Data in the Wild - project hand-in

IT University – Fall 2024

Pexels – Austin Tsai

IT UNIVERSITY OF CPH

This course - better reflection of datasets

- Broader view, how did this gata get here?

- Create a dataset
  - Find data
  - Label it
  - Process it
  - Present it
  - Share it

- Reflect on this process, problems it **solves** or **creates**

After the course, you should be able to…

- Compare different data collection/annotation/visualization methods according to their strengths and weaknesses
- Apply appropriate data collection/annotation/visualization methods in order to create novel datasets
- Find suitable connections between dataset properties, analysis methods, and research questions
- Extract insights from the data analysis and present the results with appropriate visualization and written reporting
- Discuss the findings with respect to relevant work from the literature, as well as their real-world implications

Project outline

- **Define a topic that's interesting and helps you show that you achieved the learning objectives**
- Create a dataset and show an example analysis - other people could use the dataset for different analyses
- Submit report + Github repository
- D1G exam (report submission + oral, whole group present)

Project topics

Project topic

- In previous years, you could choose almost any topic
  - Lots of interesting topics!
  - But not always feasible to create a good project with…
- The theme for this year: academic research
- Bonus: your topic might already provide you with some skills and/or content for your other courses (e.g. your 7.5 research project)

Project ideas

- The samples for your dataset can for example be:
  - research papers
  - reviews of research papers (openreview.org)
  - datasets
  - patents
  - Github repositories
  - scientists
  - conferences or journals
  - courses given at universities
  - …

Project ideas

Here are some similar studies, for URLs see LearnIT

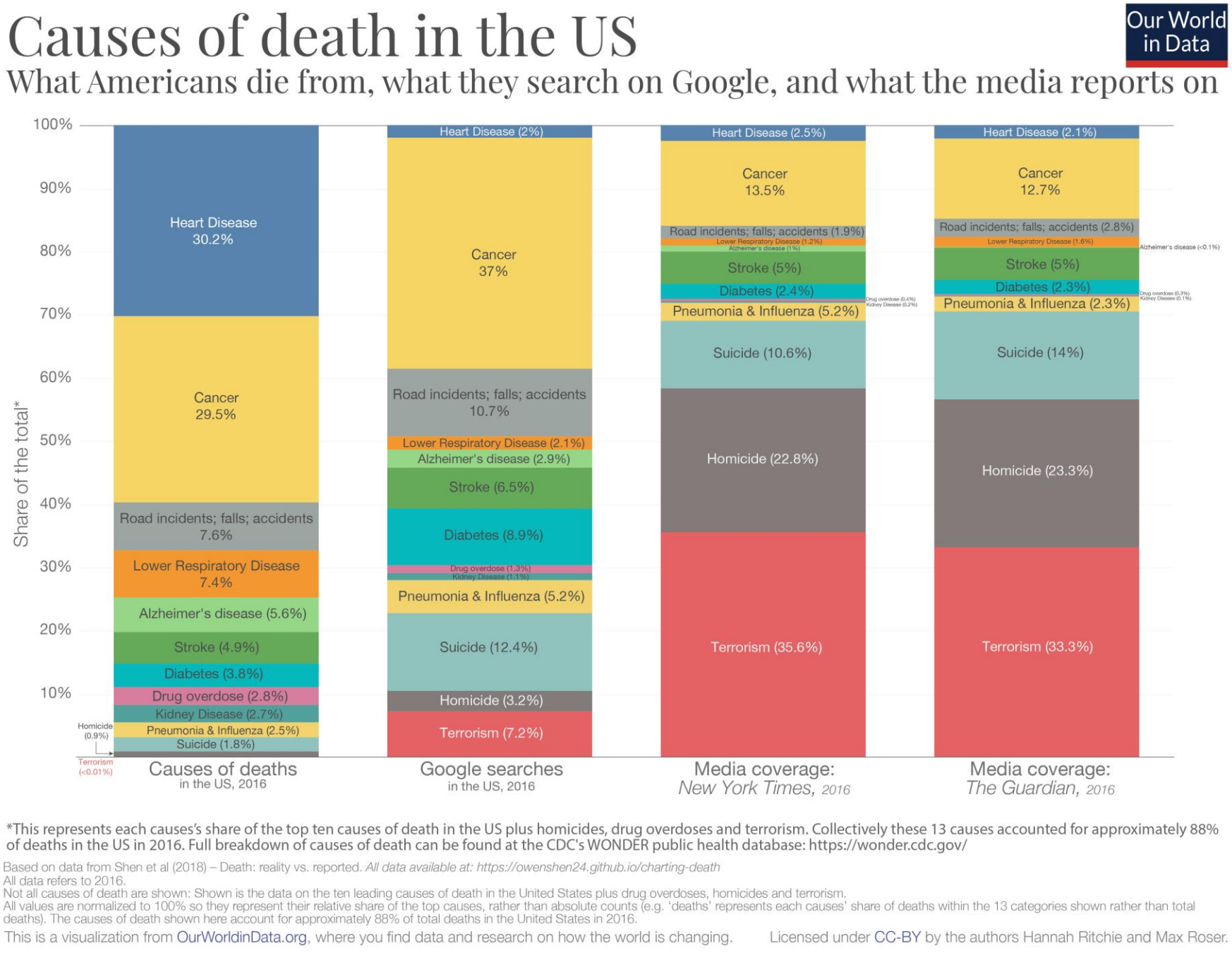
Birhane et al - Do papers at top ML conferences discuss issues like ethical concerns?

