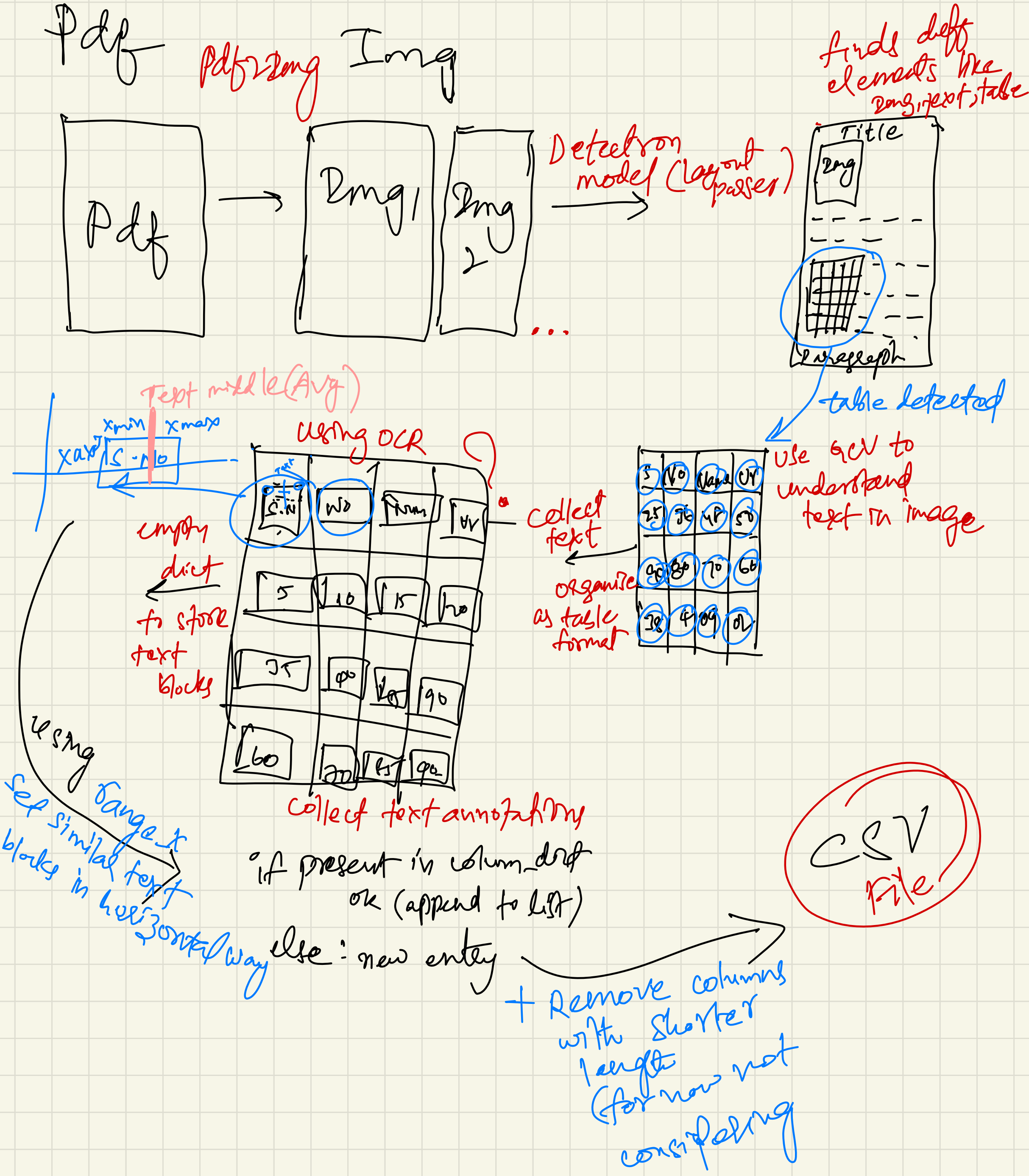
Pdf
Pdf2Img
Img
Pdf
Img1
Img 2
...
Detectron model (layout parser)
finds diff elements like Img, text, table
Title
Img
paragraph
table detected
Use OCV to understand text in image
collect text
organise as table format
S
NO
Name
Var
25
36
48
50
40
80
70
66
38
40
09
02
Using OCR
S.N
NO
Name
Var
5
10
15
20
35
40
80
90
60
20
45
90
Text middle (Avg)
xmin
xmax
xavg
S.No
empty dict to store text blocks
using range_x set similar text blocks in horizontal way
Collect text annotations
if present in column_dict
ok (append to list)
else: new entry
CSV File
+ Remove columns with shorter length (for now not considering

① drawbacks :- complex tables ?

② generalise. more

③ missing data (not able to capture all text from table)