浙江大学爱丁堡大学联合学院

ZJU-UoE Institute

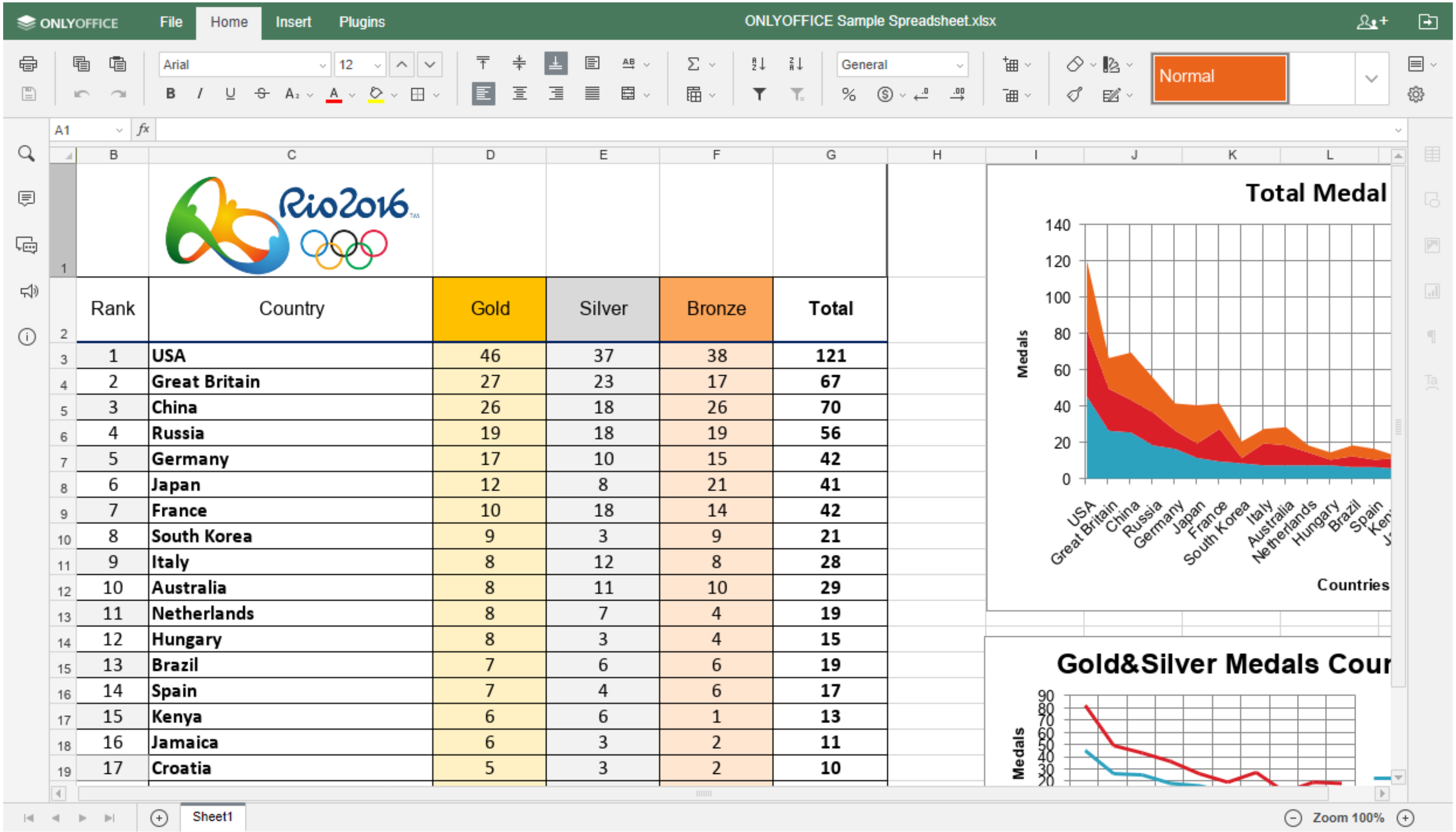
ADS2 Lecture 13
Planning a data analysis pipeline

