

# Intro les 2

Jan-Willem & Raoul Grouls

## Table of contents

<b>Get the latest changes</b>	<b>1</b>
Clean working tree . . . . .	1
Clean working tree . . . . .	2
Git pull . . . . .	2
Inspect changes . . . . .	2
<b>Quarto</b>	<b>3</b>
Using Quarto . . . . .	4
Using Quarto . . . . .	5
Installing Quarto . . . . .	5
Create pdfs . . . . .	6
<b>Visualization</b>	<b>7</b>
Explanations . . . . .	7
Proofs . . . . .	9
Proofs . . . . .	12
Confusion . . . . .	13
Confusion . . . . .	14
Toolkit . . . . .	14
Toolkit . . . . .	15
Toolkit . . . . .	15

## Get the latest changes

### Clean working tree

First, you will need to make sure you have the latest changes. To do that, you cd to the DME22 directory on your VM.

- Once you are there, you can run `git status` to check your branch.
- Your branch should be clean before you checkout to master!
- if not, make sure you are in another branch than the master branch, and either `git add` and `git commit` the things you want to keep, or remove the changes.

A fast way to revert all the changes you dont want to keep, is: `git checkout -- .`

## Clean working tree

Ok, assuming everything is clean, you should get a message like this:

```
On branch feature/les3
Your branch is up to date with 'origin/feature/les3'.
nothing to commit, working tree clean
```

## Git pull

Now, checkout to master and pull the latest changes.

```
$git checkout master
```

You should see a message that looks like:

```
Switched to branch 'master'
Your branch is behind 'origin/master' by 15 commits, and can be fast-forwarded.
(use "git pull" to update your local branch)
```

So, lets do that:

```
$git pull
```

## Inspect changes

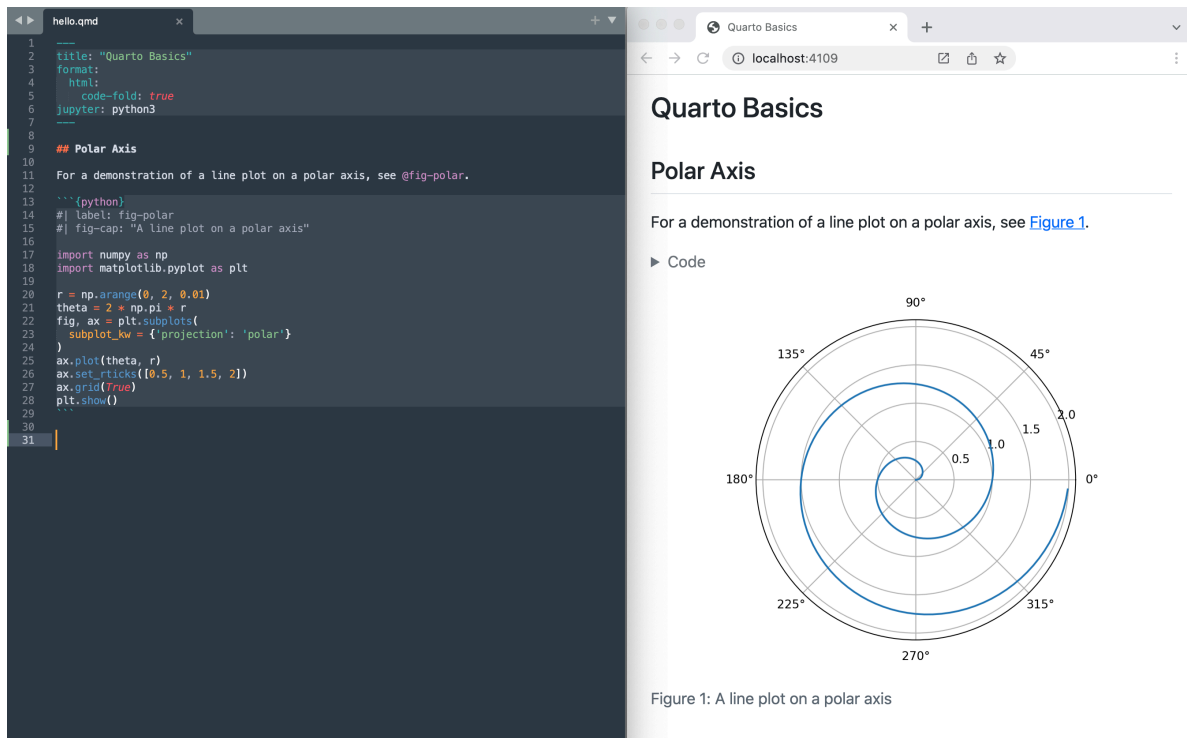
You will get a full report of all the files we have changed and added since last time you pulled. If you want to check that out in more detail, you can either use

