MIMIC-CXR

* Description:
  + dataset of chest radiographs in DICOM format with free-text radiology reports.
  + DICOM Images sometimes contain ``burned in'' annotations: areas where pixel values have been modified after image acquisition in order to display text.
  + Annotations contain relevant information including: **image orientation**, **anatomical position** of the subject, timestamp of image capture, and so on. The resulting image, with textual annotations encoded within the pixel themselves, is then transferred from the modality to PACS. Since the annotations are applied at the modality level, it is **impossible to recover the original image without annotations**.
  + Due to the burned in annotations, image pixel values required **de-identification**. A custom **algorithm** was developed which **removed** dates and patient identifiers, but **retained radiologically relevant information** such as orientation.
  + If a body of text was suspected to be PHI, all pixel values in a bounding box encompassing the PHI were set to black. For files cleaned of PHI, the DICOM header contains the location of any inserted black boxes.
  + Dataset contain both frontal & lateral view of chest in some studies.
* Folder Structure:
  + A set of **10 folders** (p10 - p19), each with ~6,500 sub-folders. Sub-folders are named according to the patient identifier “**subject\_id**”, and contain free-text reports and DICOM files for all studies for that patient
  + cxr-record-list.csv.gz - a compressed file providing the link between an image, its corresponding **study identifier**, and its corresponding **patient identifier**
  + patient has more than one study.
  + study is one record or one training example.
  + **study**\_**id**: 50,000,000 ~ 59,999,999.
  + **subject\_id**: 10,000,000 ~ 19,999,999.
  + **Where is labels** ?! in DICOM ?



* Meta-data:
  + dataset contains 377,110 images corresponding to 227,835 radiographic studies