Dr Duncan MacGregor duncan.macgregor@ed.ac.uk

Semester 1, Week 13

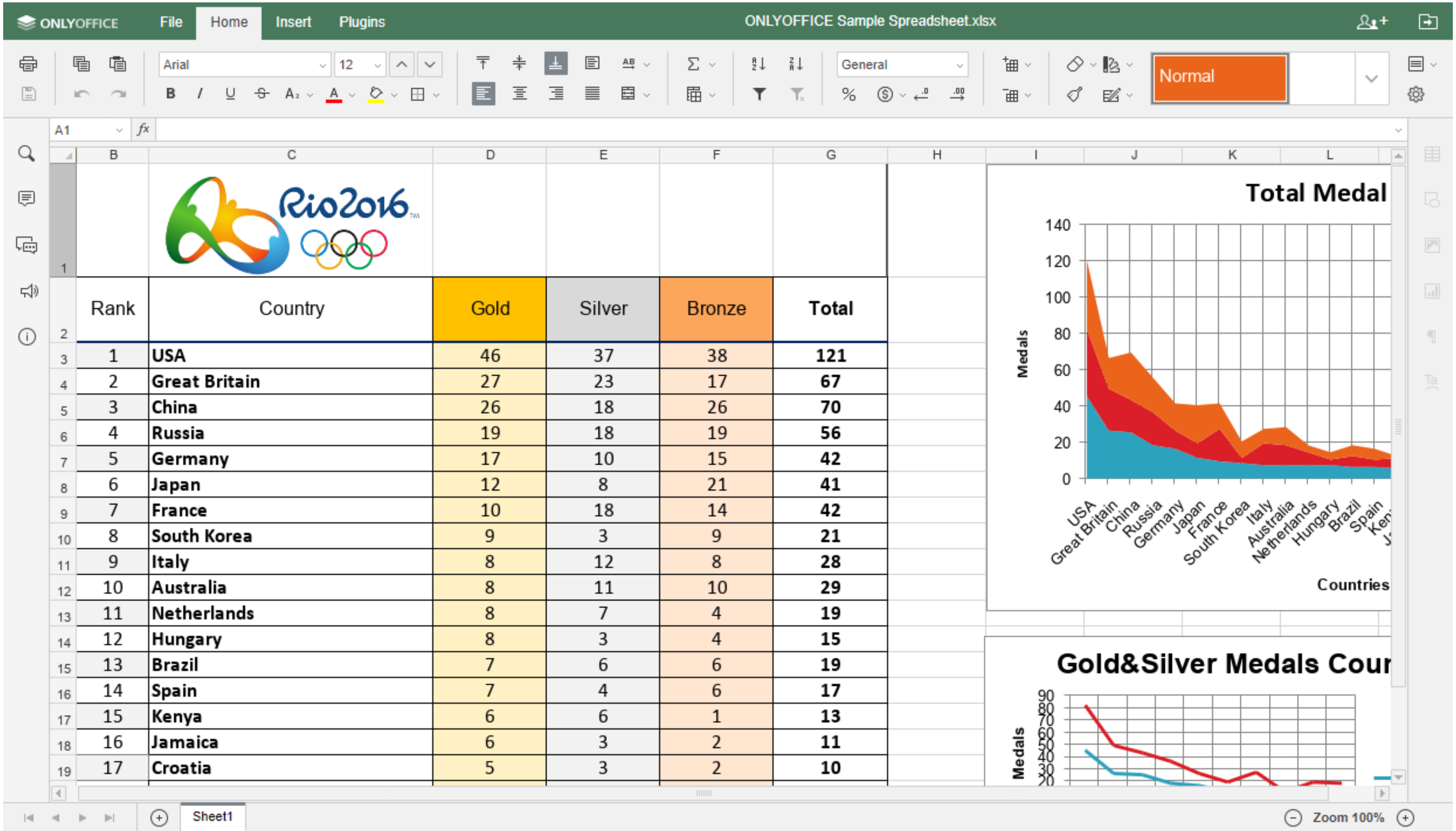
2023-24

Yay, data!



Rank	Country	Gold	Silver	Bronze	Total
1	USA	46	37	38	121
2	Great Britain	27	23	17	67
3	China	26	18	26	70
4	Russia	19	18	19	56
5	Germany	17	10	15	42
6	Japan	12	8	21	41
7	France	10	18	14	42
8	South Korea	9	3	9	21
9	Italy	8	12	8	28
10	Australia	8	11	10	29
11	Netherlands	8	7	4	19
12	Hungary	8	3	4	15
13	Brazil	7	6	6	19
14	Spain	7	4	6	17
15	Kenya	6	6	1	13
16	Jamaica	6	3	2	11
17	Croatia	5	3	2	10

Yay, data!



Rank	Country	Gold	Silver	Bronze	Total
1	USA	46	37	38	121
2	Great Britain	27	23	17	67
3	China	26	18	26	70
4	Russia	19	18	19	56
5	Germany	17	10	15	42
6	Japan	12	8	21	41
7	France	10	18	14	42
8	South Korea	9	3	9	21
9	Italy	8	12	8	28
10	Australia	8	11	10	29
11	Netherlands	8	7	4	19
12	Hungary	8	3	4	15
13	Brazil	7	6	6	19
14	Spain	7	4	6	17
15	Kenya	6	6	1	13
16	Jamaica	6	3	2	11
17	Croatia	5	3	2	10

What now?!

This lecture is about …

Why and how to plan a good data analysis pipeline

Learning Objectives

After this lecture you should be able to . . .

- Describe different steps of the data analysis pipeline
- Identify potential pitfalls in each step and good practices how to avoid them
- Generate synthetic data to test a data analysis pipeline
- Understand the importance of data, privacy, patient protection and know some relevant frameworks
- Understand the concept of open science, its benefits and challenges, relevant platforms, licensing issues