Sourget et al – Which medical images are more often used for benchmark experiments?

Bonham et al – Is there a bias in the gender of paper authors in computational biology ?

Project ideas

- Topics, vs their representation in the media
- Source
- Research papers on these diseases?



Project ideas

- Are medical papers with funny titles cited more often?
- How complete are README papers of popular Github repositories?
- How environmentally responsible are computer science conferences?

Project starting points

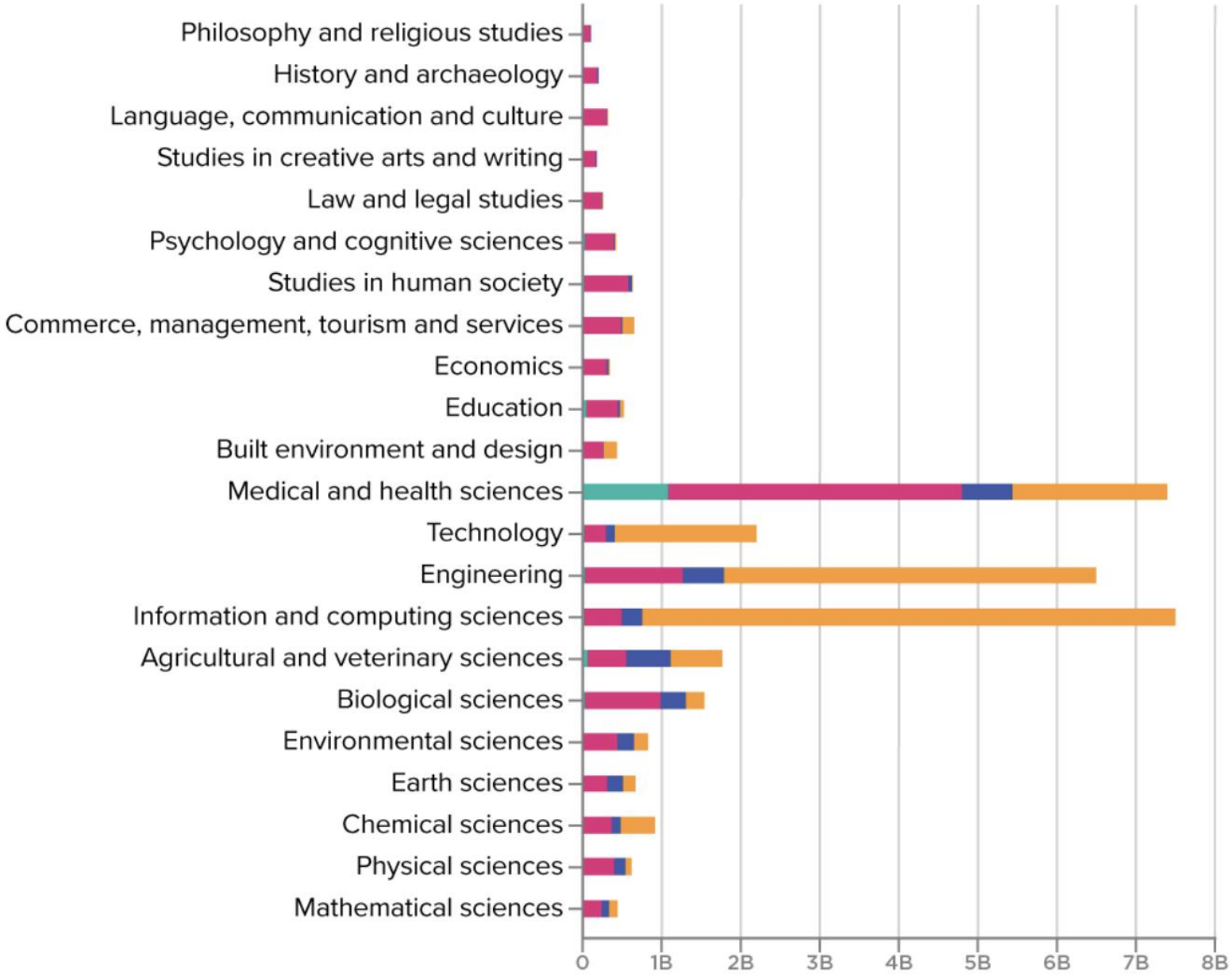
- You do not need to collect everything from scratch, can include parts of existing datasets

- Lists of top things - top cited papers, starred repositories etc.