- Git Graph inside VSCode
- or you can use the command line with a command like

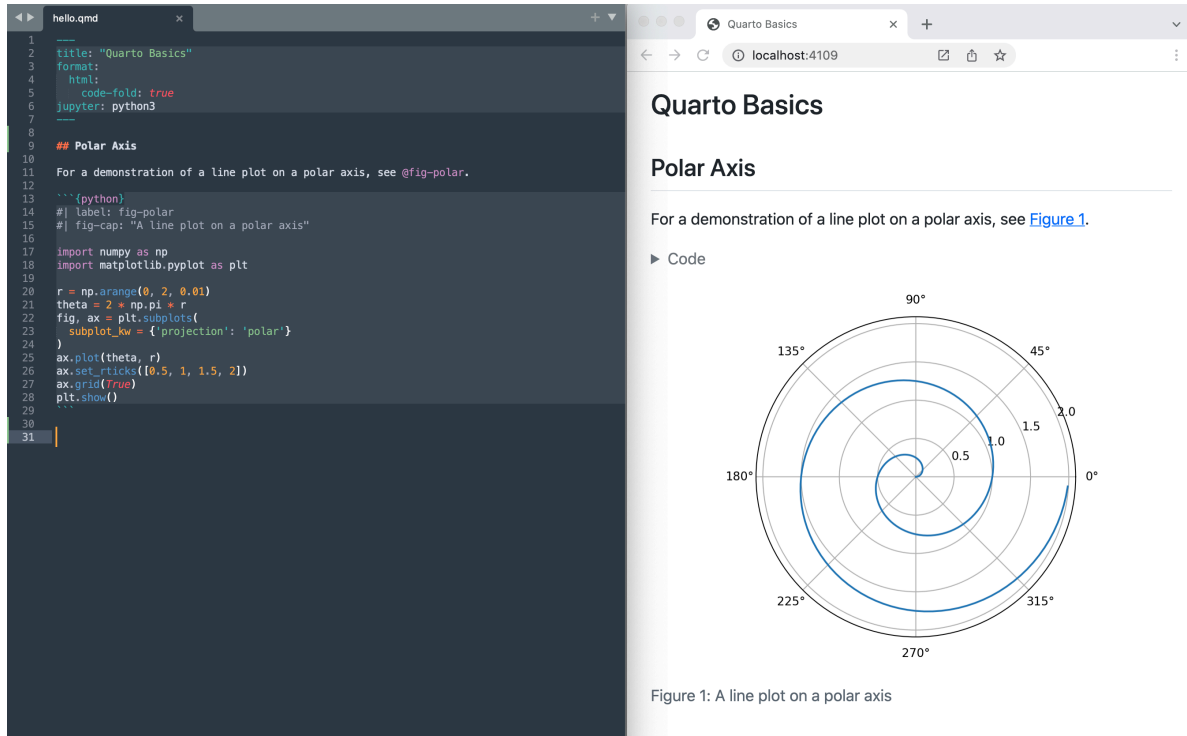
```
git log --stat --oneline
```

- You exit this view by pressing q.

