Paciscor: Design Doc

Framing:

- Given: → Set of images with "deals" (cupons)
  - → CSV of potential categories
  - → CSV of potential units of measure
- Needed: → Flyer_name
  - product_name
  - unit promo price
  - uom (unit of measure)
  - least unit for promo
  - save_per_unit
  - discount
  - organic (boolean)

⓪ Pre-processing:
→ Banners
→ Cropping
→

Diverging

- Potential Steps & Ideas (Solo Brainstorming)

① Segment the flyers {top ~~right~~ left coords, bottom right coords} ☆ not always □
→ Based on text → hard coded.
→ Based on white space
→ Based on non-uniform color patches (images of products)

② OCR
→ Specify to typed fonts
→ Specify within segments.
→ Color-agnostic

③ Are all prices red?
→ attribute in OCR output.

2.5: NLP
→ Signs of interest: {(0-9), $, /, lb, kg, g, mL, L, %}
→ Words of interest: {Organic, any of the ~1000 product names}
Regex
∄ all info every time.
→ Differentiate bolded words / text sizes.

③ Batching/Bagging algorithms } → See text from banners, include in OCR output.

④ Image recognition libraries...

⑤ Validation...?
→ We should generate our own validation set & testing procedure
☆ Autograder.