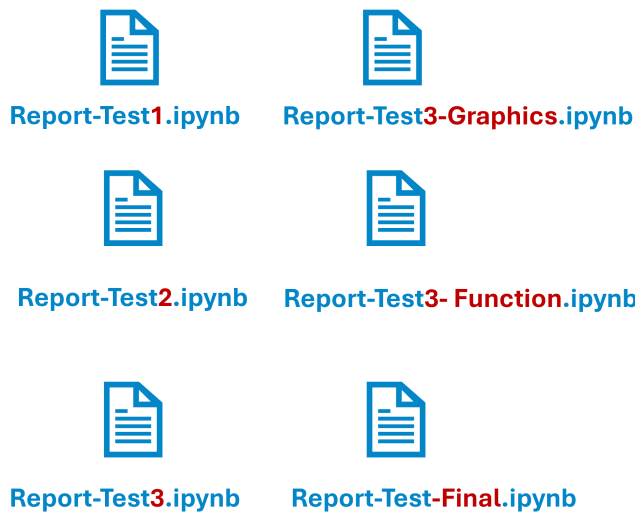


THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS

Usual ways to keep track of changes:

- New file after each change

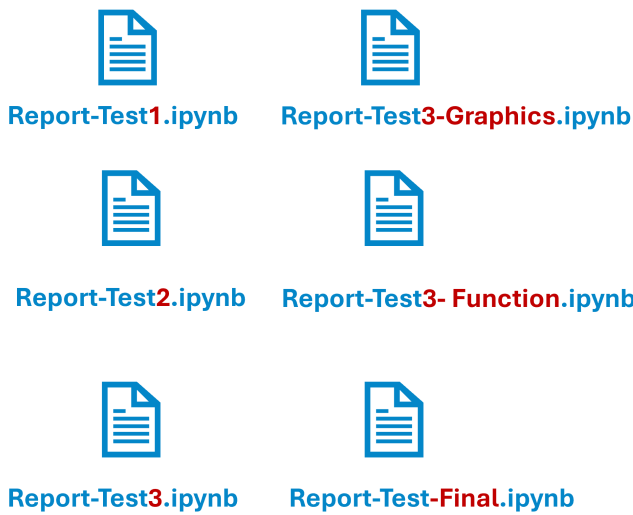
↪ Need to open each file to see the change



THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS

Usual ways to keep track of changes:

- New file after each change
- ↪ Need to open each file to see the change
- ↪ Names have to be explicit



THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS

Usual ways to keep track of changes:

- New file after each change
- ↪ Need to open each file to see the change
- ↪ Names have to be explicit
- Only the last file with lots of comments

Report-Test3-Graphics-Functions-Final-Chris.ipynb

THE EVOLUTION OF A FILE <–XXXX TO REVISE THIS

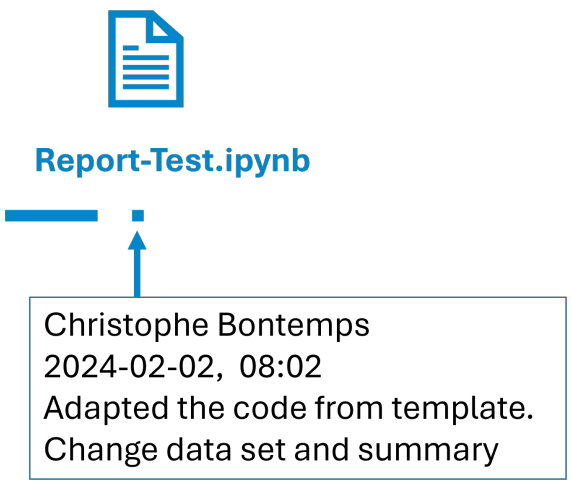
Usual ways to keep track of changes:

- New file after each change
- ↪ Need to open each file to see the change
- ↪ Names have to be explicit
- Only the last file with lots of comments
- Not fulfilling the 3 W...

Report-Test3-Graphics-Functions-Final-Chris.ipynb

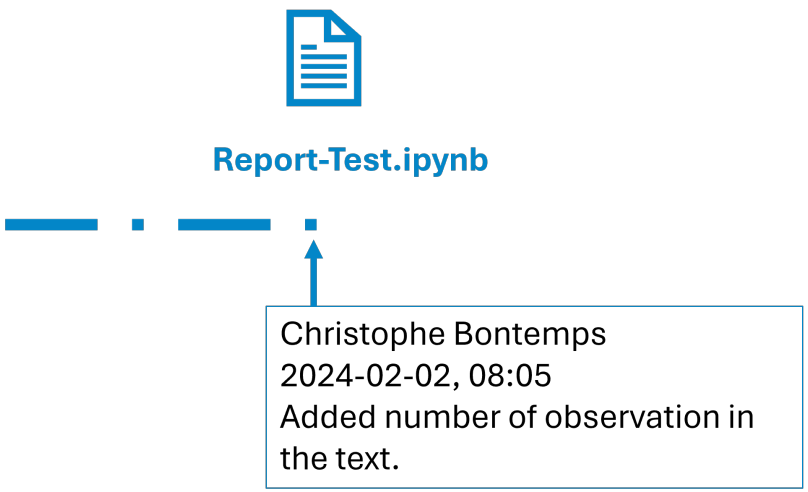
The evolution with Version Control

Record a message (*commit*) for each change!



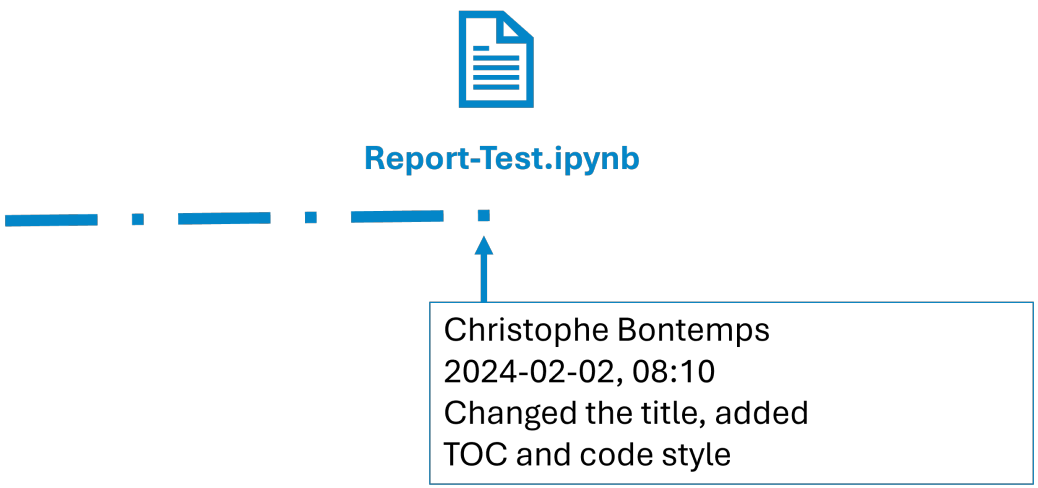
The evolution with Version Control

Record a message (*commit*) for each change!



