MODEL TRAINING PIPELINE Raw Images + Structured **Nutriment Text information** Prototype (Images stored as PIL files with Nutriments info in dataframe) Production (Images in S3 Bucket. Nutriments in csv file) **OCR Extraction** Label Alignment with **Image Processing** (PyTesseract) Nutriments Convert to BIO Tags Convert to grayscale Extract word text + confidence score from Fuzzymatch OCR tokens to Denoise **Dataset with Processed** Match Tokens to available nutriment Binarize image BIO tags for NER **Images** Deskew information Extract bounding boxes Encoding with LayoutLMv2 **Model Training** Create Custom Pytorch Processor Dataset Setup training Handles alignment of tokens **Save Trained Model** Parameters Input Ids, to Boxes Attention Mask, Train LayoutLMV2 Bounding Boxes, Truncation Labels Padding etc.