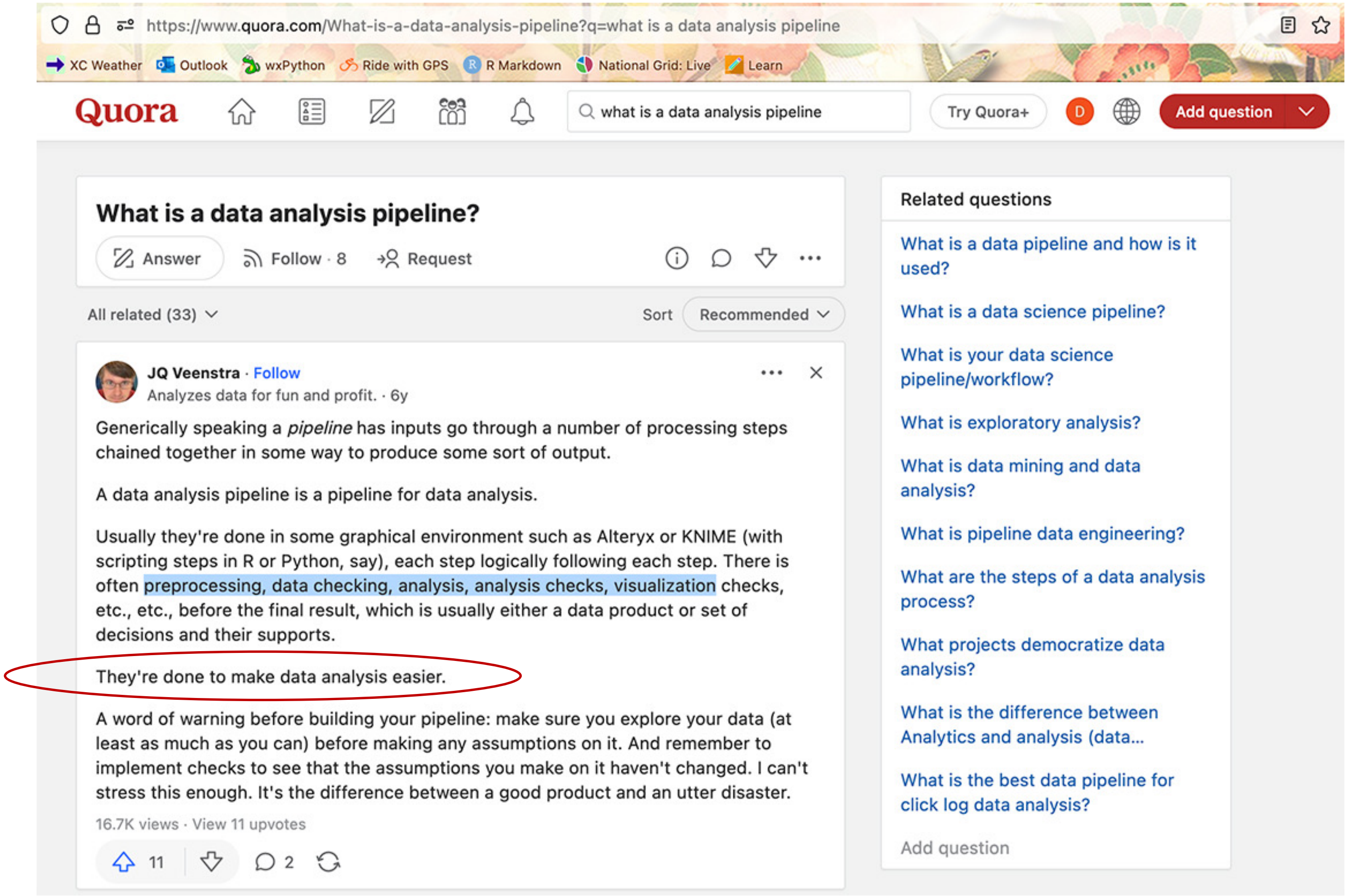
Outline

1. What is a data analysis pipeline and why do we need one?
2. Everything is OSEMN
3. Testing a data analysis pipeline
4. Open science

What is a data analysis pipeline?

Absolutely essential in data driven research

Also important in hypothesis driven research, where there are multiple and/or complicated datasets



What is a data analysis pipeline?

JQ Veenstra · Follow
Analyzes data for fun and profit. · 6y

Generically speaking a *pipeline* has inputs go through a number of processing steps chained together in some way to produce some sort of output.

A data analysis pipeline is a pipeline for data analysis.

Usually they're done in some graphical environment such as Alteryx or KNIME (with scripting steps in R or Python, say), each step logically following each step. There is often preprocessing, data checking, analysis, analysis checks, visualization checks, etc., etc., before the final result, which is usually either a data product or set of decisions and their supports.

They're done to make data analysis easier.

A word of warning before building your pipeline: make sure you explore your data (at least as much as you can) before making any assumptions on it. And remember to implement checks to see that the assumptions you make on it haven't changed. I can't stress this enough. It's the difference between a good product and an utter disaster.

Related questions

- What is a data pipeline and how is it used?
- What is a data science pipeline?
- What is your data science pipeline/workflow?
- What is exploratory analysis?
- What is data mining and data analysis?
- What is pipeline data engineering?
- What are the steps of a data analysis process?
- What projects democratize data analysis?
- What is the difference between Analytics and analysis (data...
- What is the best data pipeline for click log data analysis?

Add question

Pipelines are OSEMN

Data science pipeline – OESMN

You're awesome. I'm awesome. Data science is OESMN.

- O – **O**btaining our data
- S – **S**crubbing / Cleaning our data
- E – **E**xploring / Visualising our data will allow us to find patterns and trends
- M – **M**odelling our data will give us predictive power as a wizard
- N – i**N**terpreting our data

https://towardsdatascience.com/a-beginners-guide-to-the-data-science-pipeline-a4904b2d8ad3

Why is it important to have a well designed data analysis pipeline?

Why is it important to have a well designed data analysis pipeline?

- Systematic approach to data handling and analysis
- Documented and reproducible (for yourself and others)
- Allows for mistakes to be spotted and corrected
- . . .

Outline

1. What is a data analysis pipeline and why do we need one?
2. Everything is OSEMN
3. Testing a data analysis pipeline
4. Open science

Obtaining data

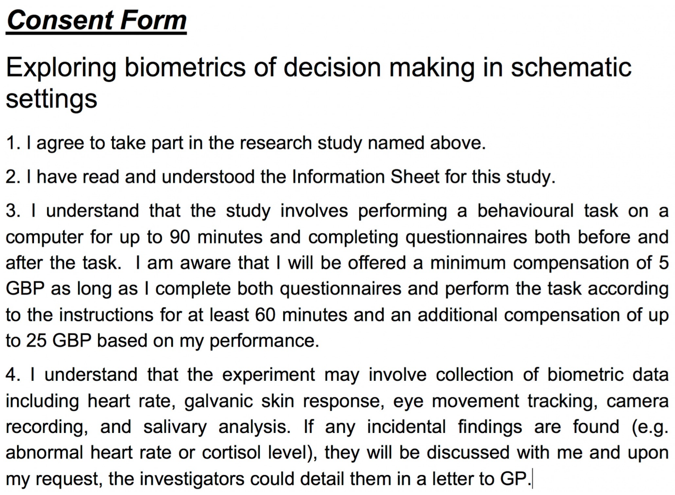
Obtaining data

Data science pipeline: obtaining data

Can data be stored in a way that is

- Efficient
- Objective
- With minimal data loss
- Secure
- With informed consent
- Protecting privacy
- Easily processable