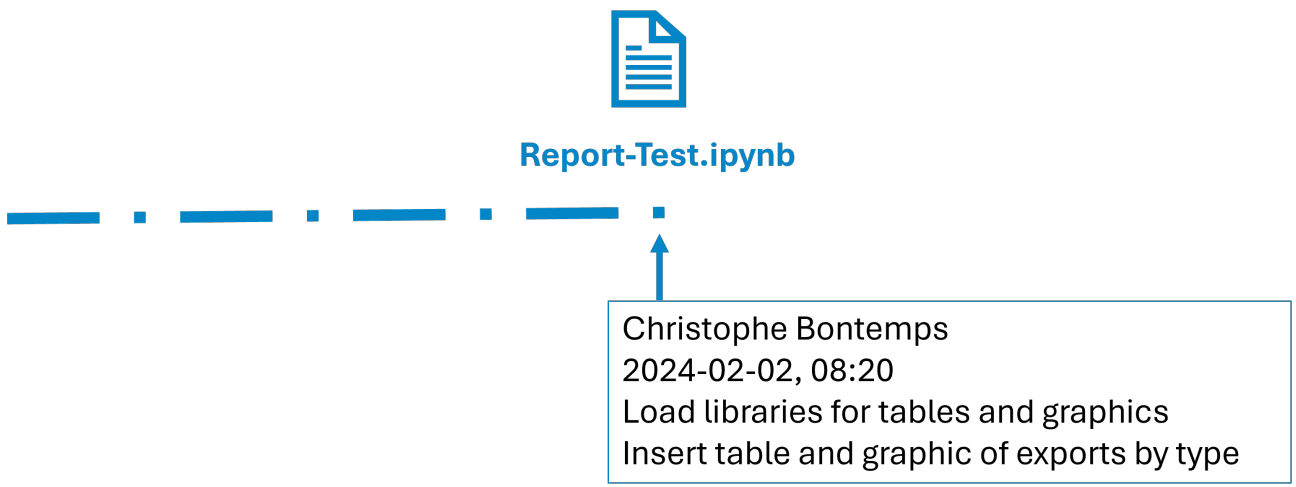
The evolution with Version Control

Record a message (*commit*) for each change!



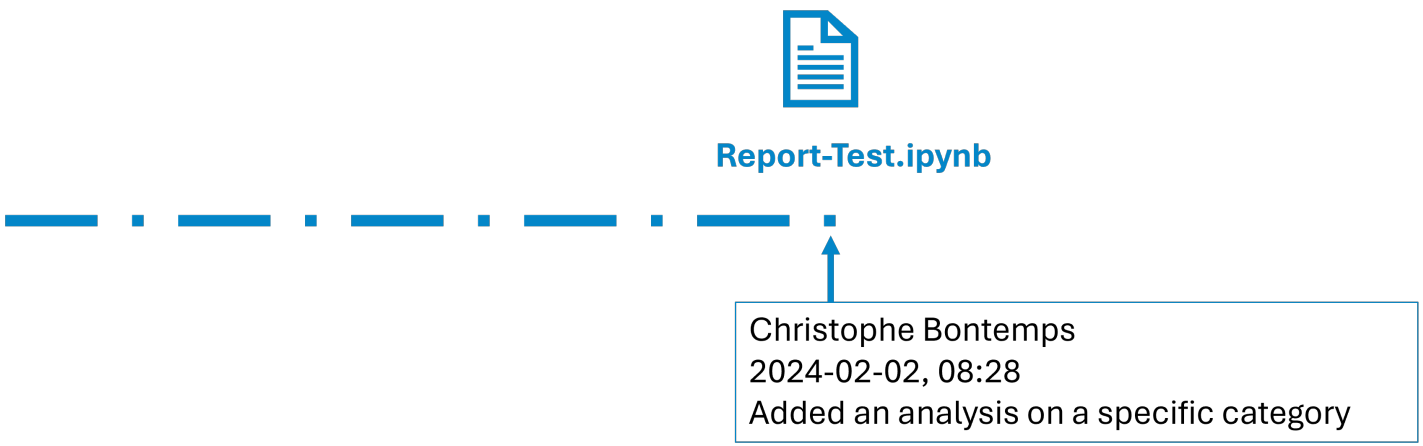
The evolution with Version Control

Record a message (*commit*) for each change!



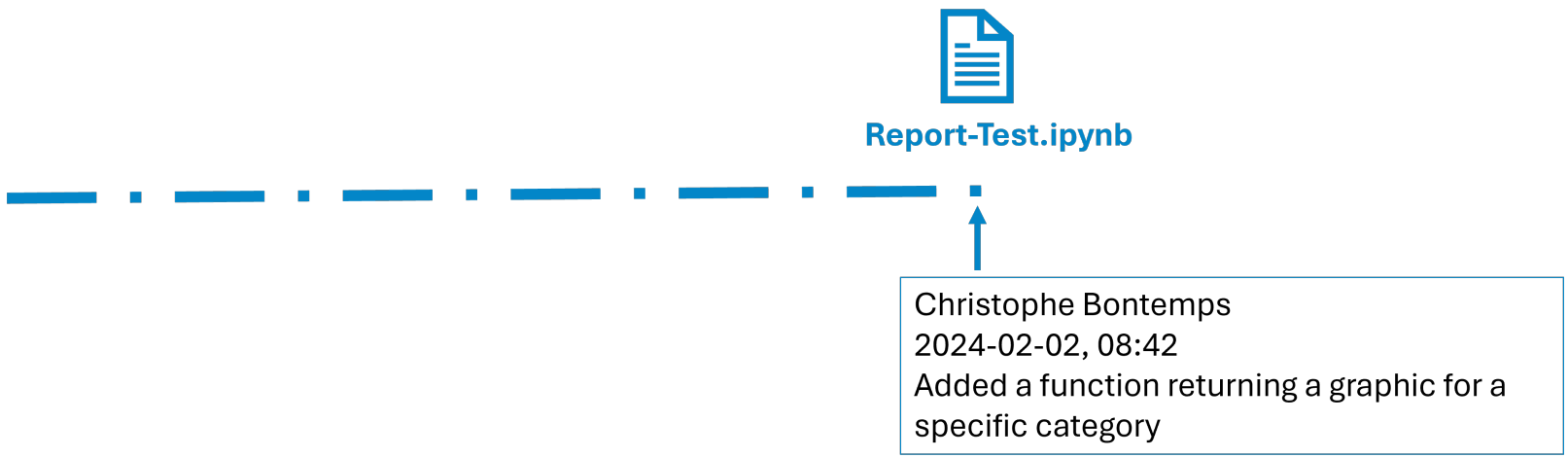
The evolution with Version Control

Record a message (*commit*) for each change!



