AGENDA

1) **A3** ORES & BIAS
2) **Feedback** from last week
3) New Assignment

---

1 | ORES & BIAS

Group 1: Q1, Q3 - **ANSA**, **MAOP**, **LUKA**
Group 2: Q2, Q4 - **XIYU, goto…, MASC**
Group 3: Q7, Q5, Q6 - **CHKI, JOWE, ALBE, SEKU**

---

1) Add your name to a group.
2) To what types of bias does **Q1** refer to?
   → ….. Preexisting biases
3) To what types of bias does **Q2** relate to?
   → ….. Group 2
4) Please add **sources of bias** and **types for** preexisting, *technical* and *emergent* bias to the table below. (Group 1 and 2)
5) Answer your questions and add your answers to the given table.

Types of Bias	**Sources of Bias**	**Types of …** ←	**Q1 -** What biases did you expect to find in the **data** (**before** you started working with it), and why?	**Q2 -** What (potential) sources of bias did you **discover** or **introduce during data processing and analysis?**

Preexisting bias	Activity Bias	**Population biases**	*baises against poorer / non-english countries and regions.* I think one could gather insights about the ralation of article quality and the level of education, access to internet or the effects of differnt regimes do have to article quality. Because countries with fewer population , fewer access internet and fewer education. Dataset is represented by a younger population	- There should be more high-quality articles from countries speaking english as they “compete” -
		Behavioral biases	harder to get good quality articles about countries with certain regimes (e.g. north korea) People in poorer countries don’t have time to write articles	- in some countries, people might simply not care about writing (good) articles.The data might be misleading - Same, other countries may not care writing articles about another country, so that data is underrepresented and appears of “low quality” just like countries with certain regimes - In some countries, there are similar platforms
		Temporal variations		
	Data Bias	Functional biases	the population data is consistent (Czechia (export_2019.csv) vs. Czech Republic (page_data.csv)).	
		Sampling bias	not clear for what the downloaded data was intentionally used for preexisting biases can not be excluded	**Data acquisition. The raw data downloaded from online source and the data collection details is not 100% clear. If the data was collected for other research purpose, it may create bias for out analysis.**

		...		
Technical bias	...	...	Missing technology leads to less data about a country / region	**Parameter choosing by ranking countries**
		Processing bias	Data has been inconsistent	
Emergent bias	...		More information (wikipedia articles) about a region, also lead to more new information about that region	

Q3 - What might your results suggest about (English) Wikipedia as a **data source**?	**Q4 -** What might your results suggest about the **internet and global society in general**?
• For the english wikipedia english speaking countries are more represented • Also first world countries were much more represented, despite not speaking english	• Honestly speaking, I don't think my results can suggest anything about the internet and global society, especially IN GENERAL. I don't think this project is valid for this purpose at all. • The anonymity of the internet makes it an extralegal place. To most parts, it seems to be impossible to find out whether things said and done online are intentionally said and done the way they were. Already in "real" global society this is not as easy, think of false-flag operations. This is so much easier online.

Q5 - Can you think of a realistic data science research situation where using these data (to train a model, perform a hypothesis-driven research, or make business decisions) might create biased or misleading results, due to the inherent gaps and limitations of the data?	**Q6 -** Can you think of a realistic data science research situation where using these data (to train a model, perform a hypothesis-driven research, or make business decisions) might still be appropriate and useful, despite its inherent limitations and biases?
• *A possible research situation could be trying to determine the quality of knowledge one countries inhabitants have about their politicians by looking at the article quality rate. Since only english Wikipedia was used, there's a possibility that a high amount of ones country's articles are written by westerners, thus not representing the countries population at all.* • *The data might not be suitable for assessing general article quality (of articles about politicians or even other articles) for any country. If ORES systematically rated articles differently depending on their relation to a certain country, it also could create problems.*	• *I think the data could be useful in many cases (especially if you are explicitly interested in the quality of English articles). For example, if the goal is to provide high quality articles for important German politicians in English, the data might help you to find articles which need to be improved. You could also use the data set to derive hypotheses (like: The amount of high quality articles in English is related to the ranking of the country in the PISA study). I think, the data can also be used if the research question is limited to countries with English as their national language.* • *Yes, by not comparing all countries but countries which are similar.*
Q7 - How might a researcher supplement or transform this dataset to potentially correct for the limitations/biases you observed?	
• Instead of getting predictions for English articles, one could retrieve predictions for articles written in the country's national language. Also, to get a complete picture one could try to get a complete list of countries. It is not trivial which politicians to choose from each country.	

2 | Feedback from last week

3 | New Assignment