Project ideas

- Cross-reference with d
  about the world

- OECD
- WHO
- OurWorldInData.org
- Local organizations



Project ideas

It is NOT mandatory to choose this kind of topic

If you have an amazing idea that you really want to do, but that is not about research, talk to us EARLY

Typically not good ideas: cryptocurrency, house prices, social media you are not allowed to scrape

Project tips – what this project is not

The datasets in these examples are NOT focused on prediction / machine learning

Scraping lots of data from many sources but without any question or analysis ≠ achieving the learning objectives

Project tips

- You will likely need to limit scope, e.g. “papers from conferences X, Y, Z”
- This can include field/topic of interest (finance, history, sports, ...)
- Start with a “minimum viable solution”, then extend during the course if time allows

[PDF] Effects of different ingredients on texture of **ice cream**

QA Syed, S Anwar, R Shukat… - Journal of Nutritional …, 2018 - researchgate.net

… Various ingredients are available in market for **ice cream** that have important effects on **ice cream** quality. These ingredients can be categorized in dairy and nondairy components from …

☆ Save 99 Cite Cited by 155 Related articles All 4 versions »

What this project is NOT about

- Primary personal data
- Data that the providers don’t want to share with you
- Lots of data from many sources but without any analysis or discussion
- ML with top accuracy / p-hacking

Project proposal

Project proposal

- Find a group of 5 people
- Ask one of the TAs to create a Teams channel for your group, titled Team-[YourTeamNameHere]
- Make a post in this channel (plain text, PDF if you prefer)
- We are **[Team Name]** consisting of @[user1] @[user2] @[user3] @[user4] @[user5]. Our project focuses on [Project Description]
- Do this before the deadline (in 4-5 weeks, see LearnIT)

Project proposal

Your proposal should describe:

- What is the motivation / background? What is it that we do not know yet, that your dataset might help answer?
- What kind of data sources might you use? You should aim to integrate multiple sources and/or multiple types of data (for example images and text)
- Potential challenges (both technical and societal/ethical)

Project proposal

- This is not a graded assignment, but to make sure you get a good start with the project

- If you complete the proposal before the deadline you will get feedback and suggestions

Project outline

- Apply 1+ techniques from each topic (collection, annotation, …) of the course

- Use feedback from project proposal/Teams and group session – you will get time slots to discuss your project closer to submission date

Data

- Multiple sources of data & ways to retrieve it (APIs, scraping..) – ideally also multiple modalities (images/text/tabular etc)

- Your data will likely be:
  - Secondary data from other data sources
  - Primary data in the form of new ways of using secondary data (your annotations)

- Discuss possible biases, ethical implications

Annotations

Best case:

- Create an annotation guide

- Ask somebody else to label a smaller part of data
  - Use LabelStudio, Taguette or other software

- Label more of the data automatically (pretrained model, etc)

- Analyse the annotators sources / accuracy / reliability …

- You can reflect on your annotation guide, ambiguities, biases etc…

Processing

This is highly dependent on the data you choose, but generally you should try to

- Apply (multiple) techniques from the course – having more types of data helps to do this!

- Motivate your choices and reflect on them

Analysis

This is highly dependent on the data you choose, but generally you should try to

- Connect your analysis to the motivation/research questions you started with
- If you cannot answer them (fully), discuss why not, and discuss potential further studies
- Discuss your findings with respect to other literature on the topic

Reporting

- Visualization
  - Use appropriate types to explain your analysis
  - This is an often overlooked part in submitted projects
- Report
  - Appropriate level of academic writing, structure, readability (see more slides below)
- Data and code
  - Proper documentation, structure – understandable to others (see more slides below)

Reporting

- Use appropriate types to explain your analysis
- This is an often overlooked part of projects

Project submission + exam overall

Project assessment

- D1G exam (report submission + oral)
- Course manager + external examiner read reports and do the oral exam
- TAs will go through your data/code
- The learning objectives are what we look at to assess your projects…

Project assessment

- What techniques (collection/annotation/analysis) does this project use / are they appropriately motivated and applied?
- What are the research questions / are the connections between questions, data, insights from analysis suitable?
- How appropriate is the presentation (visualizations/written reporting) - this includes documenting your data/code!
- How appropriate is the discussion with respect to other literature/ethical implications etc?