The evolution with Version Control

Record a message (*commit*) for each change!

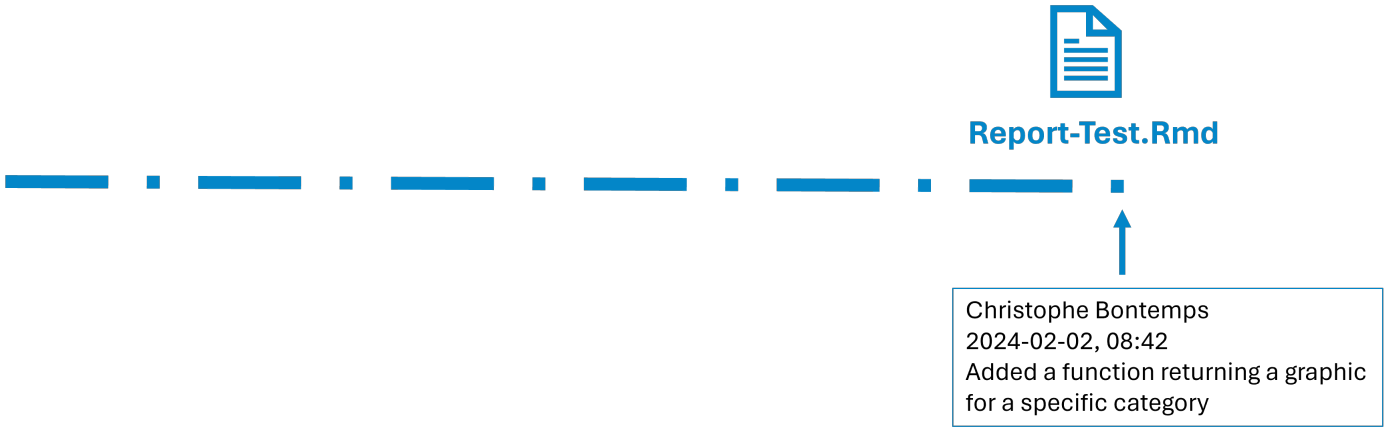


THE EVOLUTION WITH VERSION CONTROL

Record a message (*commit*) for each change!

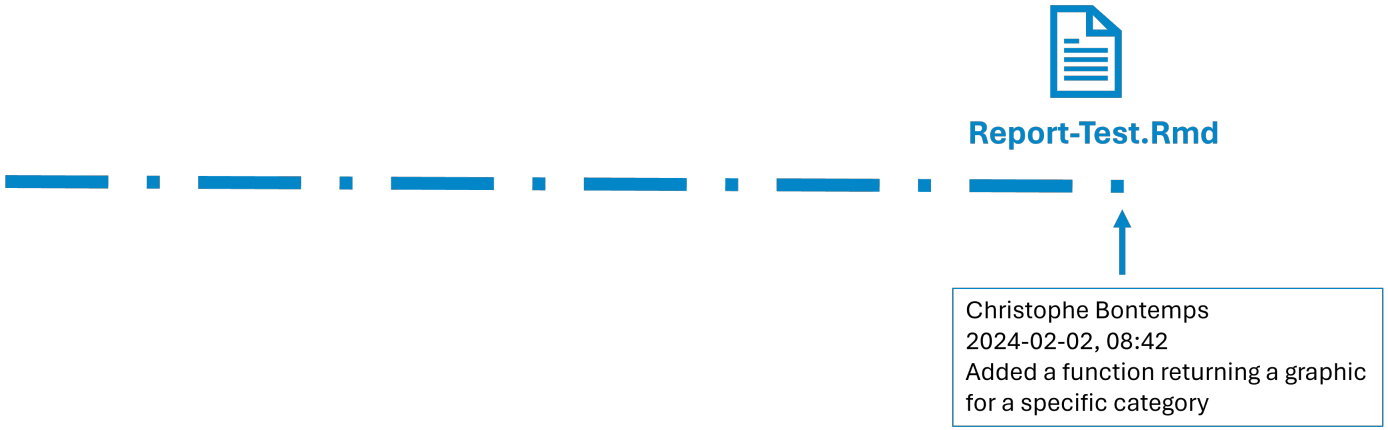
THE HISTORY OF THE FILE IS RECORDED!

Each version is documented (with *commits*)



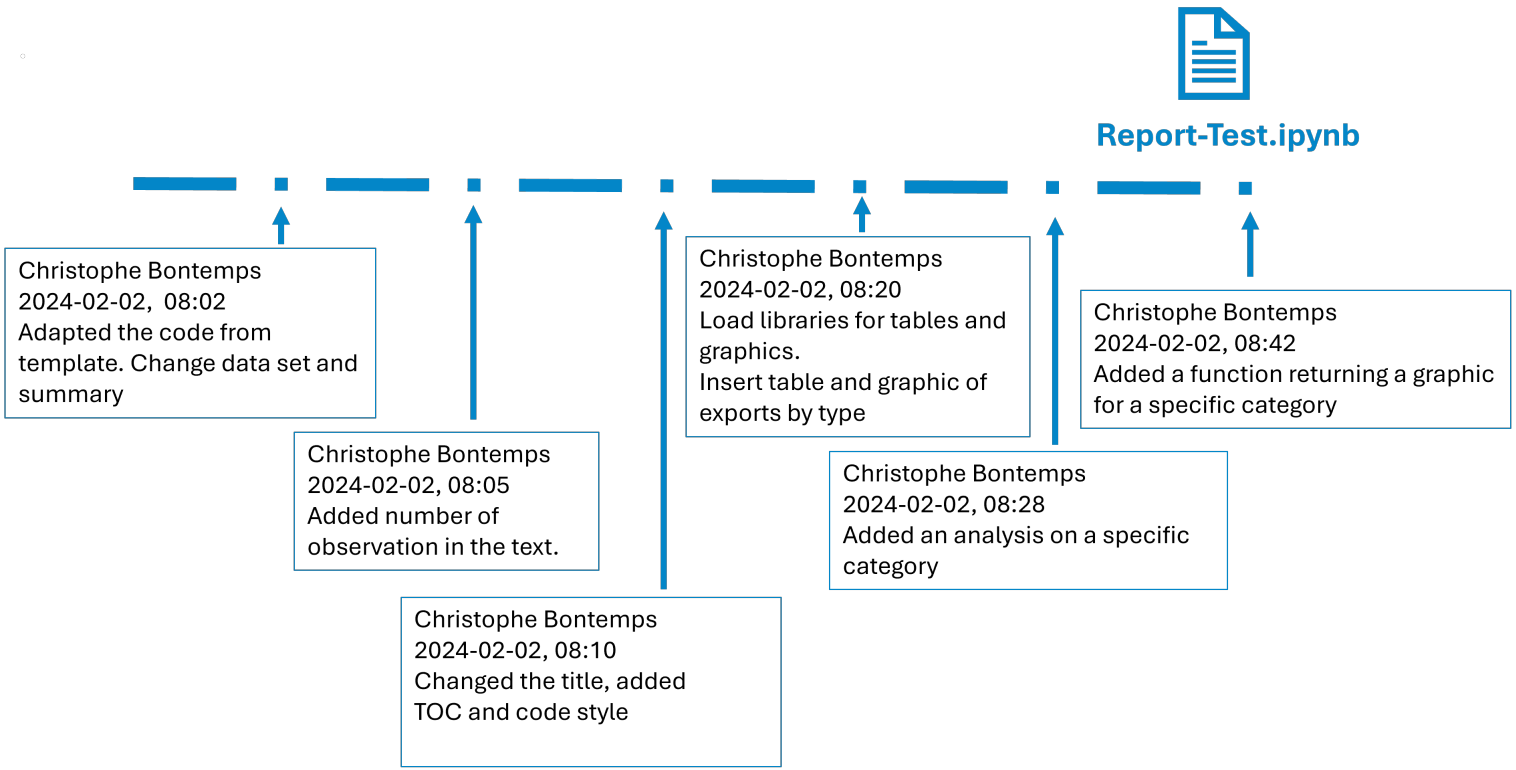
THE HISTORY OF THE FILE IS RECORDED!

