REPRODUCIBLE RESEARCH

Someone who is reading your analysis, it should be reproducible.
Will talk about how some basic principles can be employed in R-studio and coding to create better reproducible complex analysis.

Course Content

- Structuring and organizing a data analysis
- Markdown and R Markdown
- knitr / RPubs
- Reproducible research check list
- Evidence-based data analysis
- Case studies in air pollution epidemiology and high-throughput biology

If data is analyzed, reproducibility is important.
Example of music where a score or card is written to communicate how to play, what was done and how to do it.

There is no notable agreed upon system for expressing what was done and how it was done.

In Science, Replication is the most important element of the research to strengthen scientific findings.

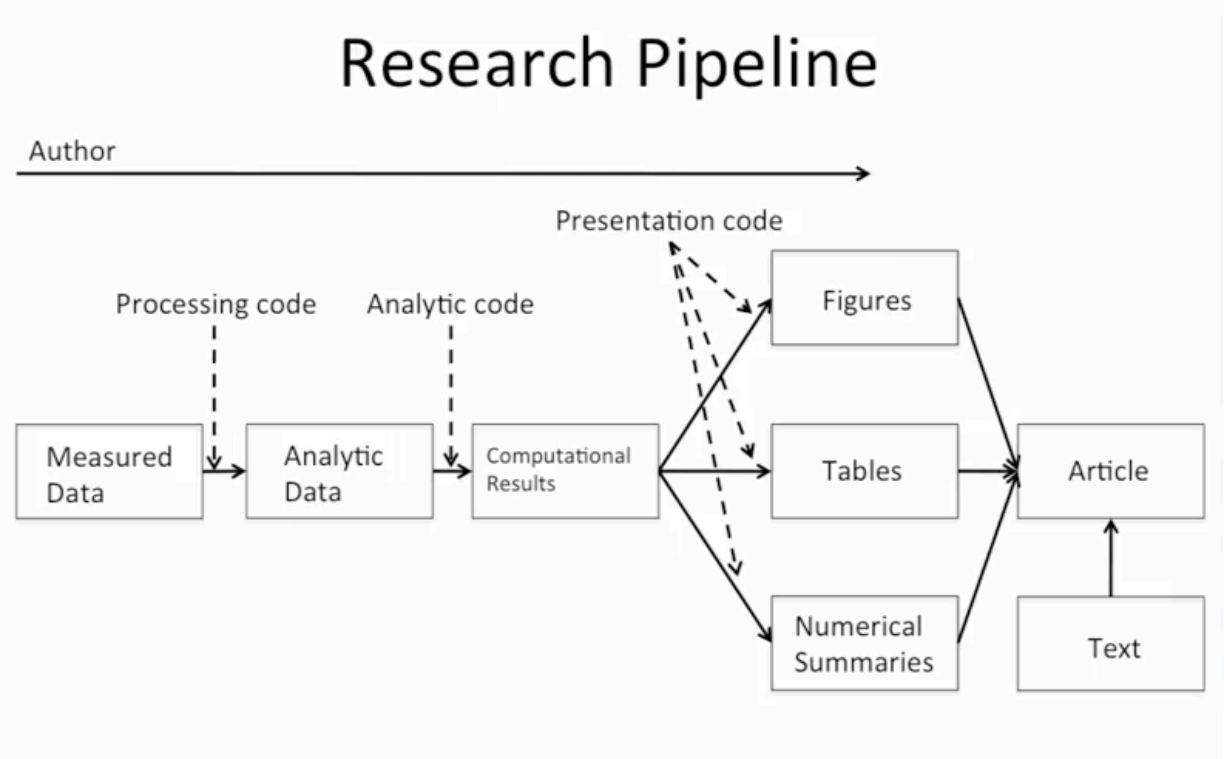
Bigger studies often are very hard to study.

Trying to create something between replication and doing nothing.