# Quarto



## Using Quarto

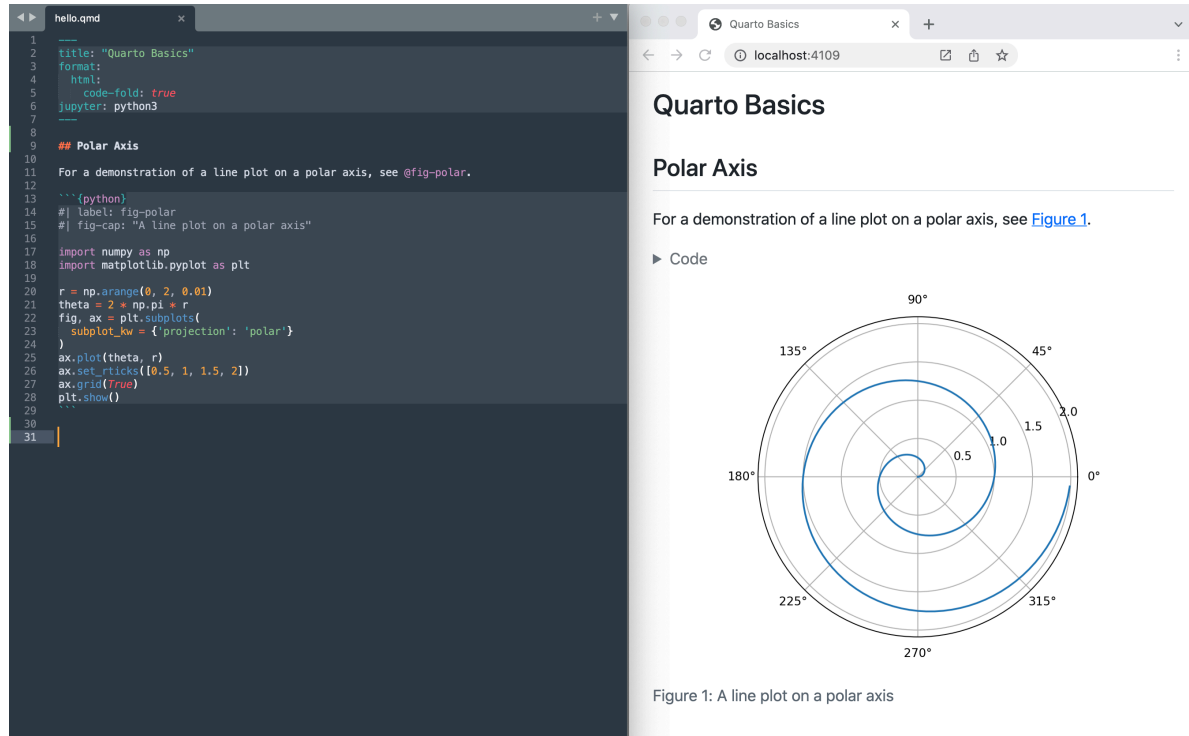


Quarto is an

“academic, scientific, and technical publishing system built on Pandoc.”

It is a very easy way to make nice reports, presentations etc from markdown files.

## Using Quarto



If you want to install it locally too, you can simply download an installer from their website. However, on the server we can't click on anything so we will need to do that with the command line.

## Installing Quarto

If you follow these steps, you have Quarto installed:

1. Go to the home directory of your VM by `cd ~`
2. Download the latest version with: `sudo curl -LO https://quarto.org/download/latest/linux/amd64/quarto-linux-amd64.deb`
3. Enter your VM password (that you wrote down somewhere during the setup of your VM) to start the download.
4. To install this `.deb` file, we need `gdebi`. You can check with `which gdebi` that it's not on the VM.
5. We can install it with `sudo apt install gdebi-core`. After that, `which gdebi` returns the location.
6. And now we can use `gdebi` to install the `.deb` file with `sudo gdebi quarto-linux-amd64.deb`