Each version is documented (with *commits*)



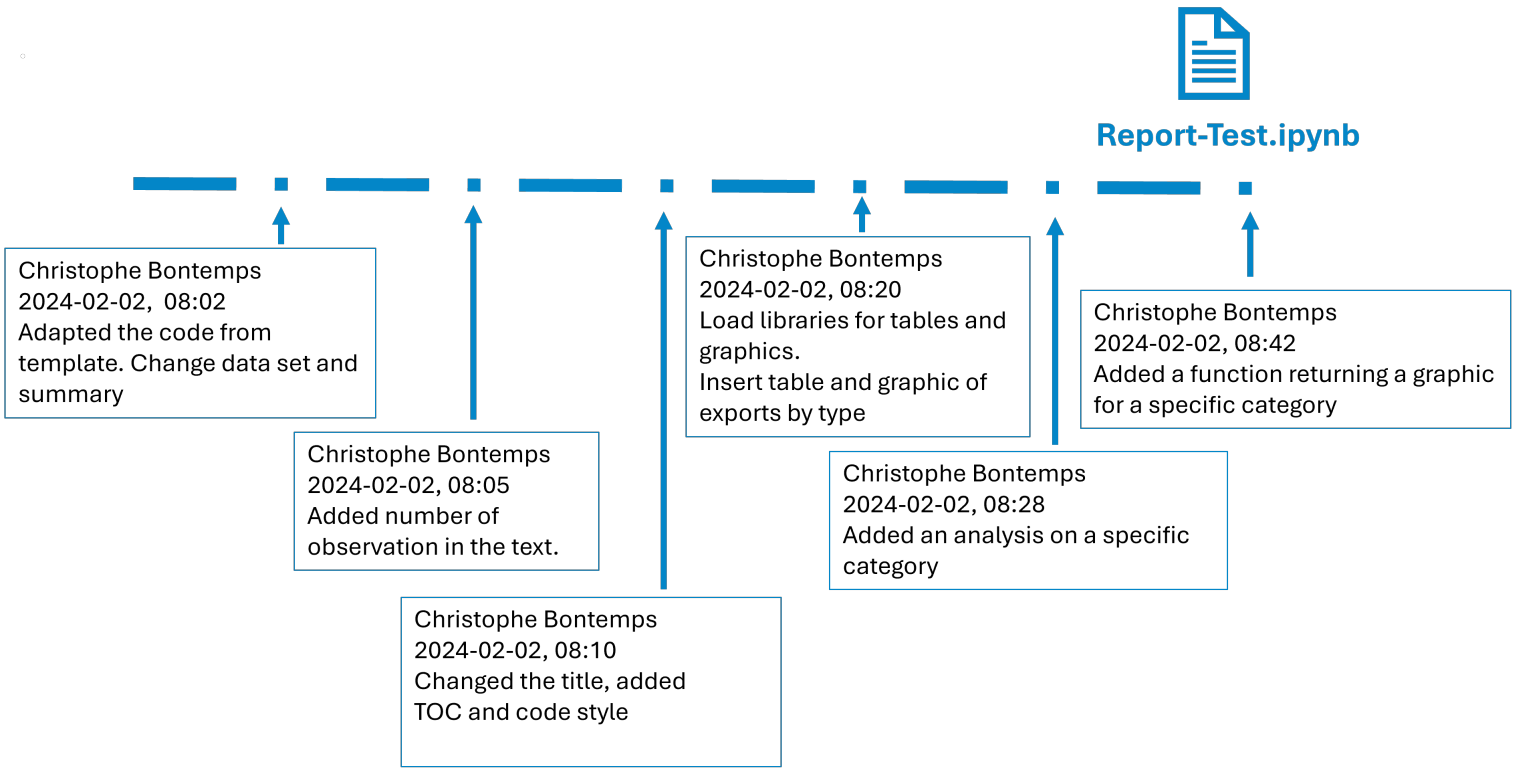
THE HISTORY OF THE FILE IS RECORDED!

Each version embeds the full history!