Example



Consent Form

Exploring biometrics of decision making in schematic settings

1. I agree to take part in the research study named above.

2. I have read and understood the Information Sheet for this study.

3. I understand that the study involves performing a behavioural task on a computer for up to 90 minutes and completing questionnaires both before and after the task. I am aware that I will be offered a minimum compensation of 5 GBP as long as I complete both questionnaires and perform the task according to the instructions for at least 60 minutes and an additional compensation of up to 25 GBP based on my performance.

4. I understand that the experiment may involve collection of biometric data including heart rate, galvanic skin response, eye movement tracking, camera recording, and salivary analysis. If any incidental findings are found (e.g. abnormal heart rate or cortisol level), they will be discussed with me and upon my request, the investigators could detail them in a letter to GP.

Obtaining data

Privacy and data protection

- Research participants, customers, etc. should be able to understand how their data stored and used
- Anonymisation is not always possible, but de-identification is
- Think about who has access to data, how long it is stored, what happens to it afterwards
- Be aware of government data regulations, such as the EU General Data Protection Regulation (GDPR)
- Research involving human participants or data needs ethics approval (institution ethics boards, may also need special licences, e.g. when working with patient data)
- Laws set minimum standards - think about the ethical implications of your work beyond that!

Data management

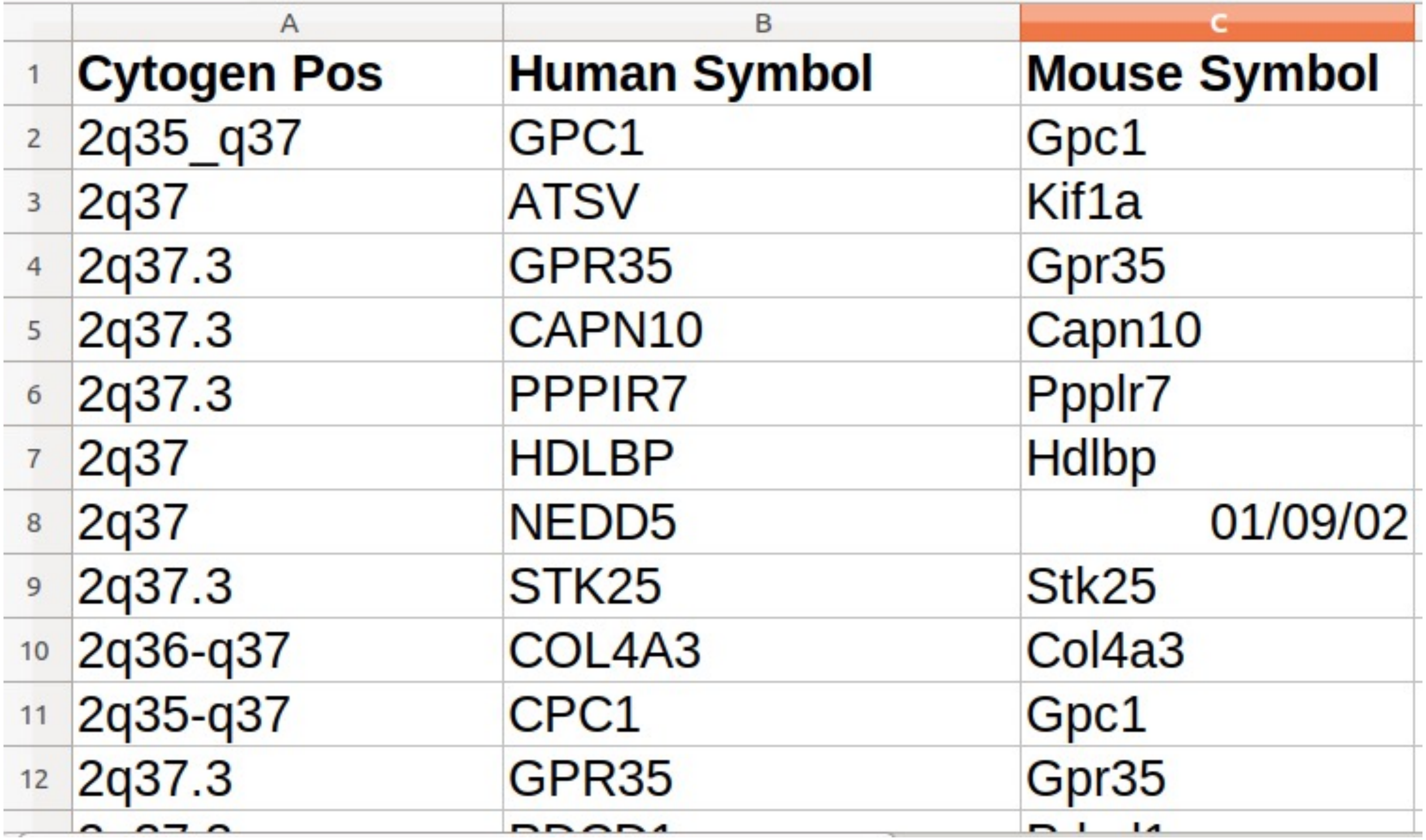
Data management plans

- Increasingly required in grant and research ethics applications
- Needs a description of data analysis pipeline
- Particular focus on data security, data sharing, privacy protection, ethics procedures
- Documentation and proper data analysis also important
- You can create a DMP by yourself:
  https://dmponline.dcc.ac.uk/