## Create pdfs

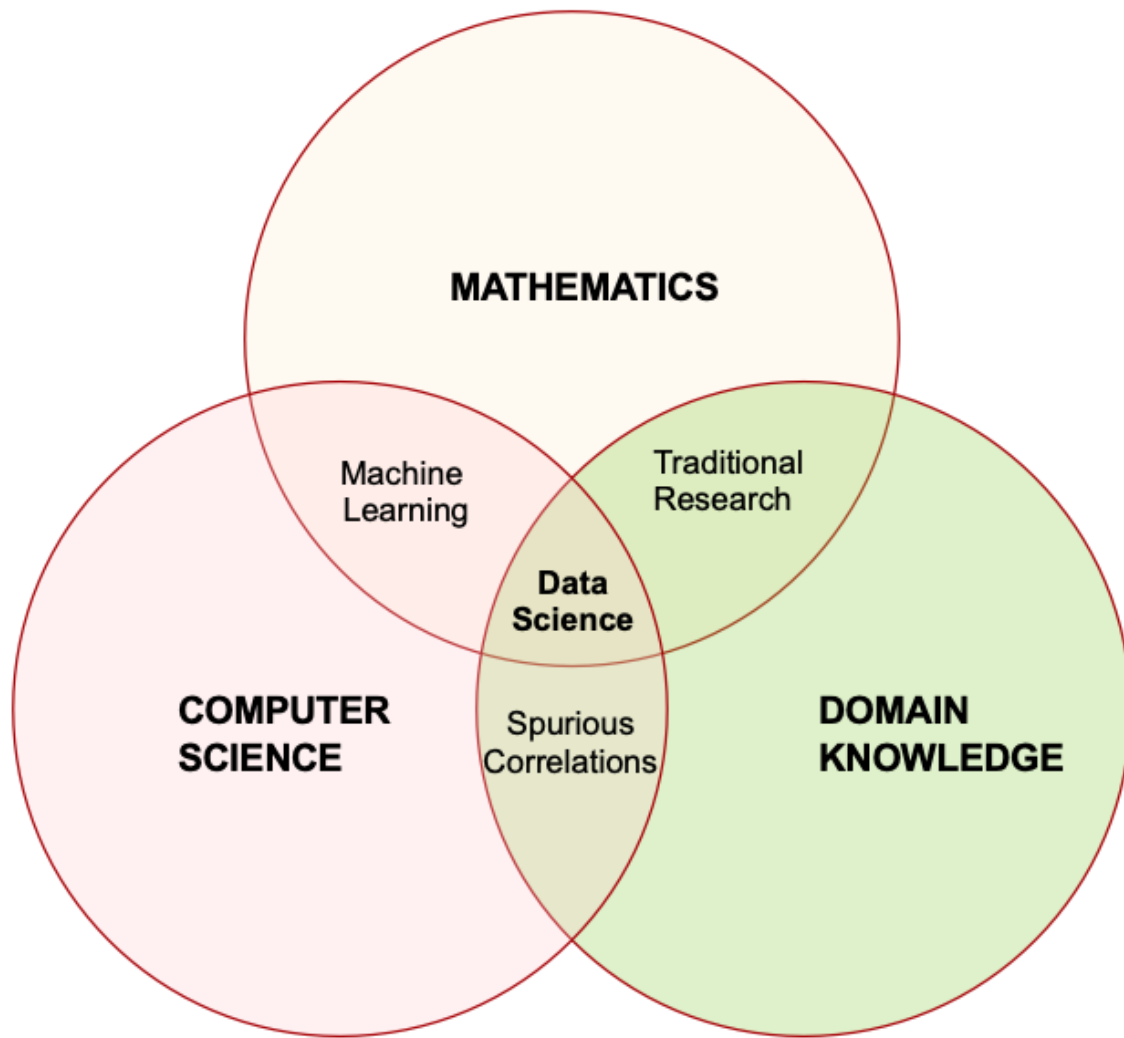
To create pdfs, we need a latex distribution. You can install a minimal version of that with

```
quarto install tool tinytex
```

If you install the `vscode-pdf` extension in VS code, you can view the pdf with VScode by clicking on it.

