Checklist: The Data Pile

This is a one-page checklist, part of the BENDER Series, with questions-to-ask-yourself *after* you have access to clinical imaging data, but, *before* you start to use it to build models.

- [ ] (1) List out pixel/voxel dimensions (number of elements) and spacing (in mm), for each subject in the data set (See here for how to list them from DICOM). We highly recommend saving this meta-data in a .csv file for each subject.
- [ ] (2) Build a reproducible pipeline to convert this raw data into processed/cleaned data in a different image format (e.g., individual 2D DICOM slices to NIfTI in case of 3D volumes). We highly recommend using cookiecutter to structure your project: especially separating raw, interim, and processed data to avoid corruption.
- [ ] (3) Depending on the modality/anatomy, consider reorienting/resampling everything to a standard space: LPS or RAS for example. Medical data includes metadata like coordinate origin and so on which is richer than natural images in computer vision. See here for more. Achtung : label data resampling cannot use linear/bicubic interpolation (they create new label categories): consider using nearest neighbour instead.
- [ ] (4) Wherever necessary, confirm that the data you have is properly anonymized (Batman, instead of Bruce Wayne; Spiderman, instead of Peter Parker, you get the drift). If you see any personally identifiable information, be sure to inform your supervisor to find out what action to take next. Here are some tools using pydicom, deid and MATLAB to anonymize DICOM files.
- [ ] (5) Plot data to visually check that the categories are correct (classification problems), and label masks are mapped to the right anatomy (segmentation problems). Verify for motion artefacts, incorrect volume of interest (e.g., missing anatomical information), etc. We recommend using 3D Slicer, ITKSnap and MATLAB Volume Viewer. Also check for duplicates in the image data (Here's a tool to automate this). If you see issues, challenge your clinical collaborator or the data provider ☺ (see Episode 02 for more!)
- [ ] (6) (optional) If your data comes from multiple sources, check if there are biases due to acquisition/processing protocols or imaging hardware vendors. Often, these biases could creep into models which can eventually lead to them learning these biases and hence generalizing worse on new data samples. Your clinical collaborator or the image metadata should have this information.
- [ ] (7) (optional) Consider registering all the subjects to a known atlas for that anatomy. For brain imaging studies, FSL lists some popular ones. This can be useful as further spatial normalization.

For questions/suggestions for improvements, please create an issue in this repository.