Project assessment

- This is not a checkbox ticking exercise
- Limitations in some parts of the project can be compensated by other parts
- Overall guidelines for assessing projects (also 7.5 ECTS research project, thesis)

Grade	Description
12	Excellent. High level of command of all aspects – no or only a few minor weaknesses
10	High level of command of most aspects – only minor weaknesses
7	Good. Good command – some weaknesses
4	Fair. Some command – some major weaknesses
02	Adequate. The minimum requirements for acceptance
00	Inadequate. Does not meet the minimum requirements for acceptance
-3	Unacceptable. Unacceptable in all respects.

Project hand-in

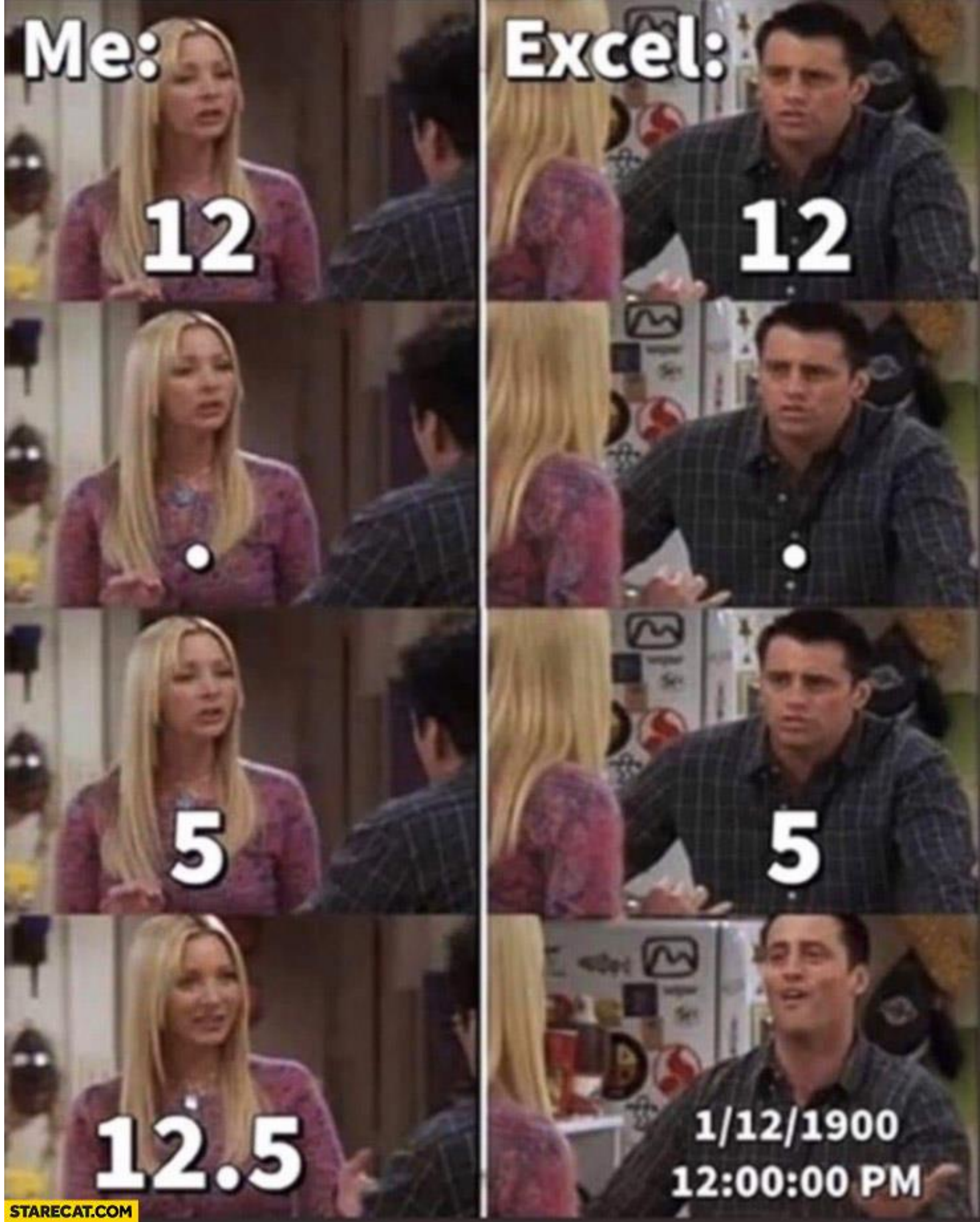
1) Data and code

Project hand-in

- Code on Github
  - Include everything needed to process/analyse your data. The code should reproduce the tables/figures in your report
  - A stand-alone README.MD with description of the data, what all the files do - should be usable for another KDS or CS student
- Data
  - Ideally, original/raw data
    - If too large, you can host the files externally
    - OR provide an example how the data looks like and how it is processed
  - Processed data

Data & code

Organized and appropriate formats for the data




So what's your idea of a perfect date?
IG: @PunHubOnline
YYYY-MM-DD
I find other formats a bit confusing.
Pun hub

Data & code

Data organization in spreadsheets [Paper]

- CSV format, no features from Excel. One table per CSV file
- Clear column names (+ separate key)
- Consistent formatting (numbers, dates)
- No summary rows
- What is empty?