Scrubbing/Cleaning Data

Scrubbing/Cleaning Data

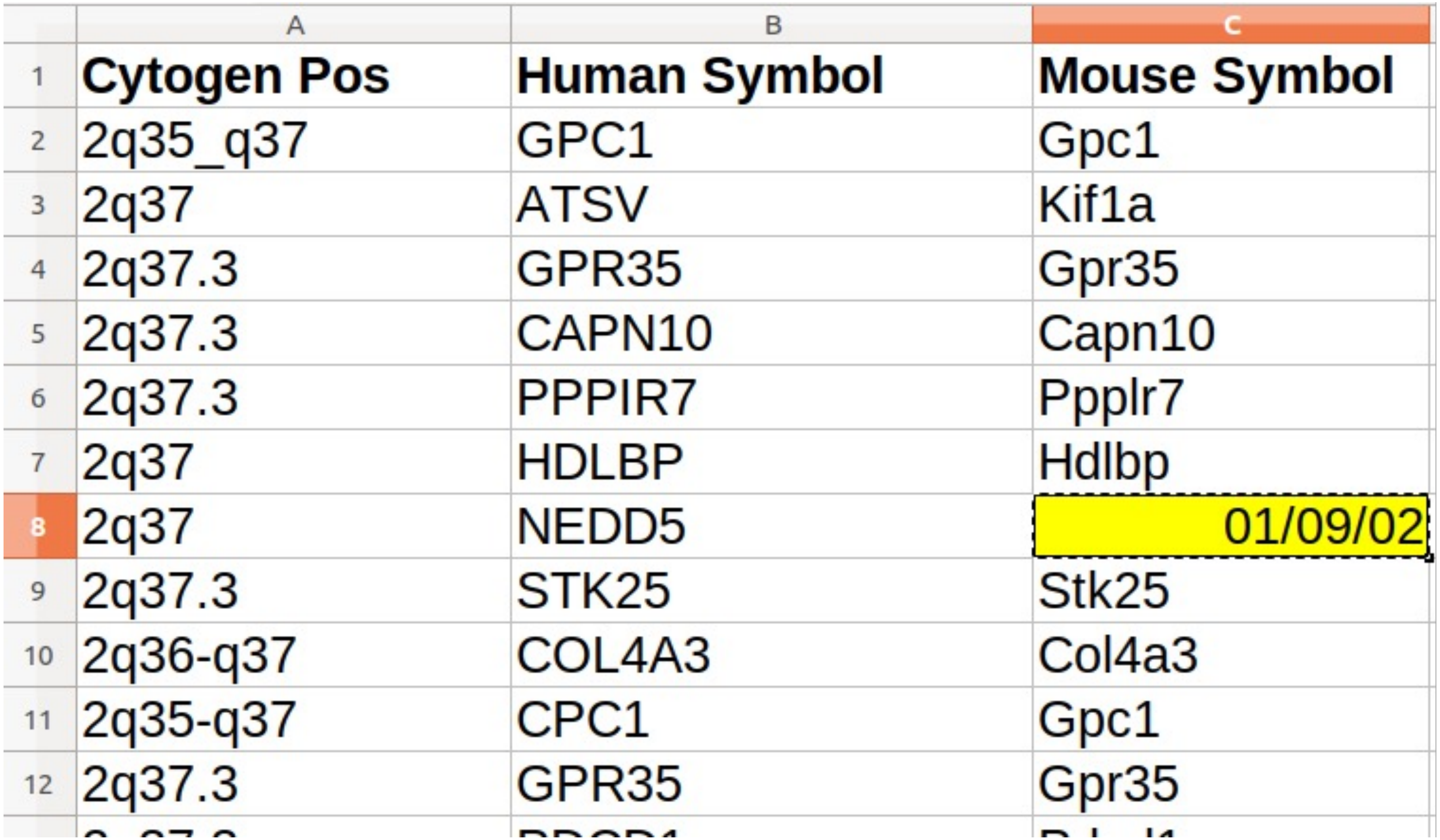
What’s wrong here?



	A	B	C
1	**Cytogen Pos**	**Human Symbol**	**Mouse Symbol**
2	2q35_q37	GPC1	Gpc1
3	2q37	ATSV	Kif1a
4	2q37.3	GPR35	Gpr35
5	2q37.3	CAPN10	Capn10
6	2q37.3	PPPIR7	Ppplr7
7	2q37	HDLBP	Hdlbp
8	2q37	NEDD5	01/09/02
9	2q37.3	STK25	Stk25
10	2q36-q37	COL4A3	Col4a3
11	2q35-q37	CPC1	Gpc1
12	2q37.3	GPR35	Gpr35

Scrubbing/Cleaning Data

What’s wrong here?



	A	B	C
1	**Cytogen Pos**	**Human Symbol**	**Mouse Symbol**
2	2q35_q37	GPC1	Gpc1
3	2q37	ATSV	Kif1a
4	2q37.3	GPR35	Gpr35
5	2q37.3	CAPN10	Capn10
6	2q37.3	PPPIR7	Ppplr7
7	2q37	HDLBP	Hdlbp
8	2q37	NEDD5	01/09/02
9	2q37.3	STK25	Stk25
10	2q36-q37	COL4A3	Col4a3
11	2q35-q37	CPC1	Gpc1
12	2q37.3	GPR35	Gpr35

But of course you remember “Getting and Cleaning Data” earlier this semester!