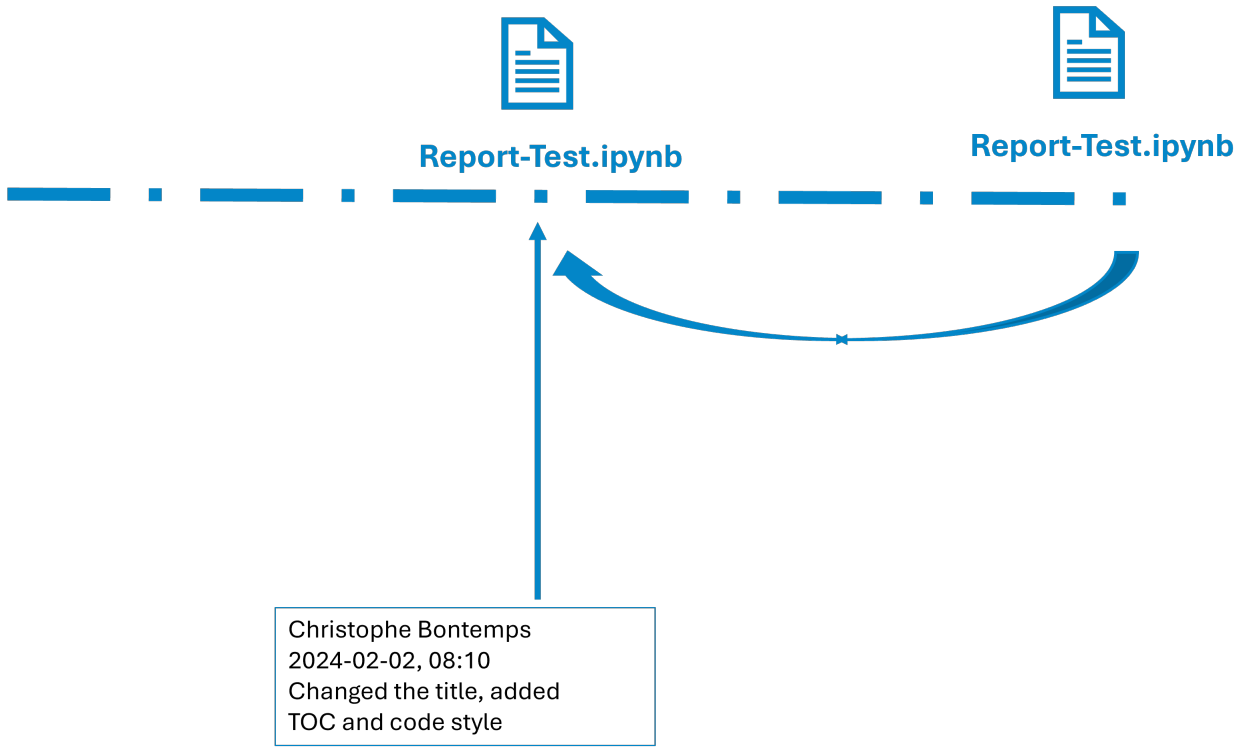
THE HISTORY OF THE FILE IS RECORDED!

Each version embeds the full history!



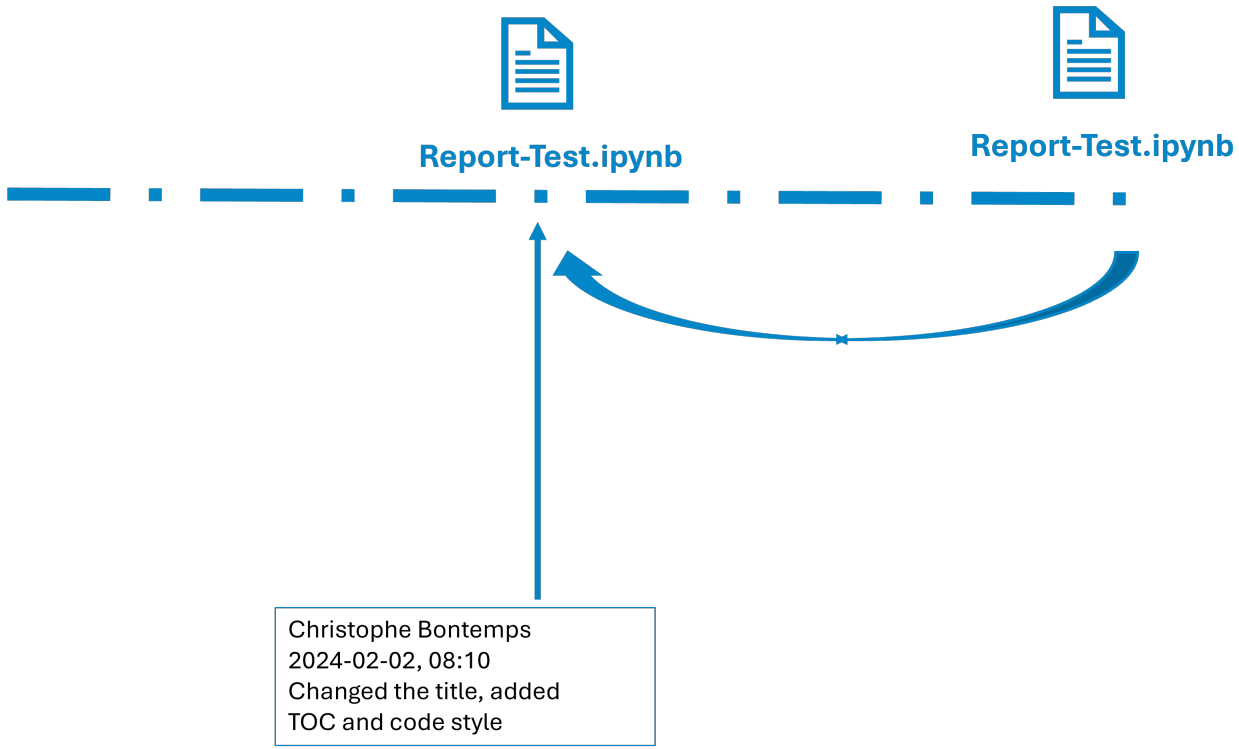
GOING BACK AND "UNDO" IS POSSIBLE

It is possible to review previous version...



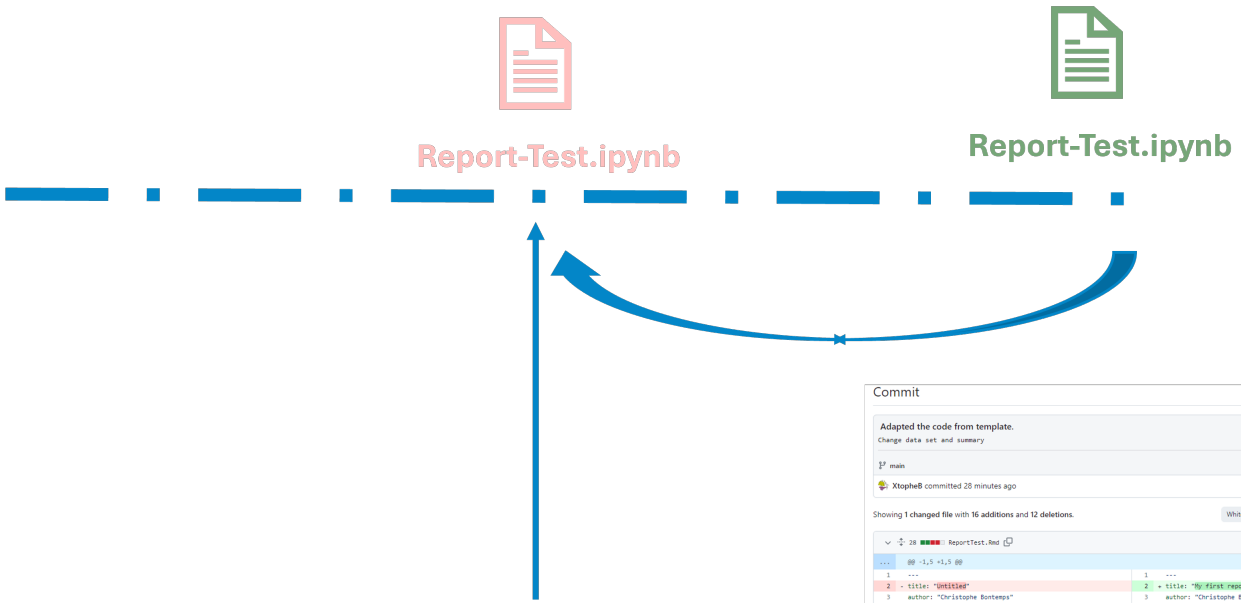
GOING BACK AND "UNDO" IS POSSIBLE

It is possible to review previous version...