Why do we need it ?
To make data and computational analysis ready for others to use. Useful to validating the data analysis

Have a computational analysis of everything that is there.

Research pipeline. Author does left to right , reader goes right to



.

Analytic data used should be available. Slightly better than raw data.
Analytic code should be available. Documentation and some standard means of Distribution.

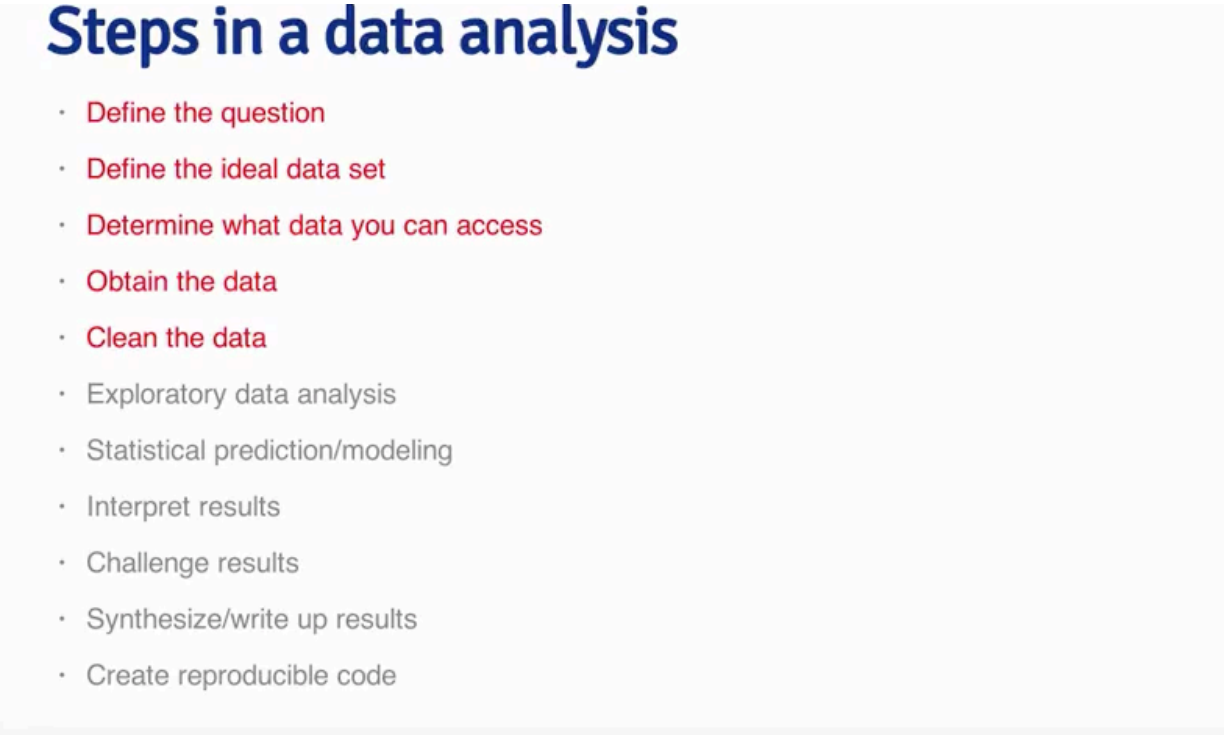
Literate Statistical Programming
Article should be described in a stream of Text and code. Analysis is divided into text and code to show things work.

Sweave - uses Latex and R documentation language. Might lack features for caching and all .

Knitr - package uses R as a programming language, Latex, Markdown and HTML can be used to create reproducible research.