(Your data can be in other formats as well, e.g. JSON, but these are good rules to follow if you have CSV)

Data & code

File names

report.docx
Vs
20211108-DataScience-group01-report.pdf

Aim: readable for humans and machines, intuitive ordering

More examples in:
https://speakerdeck.com/jennybc/how-to-name-files

NO

myabstract.docx
Joe's Filenames Use Spaces and Punctuation.xlsx
figure 1.png
fig 2.png
JW7d^(2sl@deletethisandyourcareerisoverWx2*.txt

YES

2014-06-08_abstract-for-sla.docx
joes-filenames-are-getting-better.xlsx
fig01_scatterplot-talk-length-vs-interest.png
fig02_histogram-talk-attendance.png
1986-01-28_raw-data-from-challenger-o-rings.txt

Data & code

Consistent & descriptive names

Comments

Look at repositories that you use!

90% of all code comments:

Data & code

Keep parameters & logic separated (e.g. thresholds)

define parameter file which can be linked to experiments

Keep track of the random seed

Data & code

What NOT to do (funny): https://github.com/Droogans/unmaintainable-code

Be Abstract

In naming functions and variables, make heavy use of abstract words like *it*, *everything*, *data*, *handle*, *stuff*, *do*, *routine*, *perform* and the digits e.g. `routineX48`, `PerformDataFunction`, `DoIt`, `HandleStuff` and `do_args_method`.

Misleading names

Make sure that every method does a little bit more (or less) than its name suggests. As a simple example, a method named `isValid(x)` should as a side effect convert `x` to binary and store the result in a database.

Data & code - Reproducibility

Ideal situation:

- Data and code shared
- Tables / graphs can be exactly reproduced with a few clicks
- Data is in a common format, we can use other code (robust)
- Code can be run on other data that's not in the same format (replicable)

		Data	
		Same	Different
Analysis	Same	Reproducible	Replicable
	Different	Robust	Generalisable

https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions

Project hand-in

2) Report

Report

2) Project report should describe
   - Introduction (why this dataset?)
   - Methods (what did you use, and why?)
   - Experiments/Analysis (what insights could you get from this dataset?)
   - Discussion

The sections do not need to have these exact titles, but you should cover these elements

“Data and code are available on …” on the first page (e.g. last sentence of abstract, or a footnote)

Report

- This is not a writing course BUT written reporting is part of the learning objectives

- Recommended: +/- 10 pages in a double column format
  - Overleaf has templates, for example IEEE or AAAI or ACM conference proceedings

  - You can add additional analyses to the appendix (but the main report must be readable stand-alone)

Report

- Further slides in this presentation are general slides, not specifically for Data in the Wild course

- But could help you in writing a concise & clear report

Report - What to include

(General slides, not specifically for Data in the Wild course)

Tip #1: Learn from examples

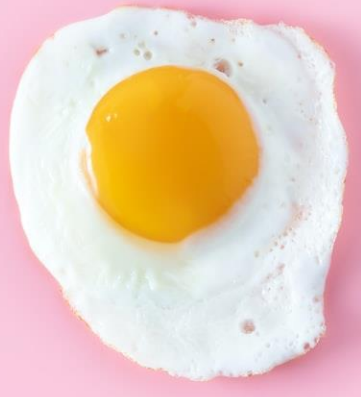
- Look at other papers
- “Three minute thesis” on YouTube
- Ask what is needed, there could be exceptions

General structure

- Title/abstract
- Introduction
- Related work
- Methods
- Experiments
- Results
- Discussion
- Conclusion

- (Depending on the project, you will not need to write all these)

[Image Leti Kugler]



Introduction

- **What is the problem?**

Eating raw eggs can be dangerous…

- **How are others solving the problem?** Why is it not enough?

Humans have been cooking eggs for centuries using different methods, including boiling [1] and frying [2,3]. However, it is unclear what the optimal time is for cooking eggs of different sizes…

- **How do we propose to solve it?**

We investigate different cooking strategies and measure the temperature of the cooked egg..

Related work

- Can be included in the middle of the introduction -OR-
- A short summary in introduction + own section (often Section 1.1, or Section 2) with details

Related work

Overall strategy:

1) Start general

Egg cooking is typically addressed by boiling [1-3], frying [4,5] or poaching [6,7]

2) More specific paragraphs for different topics

The traditional method is boiling. In [1,2], the authors fill a pan with cold water and ….. On the other hand, [3] boils the water with a kettle, and ….

3) Distinguish what you will do

We investigate a frying method, looking more specifically different types of oil ..