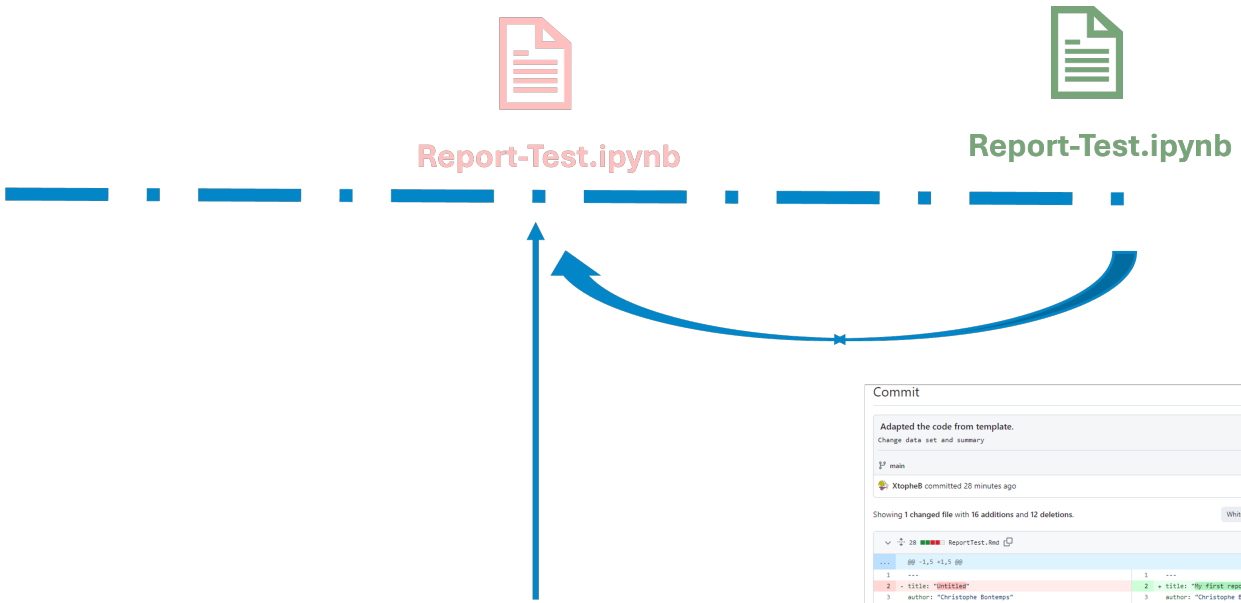
GOING BACK AND "UNDO" IS POSSIBLE

...to compare the changes...



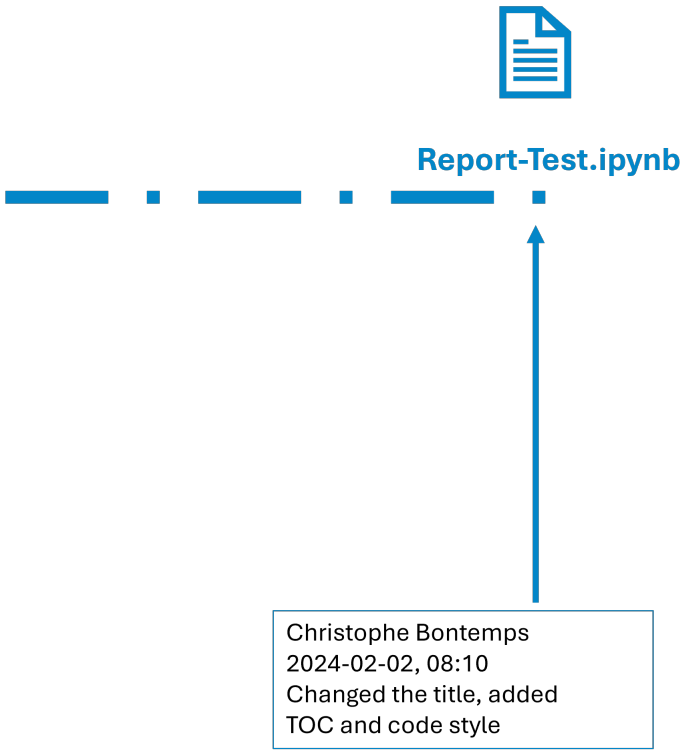
GOING BACK AND "UNDO" IS POSSIBLE

...to compare the changes...



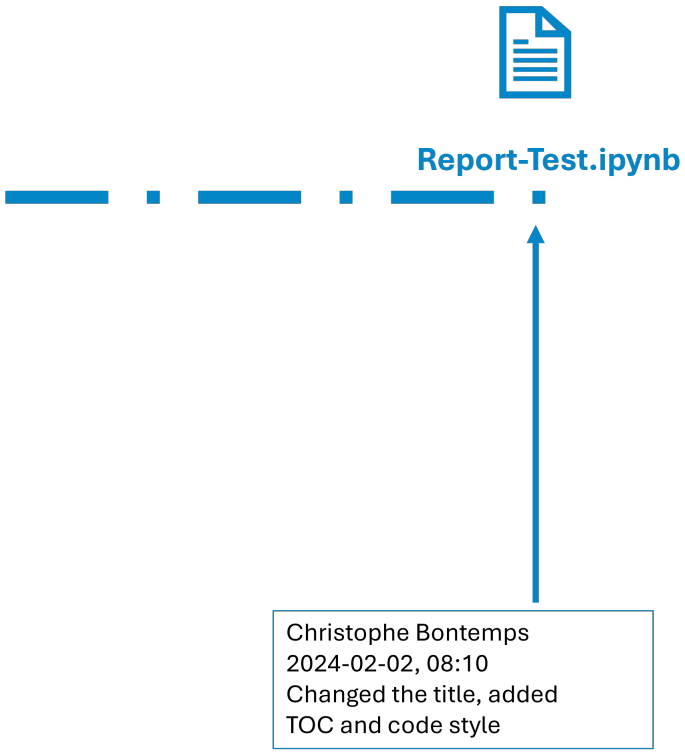
GOING BACK AND "UNDO" IS POSSIBLE

... and to revert to a previous version...



