De-obfuscation and reusability of scientific publications

An Optical Character Recognition and heuristic approach

Luke Darlow

How and why things have changed

With research came… more questions!

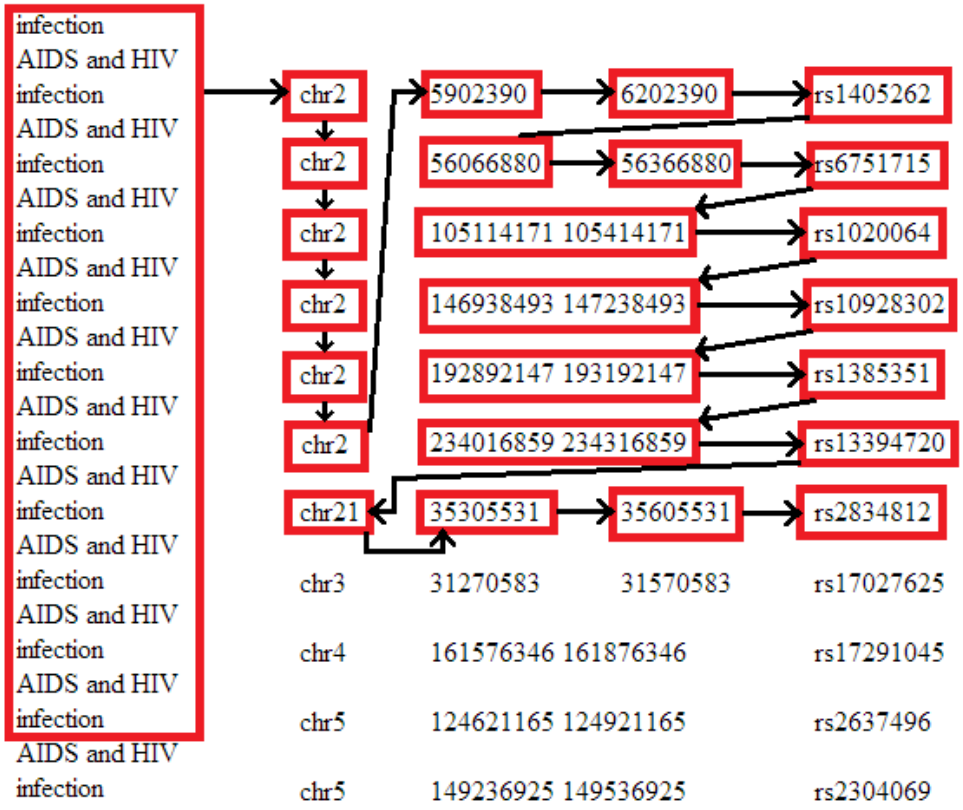
- Found that bioinformatics supplementary data wasn't stored in an easily reusable way
  - PDFs – Extracting data is a nightmare
  - Reusability and repetition are core to the scientific process

Goals

- Build a proof of concept system for supplementary data extraction
  - Finding the supplements: web scraping
  - Extracting the data (Largest chunk of research) assuming tables and that a PDF page only contains this
    - Excel (easy) and PDFs (not so easy)
  - Providing reusability
  - Allow for user intervention
- Explore different techniques (OCR) and test viability
- Learn where things can change and improve

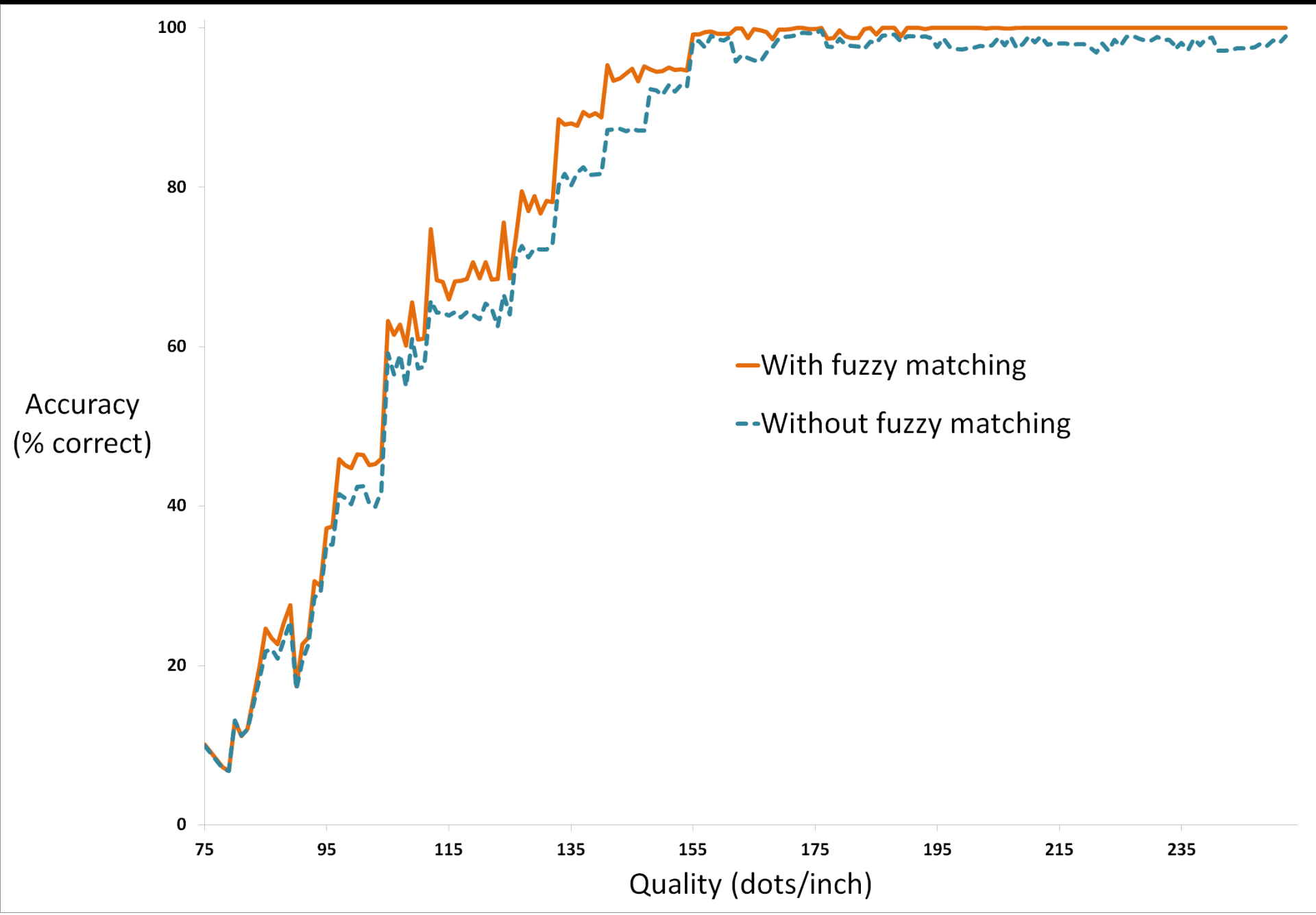
What is being done?