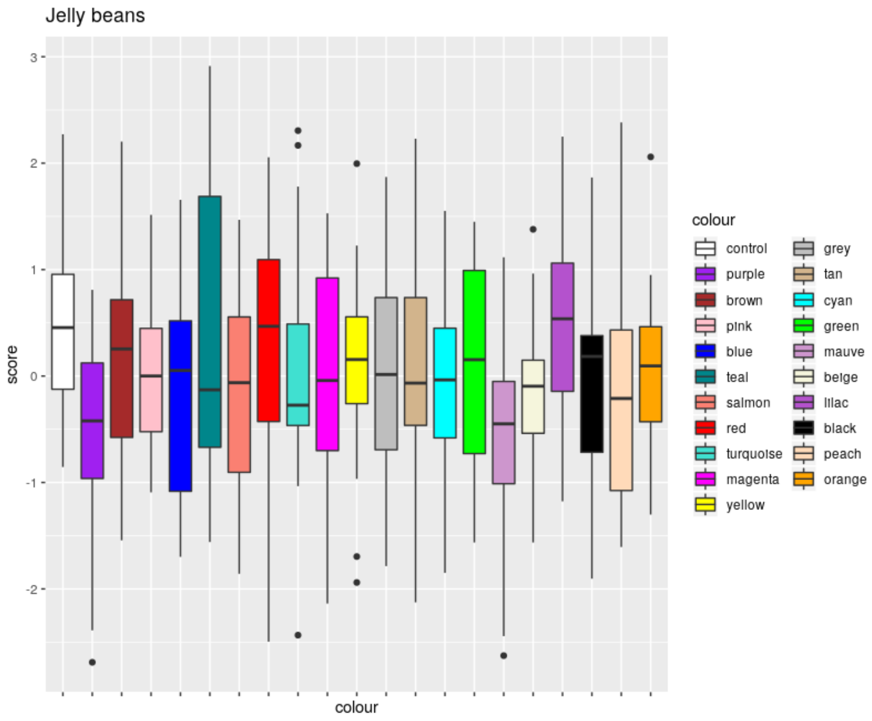
Exploring Data

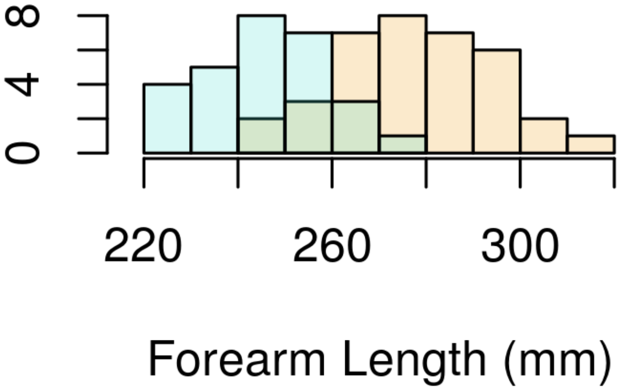
What does data exploration entail?

Exploring Data

What does data exploration entail?

Summary statistics, plots, “looking at the data”