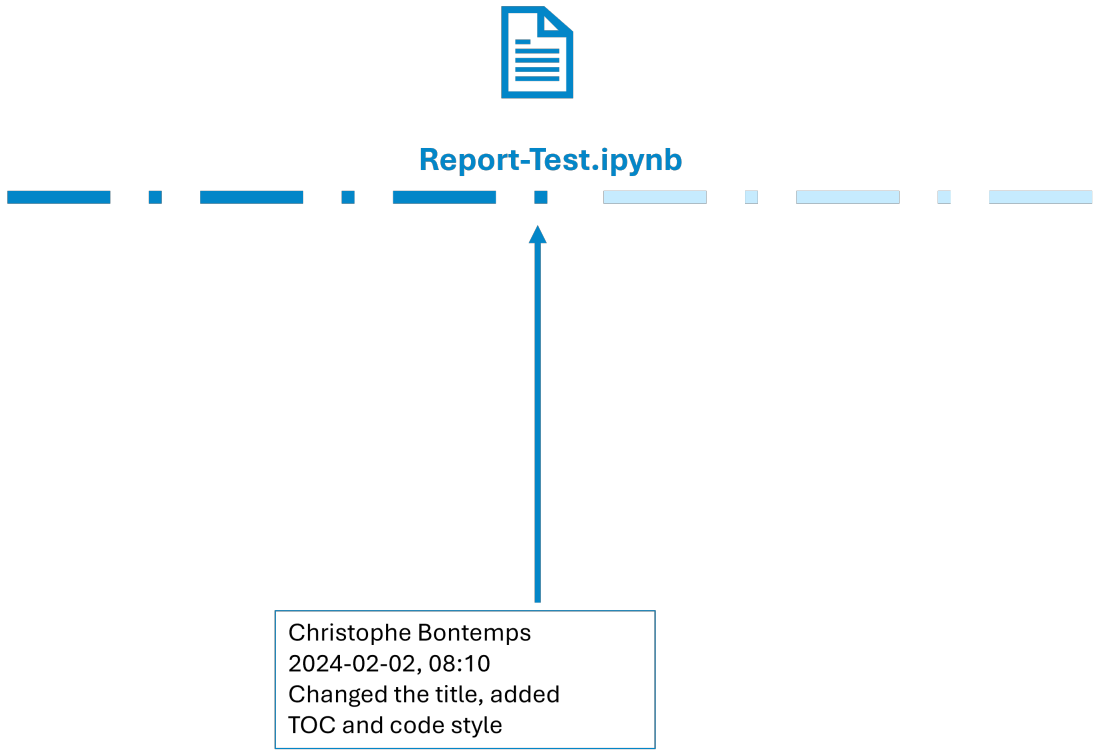
GOING BACK AND "UNDO" IS POSSIBLE

... and to revert to a previous version...



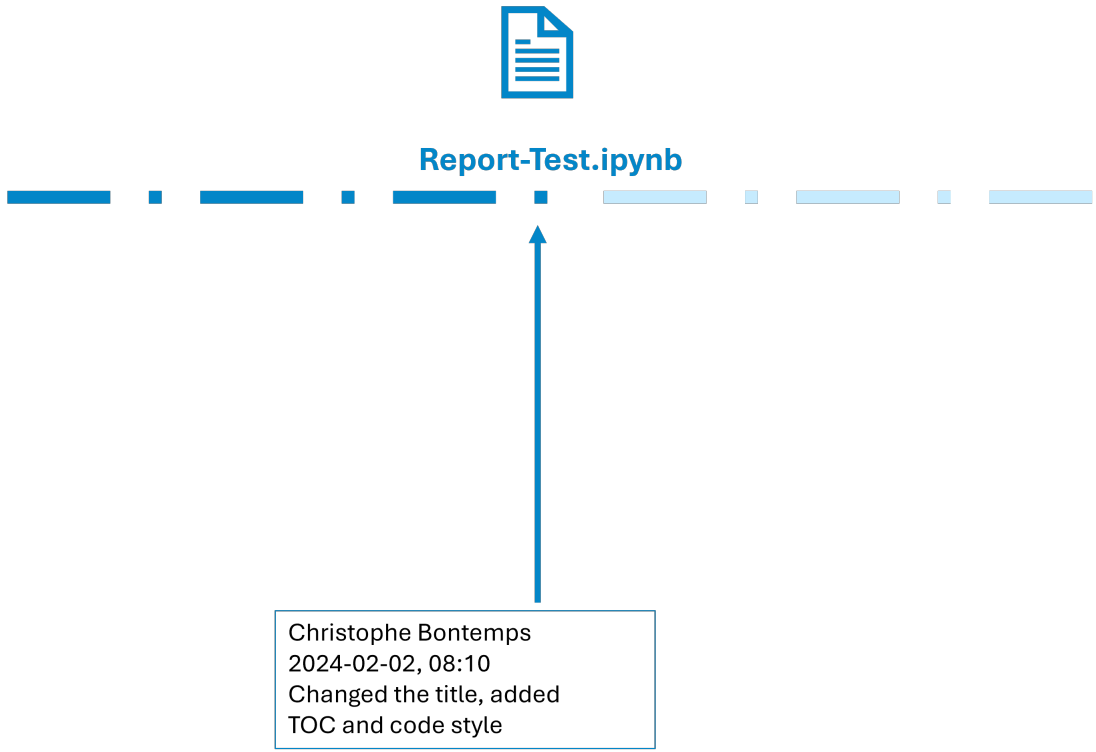
GOING BACK AND "UNDO" IS POSSIBLE

... or *undo* as if nothing happened



GOING BACK AND "UNDO" IS POSSIBLE

... or *undo* as if nothing happened



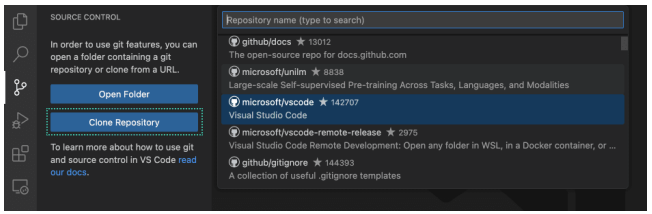
GOOD AND BAD NEWS <–XXXX TO REVISE THIS

Real life is more complex:

GOOD AND BAD NEWS <–XXXX TO REVISE THIS

Real life is more complex:

+ Version Control is integrated in Visual Studio (*& RStudio*)



"Happy Git and GitHub for the useR"

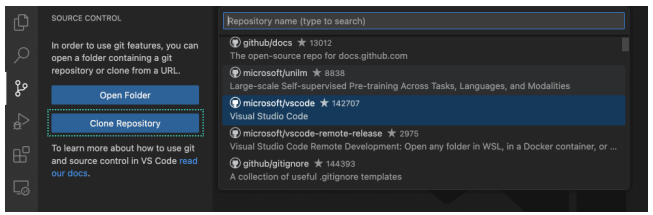
(Jennifer Bryan)

GOOD AND BAD NEWS <–XXXX TO REVISE THIS

Real life is more complex:

+ Version Control is integrated in Visual Studio (*& RStudio*)