Methods

- The general “recipe” of what you do
- Overview of the steps (often a flowchart) + recipe for each step

We boil each egg for M minutes

- Can use bullet point lists, or a box with pseudocode

Algorithm 1 Sample Algorithm

```
1: i ← 1
2: while i > 10 do
3:     statement
4:     if condition then
5:         statement
6:         statement
7:     else
8:         statement
9:     end if
10: end while
```

Experiments

- One or more examples of applying the recipe in a specific case

We apply the method to batches of 12 eggs from…
We experiment with M=2, M=4, M=8 …
We evaluate each setup by measuring the average temperature…

- Specific hardware/software used

Results

- Summarize what the experiments show

We show the average temperature as a function of the time M in Fig. 1. We see that the temperature is relatively constant …

- Note “In Fig.1” vs “also in this figure”

Discussion

- How do your results compare to what you expected?

We did not find differences in temperature in our experiments. This is contrast to [3,4] where higher values of M lead to …

- What could you have done differently?

One reason for our results could be a too small range of M. In future work we plan to investigate…

Discussion

- What are some general implications of this research?

An important issue to consider in egg boiling experiments is the carbon footprint of animal products…

Presentation

- For a 10 minute presentation on one project you do not need a slide with “presentation structure” (but you could include “breadcrumbs”)
- You do not need to cover everything in your report
- Focus on:
  - Motivation
  - Main ideas behind method
  - Results
  - Discussion
- You should use less text than these slides ☺ (different purpose of slides)

Report - Help the reader

(General slides, not specifically for Data in the Wild course)

Remove clutter

- Empty sentences
  - “As has been previously stated …”

- Watch examples from Coursera course (Unit 1) https://www.coursera.org/learn/sciwrite

Remove clutter

- Use simpler words
  - utilize → use
  - [Checker]

THE UP-GOER FIVE TEXT EDITOR

CAN YOU EXPLAIN A HARD IDEA USING ONLY THE TEN HUNDRED MOST USED W
VERY EASY. TYPE IN THE BOX TO TRY IT OUT.

Active voice

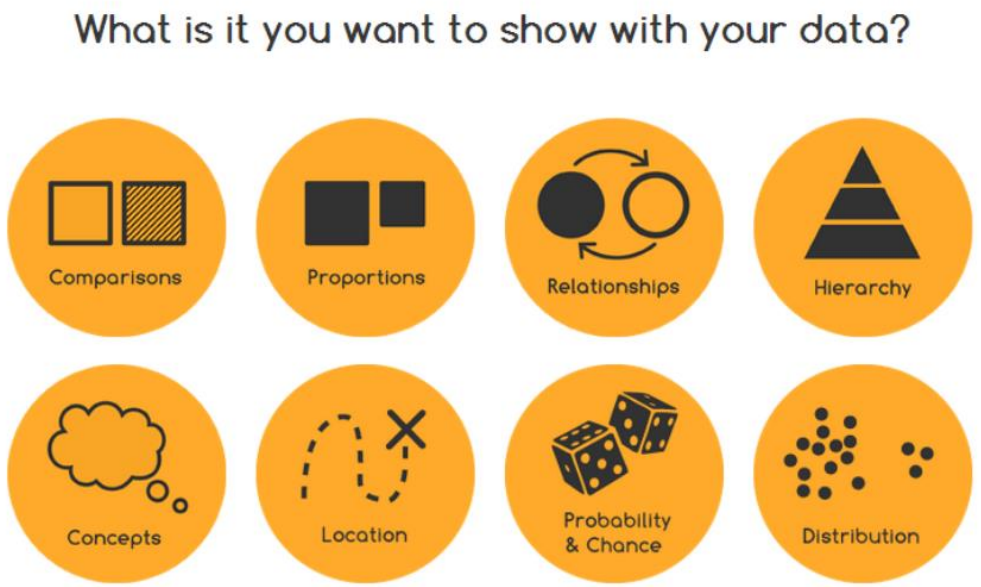
- “It is shown” → we show
- Noun + verb
- Watch examples from Coursera course (Unit 2) https://www.coursera.org/learn/sciwrite

Paragraphs

- One main idea per paragraph
- The first sentences of the paragraphs should make a coherent story
- Paragraphs should not be too long, use lists where appropriate (for example steps of the method)

