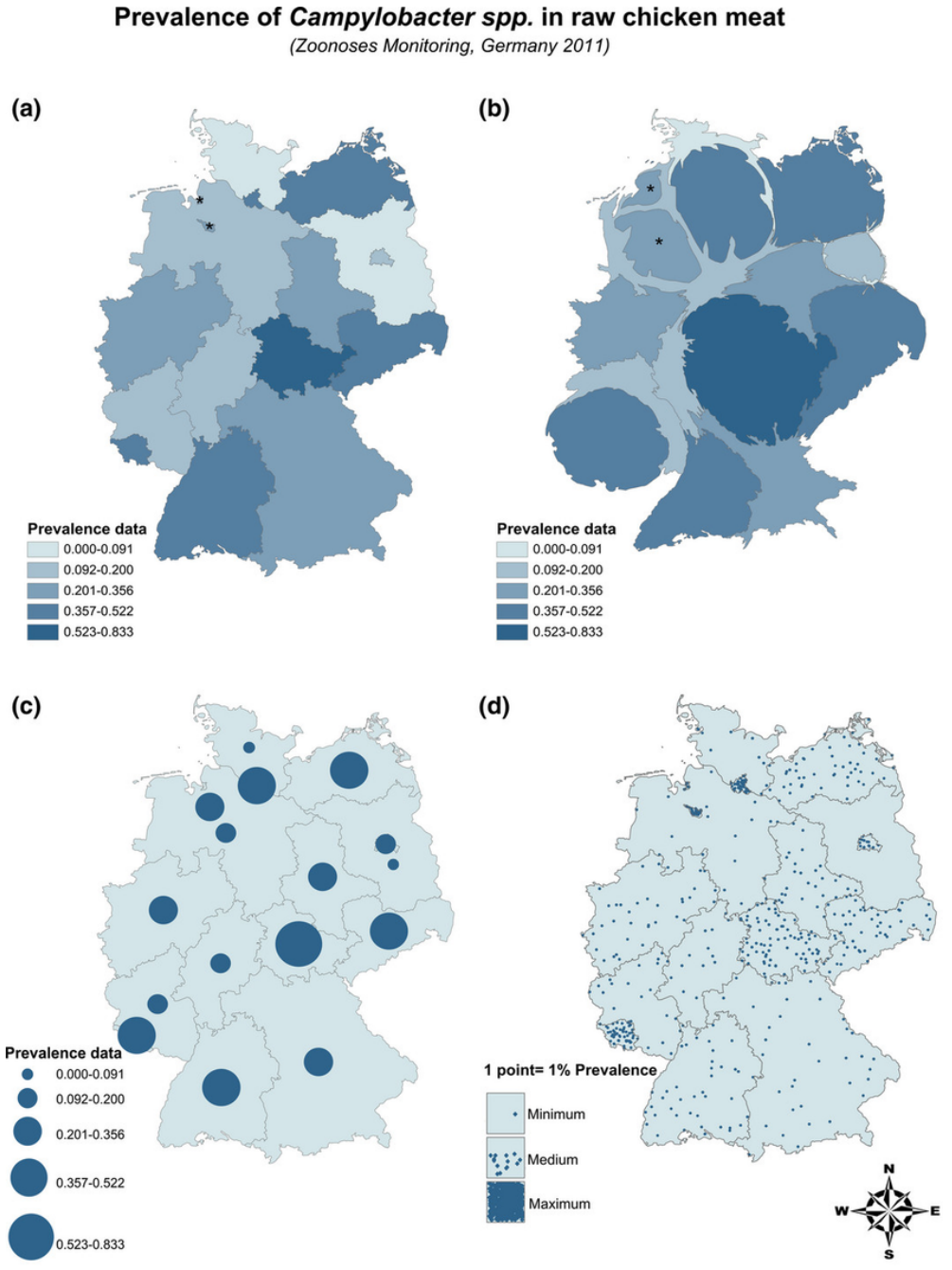
University of Trento --- M.Sc. Data Science

Data Visualization Lab Exam
10 January 2022

> Create a Jupyter and/or R Notebook, named
> *name_surname.*{*ipynb,Rmd*}

Answer the questions (in a Markdown cell/ as plain text) and solve the exercises listed hereafter:

1. [0-5 points] Describe in detail the meaning of the visual encoding elements in the following infographic reporting data concerning the presence of a bacterium in raw chicken meat in Germany.