↪ Simple operations are easy



"Happy Git and GitHub for the useR"

(Jennifer Bryan)

Good and Bad News <–XXXX To revise this

Real life is more complex:

+ Version Control is integrated in Visual Studio (*& RStudio*)

↪ Simple operations are easy

+ Collaborate on a project

Good and Bad News <–XXXX To revise this

Real life is more complex:

+ Version Control is integrated in Visual Studio (*& RStudio*)

↪ Simple operations are easy

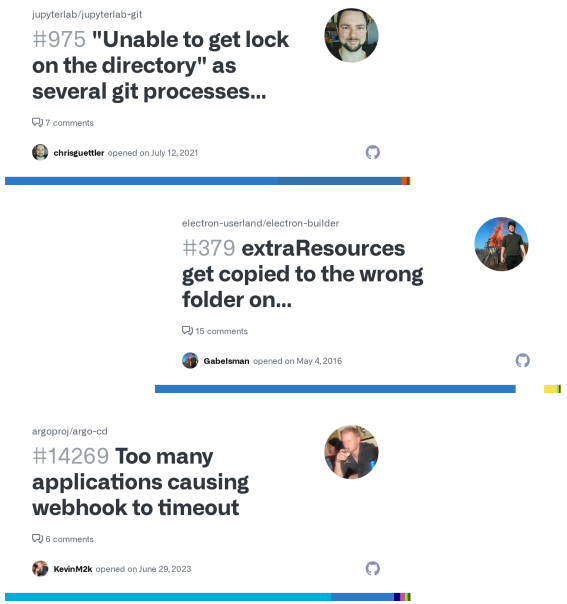
+ Collaborate on a project

↪ Track changes of others

Good and Bad News <–XXXX To revise this

Real life is more complex:

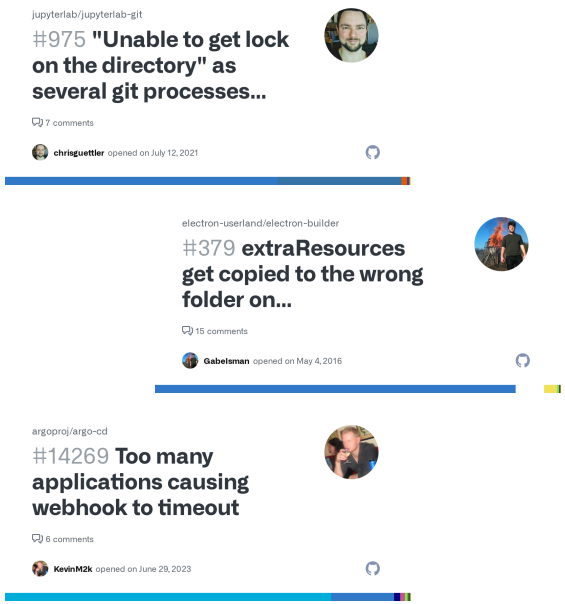
+ Version Control is integrated in Visual Studio (*& RStudio*)
↪ Simple operations are easy
+ Collaborate on a project
↪ Track changes of others
- Git is a bit "*unfriendly*"



GOOD AND BAD NEWS <–XXXX TO REVISE THIS

Real life is more complex:

+ Version Control is integrated in Visual Studio (*& RStudio*)
↪ Simple operations are easy
+ Collaborate on a project
↪ Track changes of others
- Git is a bit "*unfriendly*"
↪ Complex situations appear easily



GOOD AND BAD NEWS <–XXXX TO REVISE THIS

Real life is more complex:

+ Version Control is integrated in Visual Studio (*& RStudio*)

↪ Simple operations are easy

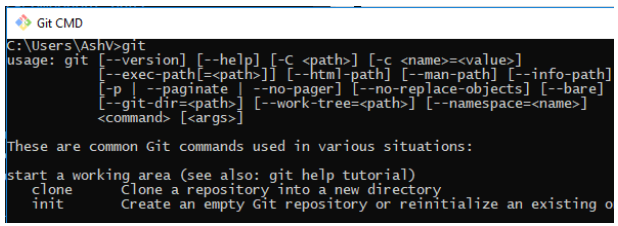
+ Collaborate on a project

↪ Track changes of others

- Git is a bit "*unfriendly*"

↪ Complex situations appear easily

- Git works *mostly* in command mode



Ashish Vishwakarma

VERSION CONTROL IN A NUTSHELL <–XXXX TO REVISE THIS

Version control system:

- Allows to travel back in time

GitHub logo

VERSION CONTROL IN A NUTSHELL <–XXXX TO REVISE THIS

Version control system:

- Allows to travel back in time
- Keeps track of all changes

GitHub logo

VERSION CONTROL IN A NUTSHELL <–XXXX TO REVISE THIS

Version control system:

- Allows to travel back in time
- Keeps track of all changes
- Allows to "undo" at any point

GitHub logo

VERSION CONTROL IN A NUTSHELL <–XXXX TO REVISE THIS

Version control system:

- Allows to travel back in time
- Keeps track of all changes
- Allows to "undo" at any point
- Allows reviewing stages of development

GitHub logo

VERSION CONTROL IN A NUTSHELL <–XXXX TO REVISE THIS

Version control system:

- Allows to travel back in time
- Keeps track of all changes
- Allows to "undo" at any point
- Allows reviewing stages of development
- Allow collaborating on projects

GitHub logo

VERSION CONTROL IN A NUTSHELL <–XXXX TO REVISE THIS

Version control system:

- Allows to travel back in time
- Keeps track of all changes
- Allows to "undo" at any point
- Allows reviewing stages of development
- Allow collaborating on projects
- Backups your work

GitHub logo

TAKEAWAYS <–XXXX TO REVISE THIS

TAKEAWAYS <–XXXX TO REVISE THIS

- Version control is very useful 

TAKEAWAYS <–XXXX TO REVISE THIS

- Version control is very useful 
- Version control requires patience, training and experience

TAKEAWAYS <–XXXX TO REVISE THIS

- Version control is very useful 
- Version control requires patience, training and experience
- Version control is essential for transparency, and reproducibility of any project

TAKEAWAYS <–XXXX TO REVISE THIS

- Version control is very useful 
- Version control requires patience, training and experience
- Version control is essential for transparency, and reproducibility of any project
- Installing Git within RStudio can be tedious but worth it!

TAKEAWAYS <–XXXX TO REVISE THIS

- Version control is very useful 
- Version control requires patience, training and experience
- Version control is essential for transparency, and reproducibility of any project
- Installing Git within RStudio can be tedious but worth it!
- There is guidance, tutorials and helping blogs...

Useful resources

- The UK government RAP website.
- UK best practice documentation.
- A free RAP course to teach you all you need to know.
- How the Data Science Campus sets its coding standards.
- A new open-source book from the Alan Turing institute setting out how to do reproducible data science.