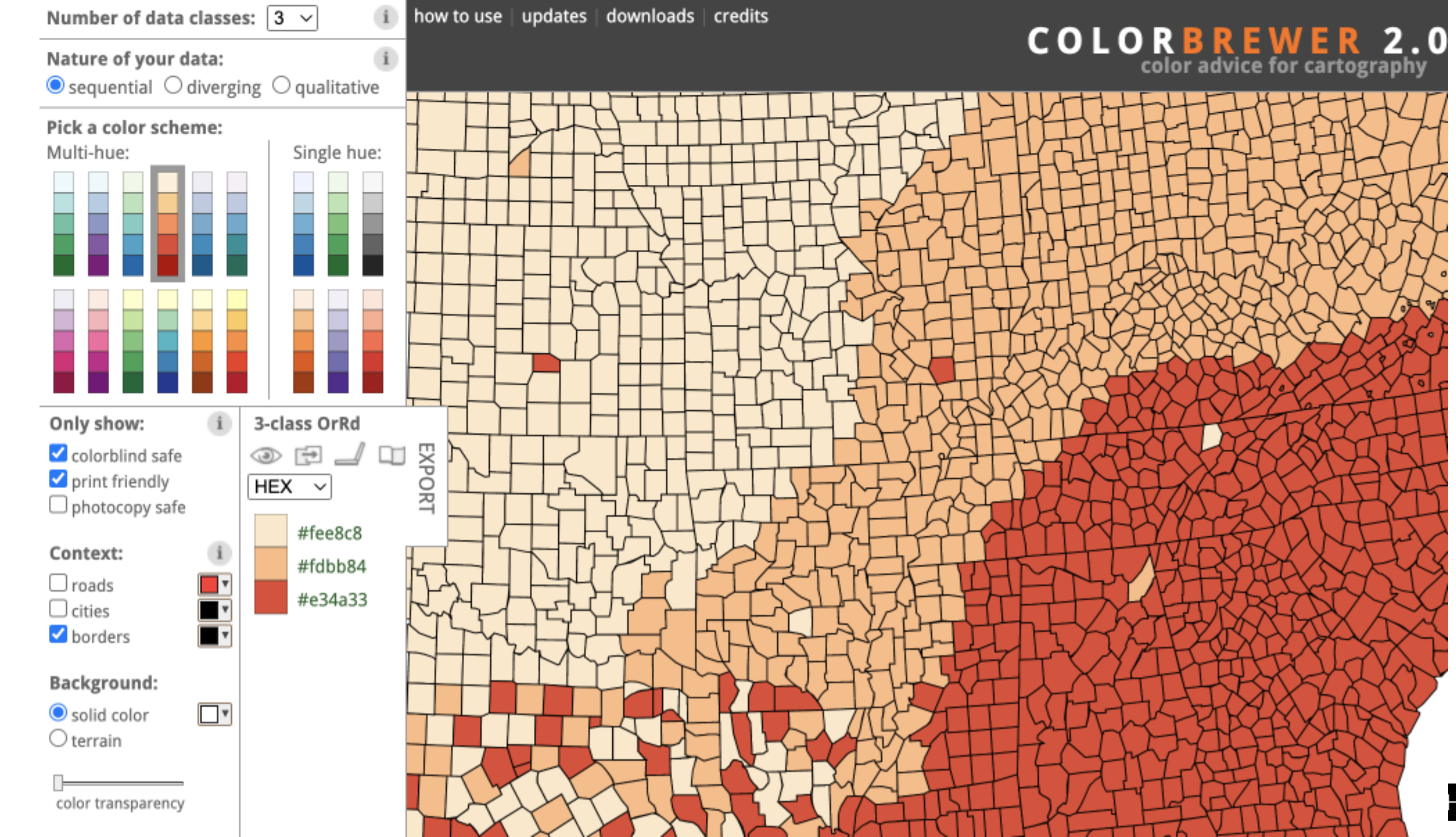
Figures

- What is the message? https://github.com/widged/data-for-good/wiki/Visualisation-::-Choosing-a-chart