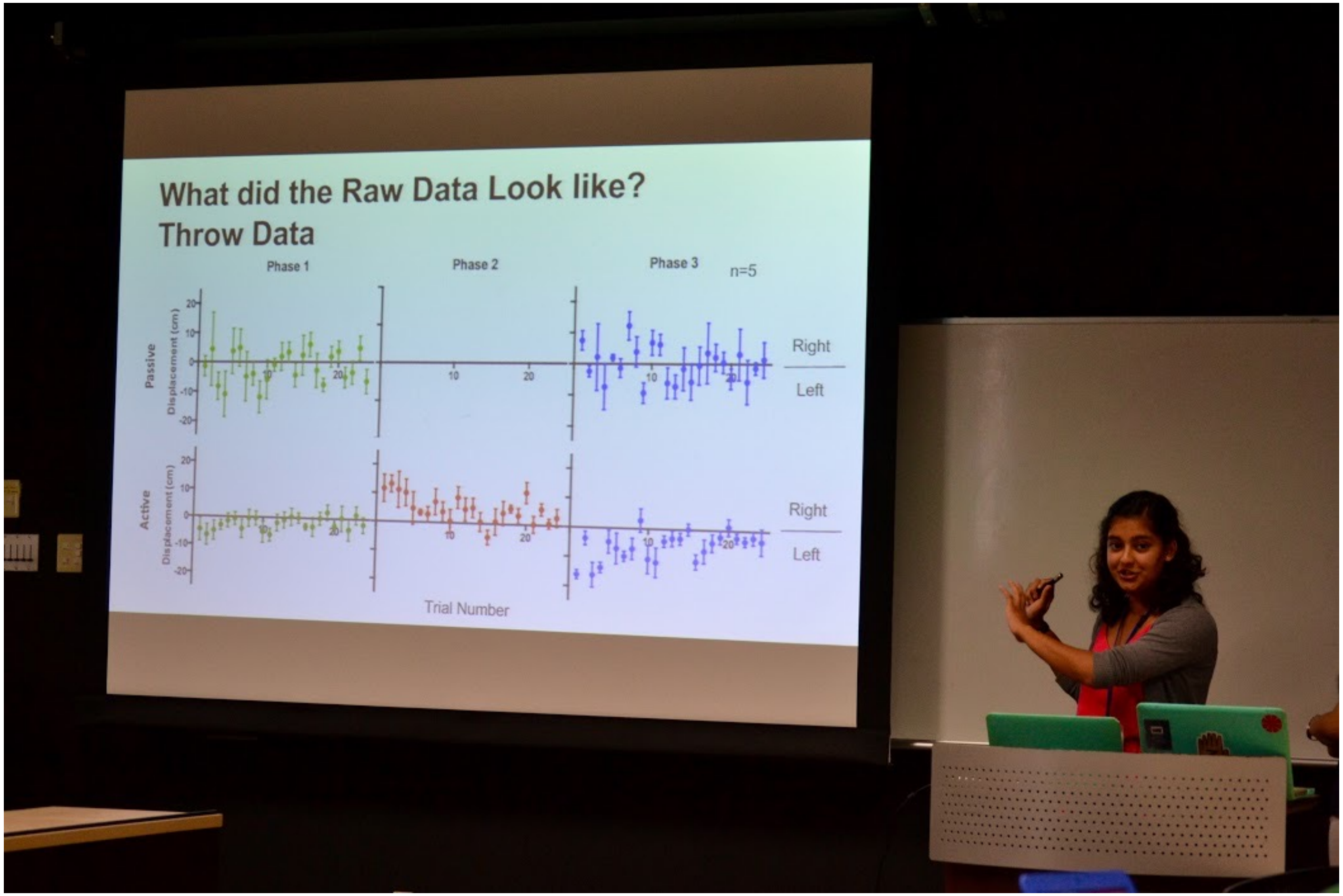
Modelling and i**N**terpreting Data

Learning stuff from data

Modelling and iNterpreting Data

Learning stuff from data

Hypothesis tests, regression, machine learning, model fitting, . . .



What did the Raw Data Look like?

Throw Data

Outline

1. What is a data analysis pipeline and why do we need one?

2. Everything is OSEMN

3. Testing a data analysis pipeline

4. Open science

How would you test a data analysis pipeline?

How would you test a data analysis pipeline?

Synthetic data: Dataset(s) that you create to test the pipeline

Can help address the following questions

- Is my pipeline working as expected?
- Do my anonymisation/de-identification protocols work ok?
- Can I detect a difference when there is one (and none when there is none?)
- How big a dataset would I need? What format do I need the data in?
- How does my pipeline deal with problems (outliers, incomplete data, typos etc.)
- How long does my pipeline take to run?

How would you test a data analysis pipeline?

Synthetic data: Dataset(s) that you create to test the pipeline

Can help address the following questions

- Is my pipeline working as expected?
- Do my anonymisation/de-identification protocols work ok?
- Can I detect a difference when there is one (and none when there is none?)
- How big a dataset would I need? What format do I need the data in?
- How does my pipeline deal with problems (outliers, incomplete data, typos etc.)
- How long does my pipeline take to run?

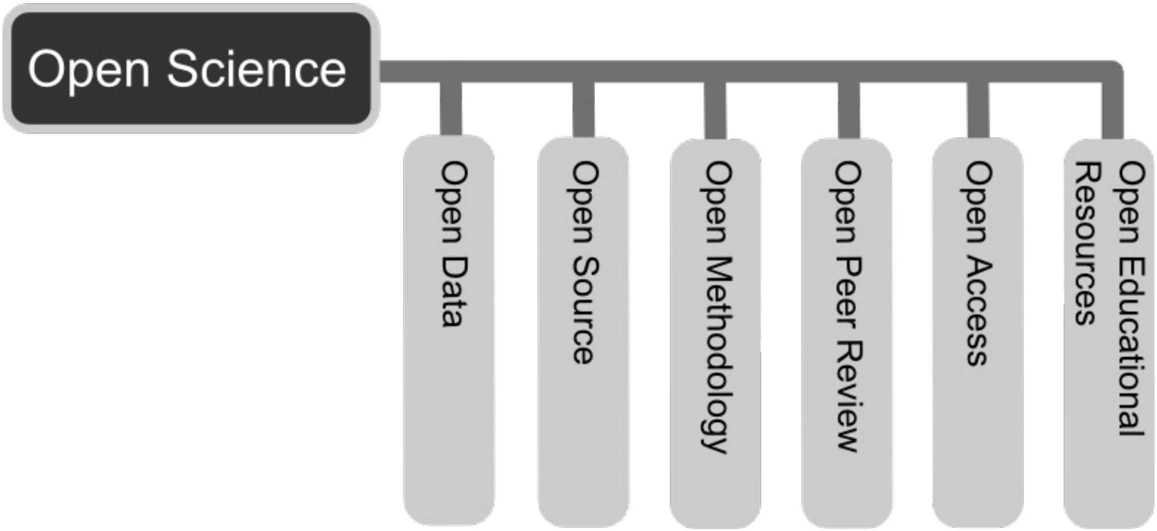
The synthetic dataset can be created (maybe based on pilot data) before data collection is complete!

Outline

1. What is a data analysis pipeline and why do we need one?
2. Everything is OSEMN
3. Testing a data analysis pipeline
4. Open science

Open science

https://szpylbert.com/tag/ivory-tower/



Open Science

Open Data

Open Source

Open Methodology

Open Peer Review

Open Access

Open Educational Resources

https://en.wikipedia.org/wiki/Open_science

- Increasingly popular and generally very welcome. But...
- More easily said than done: various issues to be addressed