## Visualization

### Explanations



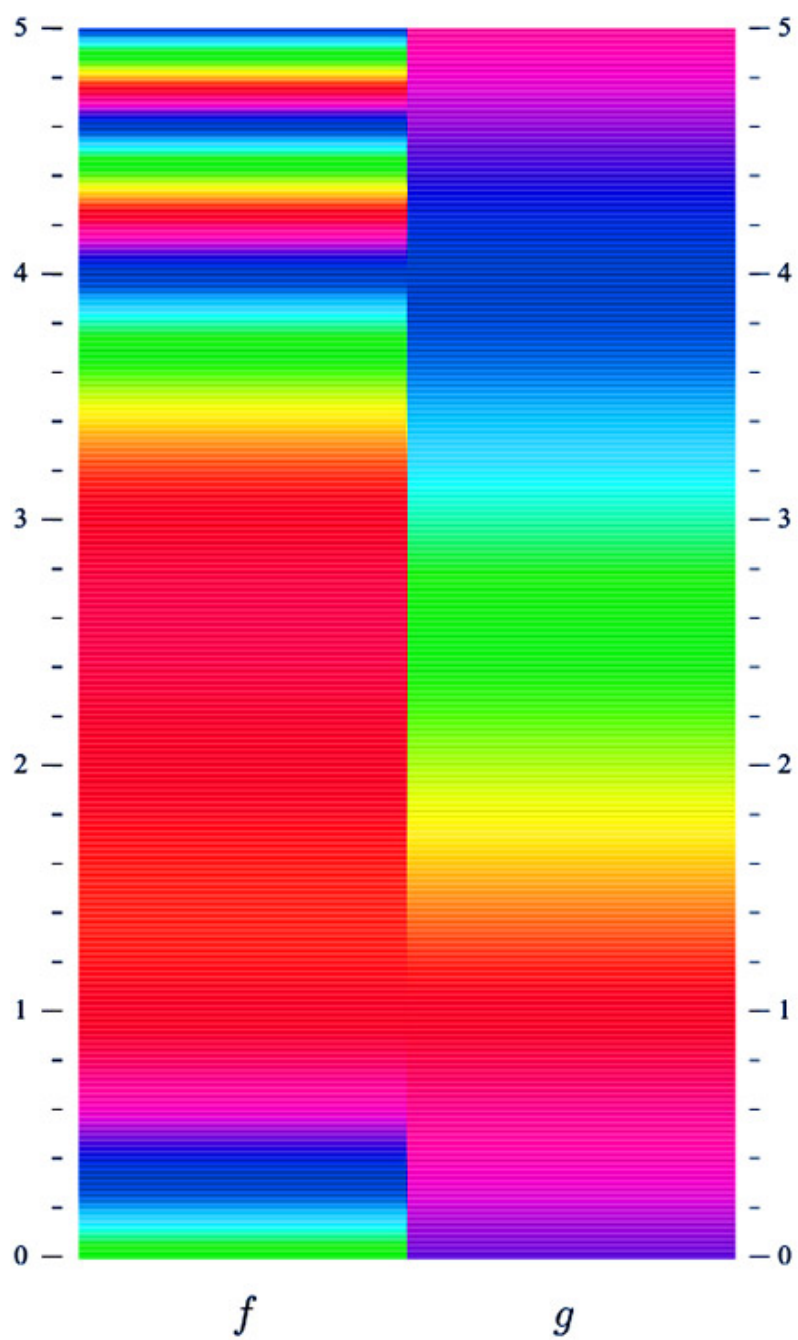
As humans, we process images much faster than text. Without an image, the information on the left would be harder to explain.





## Proofs

## Chromatic



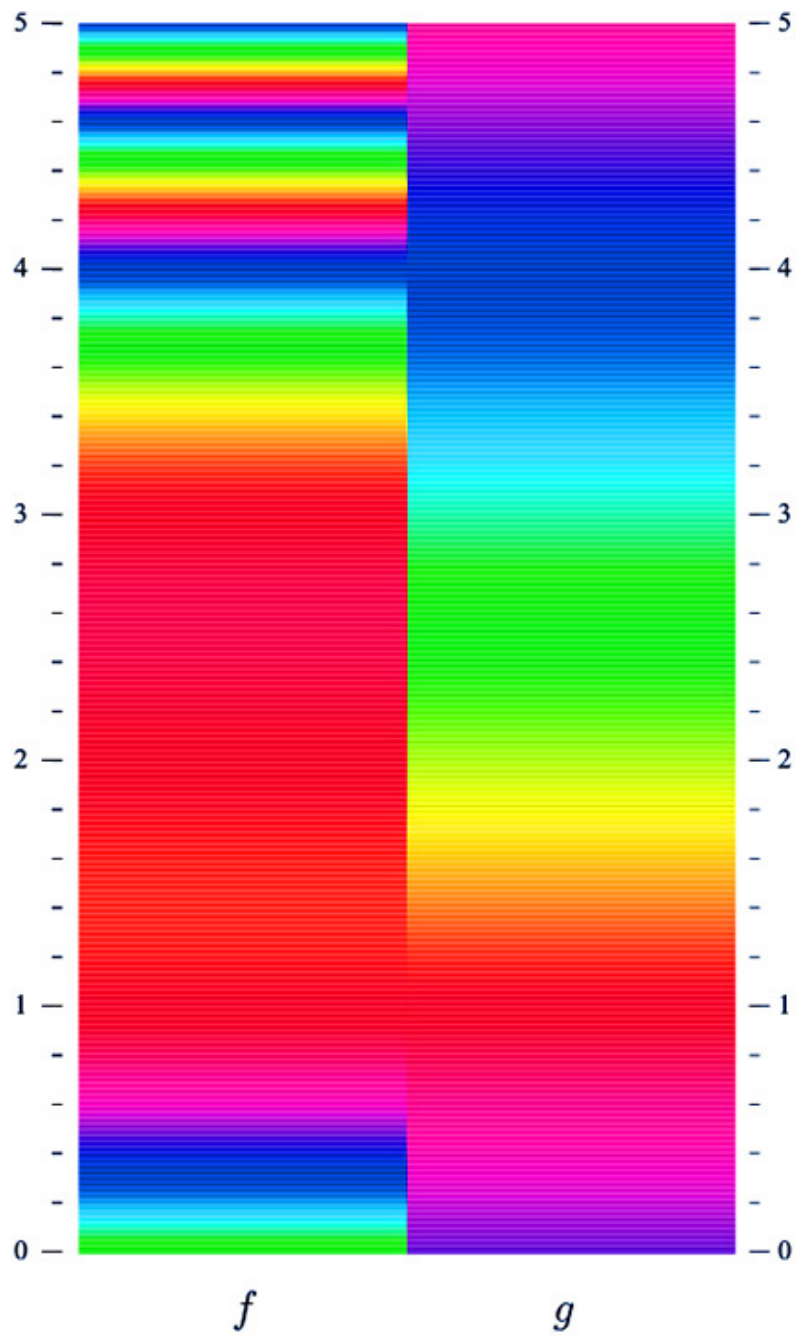
How people create mathematical proofs is something that changes over the centuries.