*Federal State Bremen consists of two cities surrounded each by the Federal State Niedersachsen. The original two small areas in choropleth map become two big regions in cartogram.

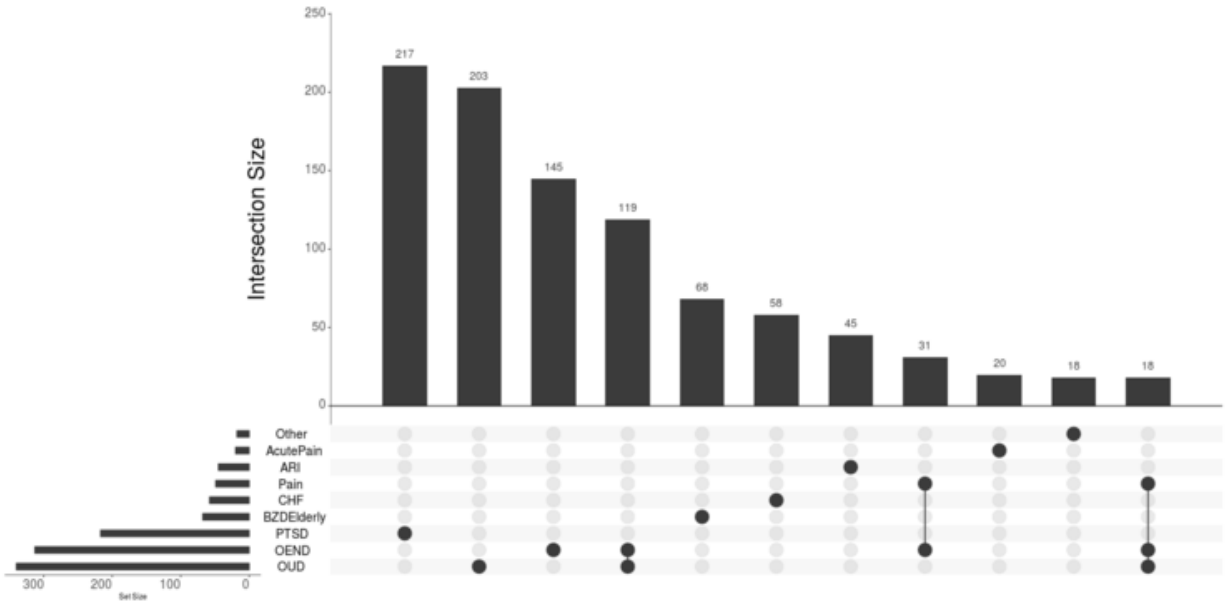
2. [0-5 points] Discuss in detail the concept of data-ink ratio in a graph and provide at least 2 practical examples different from those shown in the course slides.

3. [0-5 points] Discuss the meaning and the importance of the most important parameters of the t-SNE algorithm, providing examples showing graphical outputs corresponding to different values of parameters.

4. [0-7 points] The datafile us_states.csv collects information about recent demography of US States. Prepare an infographic with two panels describing some patterns inferred by (a subset of) the dataset, including a chloroplet map.

5. [0-7 points] Consider the datafile dermatology_data.csv, describing the histopathological and clinical data of 366 patients by 34 numerical features (with some missing values). The last column labels each patient according to his pathology, by the coding *1-psoriasis, 2-seboreic dermatitis, 3-lichen planus, 4-pityriasis rosea, 5-cronic dermatitis, 6-pityriasis rubra pilaris.*
   Prepare a data visualization consisting of 3 panels, showing the 2-D projection of the dataset by the MDS, the UMAP and the t-SNE algorithm; label each point with the corresponding sample name and color the points according to their pathology. What algorithm better separates the six classes of patients? Are there groups of diseases whose mutual distance change drastically in the three visualizations?

6. [0-7 points] Using the dataset visits.csv try to replicate the following plot.



Email the notebook(s) to *giuseppe.jurman@unitn.it* and please **wait for confirmation of correct receipt of the files before leaving the room**.

Notes:

- Exam is passed when at least 18 points are earned.
- If more than 30 points are achieved, the corresponding mark will be "30 cum laude"
- Use of the internet is allowed, but the candidate is expected to work individually.