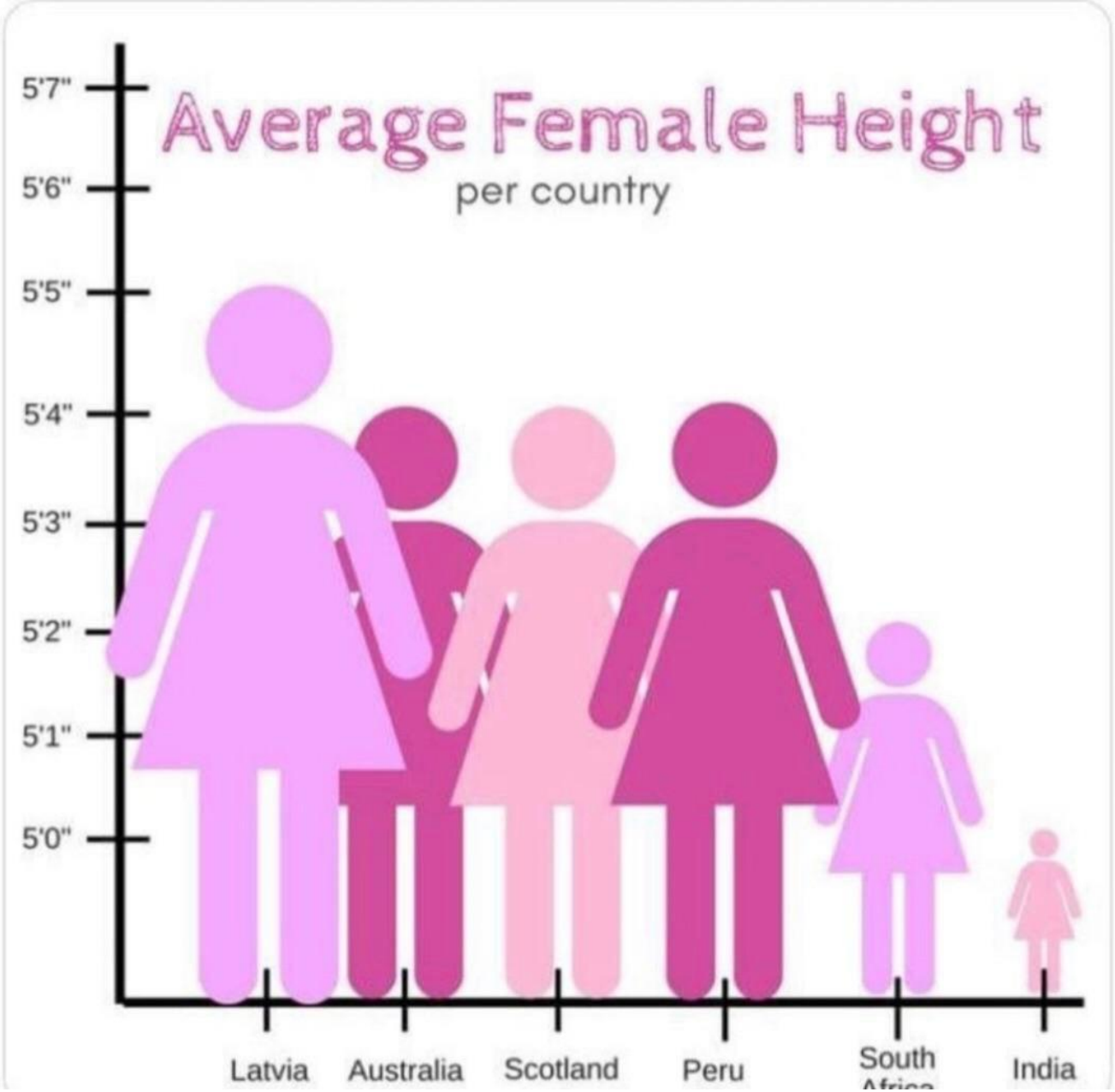
Figures

- Descriptive labels/captions, readable stand-alone
- Don’t forget axis labels and legend
- Ideally: check color scheme (color-blind, printer-friendly)


Number of data classes: 3
Nature of your data:
sequential
diverging
qualitative
Pick a color scheme:
Multi-hue:
Single hue:
Only show:
colorblind safe
print friendly
photocopy safe
Context:
roads
cities
borders
Background:
solid color
terrain
color transparency
3-class OrRd
HEX
#fee8c8
#fdbb84
#e34a33
EXPORT
how to use
updates
downloads
credits
COLORBREWER 2.0
color advice for cartography

Figures

- Pay attention to what scales represent



Figures

- Do not take screenshots of code and add it as a figure!
- If you want to add code, the best is to add *pseudocode*, *e*.g. \usepackage{algorithmic}

```
\documentclass{IEEETran}
\usepackage{algpseudocode}
\usepackage{algorithm}

\begin{document}
\begin{algorithm}
\caption{Sample Algorithm}
\begin{algorithmic}[1]
\State $ i \gets 1 $
\While {$ i > 10 $}
\State statement
\If {condition}
\State statement
\State statement
\Else
\State statement
\EndIf
\EndWhile
\end{algorithmic}
\end{algorithm}
\end{document}
```

Algorithm 1 Sample Algorithm

```
1: i ← 1
2: while i > 10 do
3:     statement
4:     if condition then
5:         statement
6:         statement
7:     else
8:         statement
9:     end if
10: end while
```

Tables

- What is the message?
- Is there something you want to highlight? (Best result per row/column, etc)
- Descriptive labels/captions, readable stand-alone
- Unit 5 of Coursera course

Figures & Tables

- Generate with your Python code!
  - run_experiment.py saves the results (e.g. .csv)
  - print_results.py creates the figures and tables

  - Figure: load csv, plot it, save image
  - Table: load csv, generate LateX (e.g. with pandas.DataFrame.to_latex), use \input{table.tex} in your report

- Avoid errors or typos due to copy pasting

- Easy to change formatting

References

- References - there are different styles for different journals etc
- This report: any style, as long as references are consistent & complete
  - Title
  - Author
  - Year
  - Venue (journal/conference with pages or URL for website)
- Typical bibtex entries: @article, @inproceedings, @misc