A proof has a social aspect: you need to convince your listener.

That means that in some cases visualisations are an acceptable proof.



# Chromatic



In the image, you can see that

$$x^3 - 6x^2 + 11x = 2x - 2$$

will be true for  $x = 4$  and  $x = 1$  because the colors are the same. On the left you see  $x^3 - 6x^2 + 11x$  and on the right  $2x - 2$ .

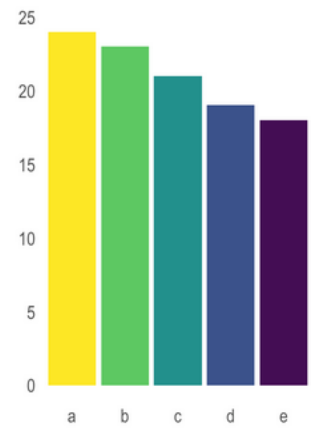
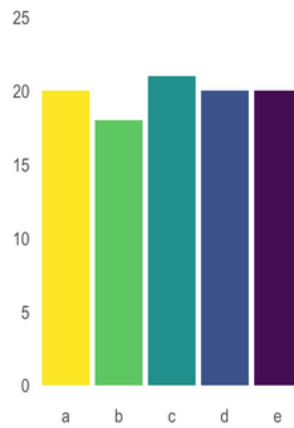
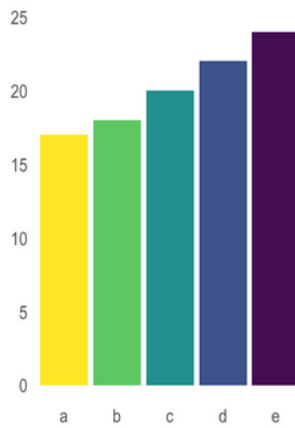
### Confusion



Unfortunately, visualisations can be confusing too.

Can you tell how these groups have evolved over time?

## Confusion



CODE

Now compare that to this.

Much better, right?

You can make it even worse by adding: - 3d effects - percentages that do not sum to 100 - too many items

## Toolkit

These charts are probably good enough for 80% of what you need to do as a data scientist:

- scatterplot
- lineplot
- histogram / boxplot

## Toolkit

With these three, you can spot things like:

- clusters of datapoints
- white spots (missing data)
- trends over time
- differences between distributions

## Toolkit



In the CRISP-DM cycle, visualisation will help you with:

- data understanding
- modeling
- business understanding

Keep in mind that most of your communication with the business will be visual. You can use visualisations to check assumptions and to find anomalies in the data.