Open science

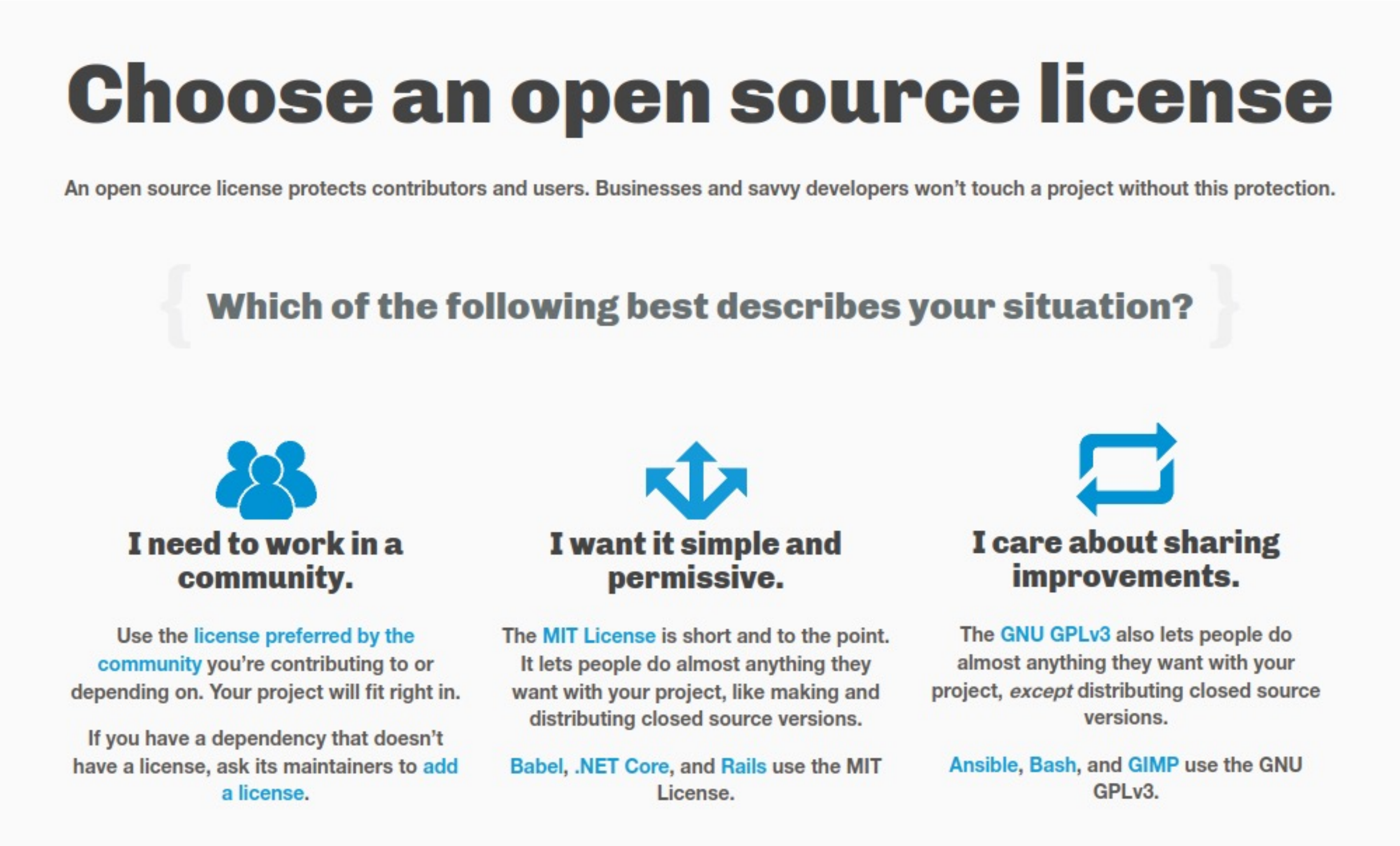
What are advantages of open science?

Can you think of times when open science conflicts with other considerations?

Licensing

- If you share your software (e.g. on GitHub), you need a license. Without a license, your code is automatically copyrighted, and nobody else can use it.
- If your code makes use of other people's code, that may dictate what kind of license you can and can't use.
- Choosing any license is better than having none. Use tools like https://choosealicense.com to help you decide.

Licensing



Choose an open source license

An open source license protects contributors and users. Businesses and savvy developers won't touch a project without this protection.

{ Which of the following best describes your situation? }

I need to work in a community.

Use the license preferred by the community you're contributing to or depending on. Your project will fit right in.

If you have a dependency that doesn't have a license, ask its maintainers to add a license.

I want it simple and permissive.

The MIT License is short and to the point. It lets people do almost anything they want with your project, like making and distributing closed source versions.

Babel, .NET Core, and Rails use the MIT License.

I care about sharing improvements.

The GNU GPLv3 also lets people do almost anything they want with your project, *except* distributing closed source versions.

Ansible, Bash, and GIMP use the GNU GPLv3.

Review

Now, you should be able to do the following:

- Describe different steps of the data analysis pipeline
- Identify potential pitfalls in each step, and good practices to avoid them
- Generate synthetic data to test a data analysis pipeline
- Understand the importance of data, privacy, patient protection, and know some relevant frameworks
- Understand the concept of open science, its benefits and challenges, relevant platforms, licensing issues

Acknowledgements and Image credits

This lecture uses materials from ADS2 lectures from previous years by Gedi Luksys and Melanie Steffan. Where not otherwise indicated, images are also from those lectures.

- Box plots. Melanie Steffan, CC BY-SA 3.0, 2019.
- Clipboard. By Schzmo, Public domain, via Wikimedia Commons
- Histogram. Melanie Steffan, CC BY-SA 3.0, 2017.
- Screenshot of a list of gene symbols in mice and humans. My own work, based on a figure from BR Zeeberg, J Riss, DW Kane, KJ Bussey, E Uchio, WM Linehan, JC Barrett & JN Weinstein (2004). Mistaken Identifiers: Gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics 5,80.
- Spreadsheet. By Softwaregenius, CC BY-SA 4.0, https://creativecommons.org/licenses/by-sa/4.0, via Wikimedia Commons
- Student presenting data. By Yuuki Guzman and Agoston Tyll (Okinawa Institute of Science and Technology), 2015.
- Three students working on computers. By Yuuki Guzman and Agoston Tyll (Okinawa Institute of Science and Technology), 2015.
- Tool for choosing an open-source license. Screenshot from `https://choosealicense.com`, 2021.