1. Script Everything: Write everything down, score for data analysis. In Studio, save the R script in a blank test file.

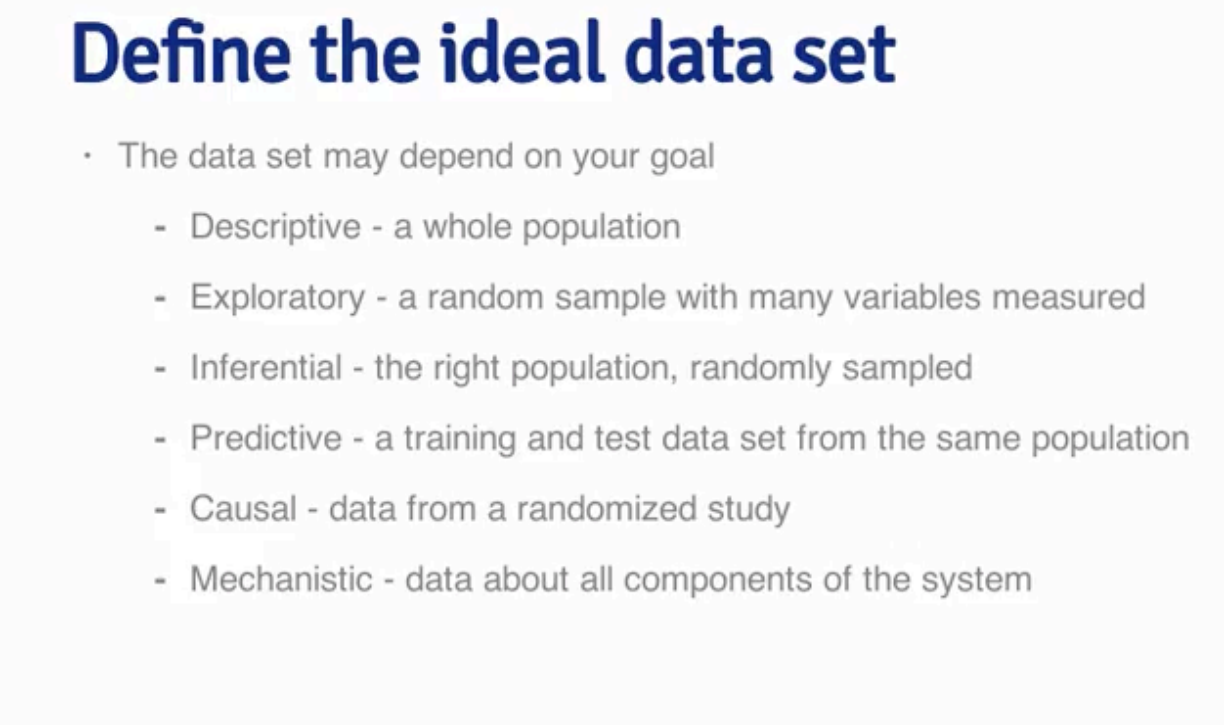
What are the pieces of data analysis ?



First step is to come up with a proper question. This is the most powerful dimension reduction tool you can come up with. If you narrow down the question it will help you reduce the noise that you would see in a dataset. Think about what kind of question you are asking before you dive into the analysis.

Scientific questions, can I do this ? or that ? make questions more concrete and translate it in the lingo of the data analysis quantitative tools .

Ideal data set might be.



Think about what are the kind of data that you would be able to access. Think about what is there in net for free, or something else that you can generate.

Obtain raw data, keep track of where it came from ,list sources. If internet source, record Url and time and date of when the data was available. Document how you got it.

Clean the data. Understand how the preprocessing was done, how the data came. Reformat the data, sub sample the data set if it is very large.

<<check ipnyb notebook for example>>

Organizing a data analysis:
Data, figures, Rcode and Text,

The Raw data can come in many forms. Use some tools to generate proper formatting and store in analysis folder.

Processed data would be more clearer and often comes in a table. Processing script is necessary to document what code was used to transform the raw data into processed data.

Exploratory figures need not be a part of the final report.
final scricpts would be much more commented with processing data and all to see

the chain of events that was used.

R markdown files are useful to embed code and show scripts and their results.

Add readme files if not using markdown files.