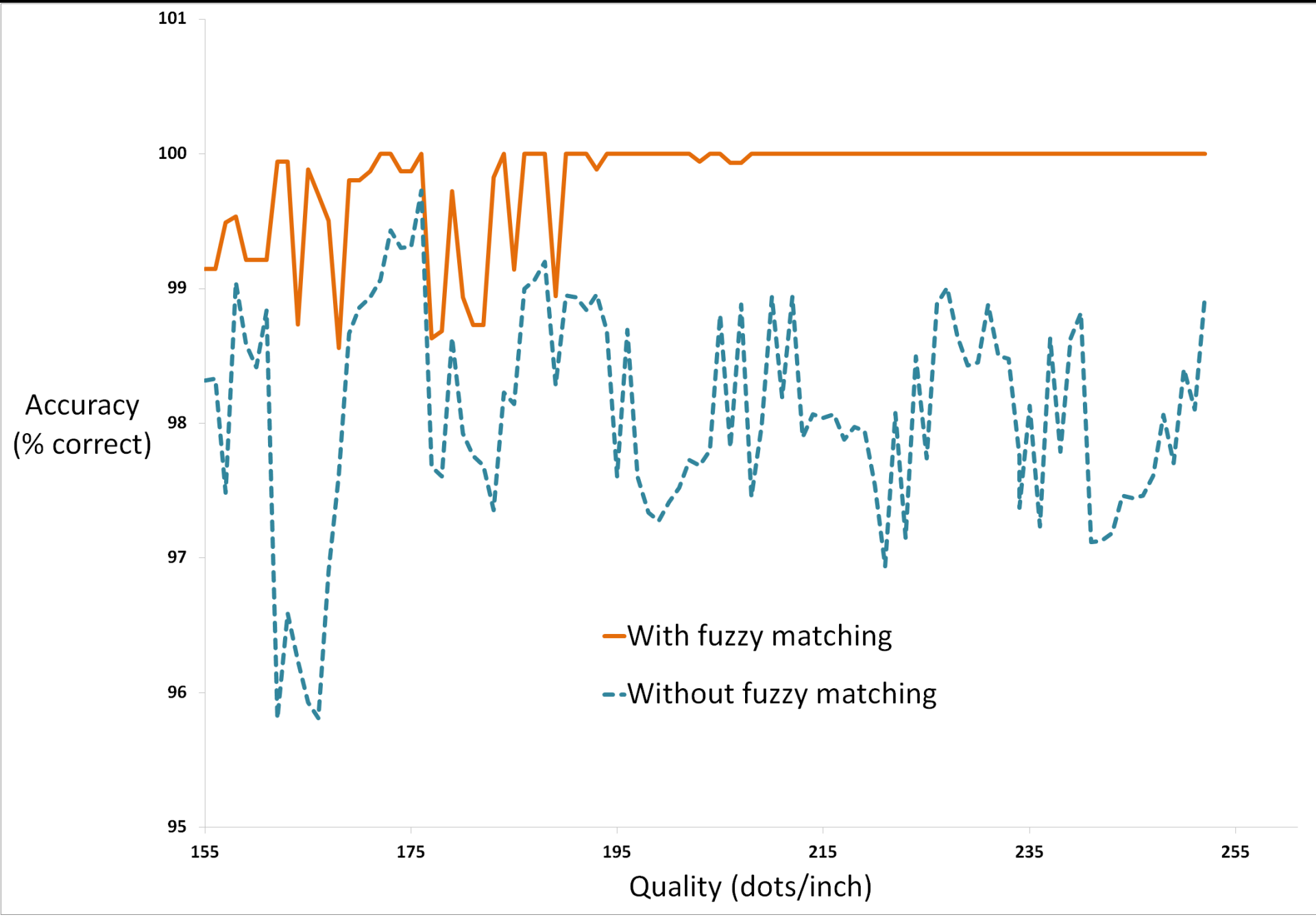
- Current default techniques fail unless carefully customized:
- Nobody uses OCR or image processing

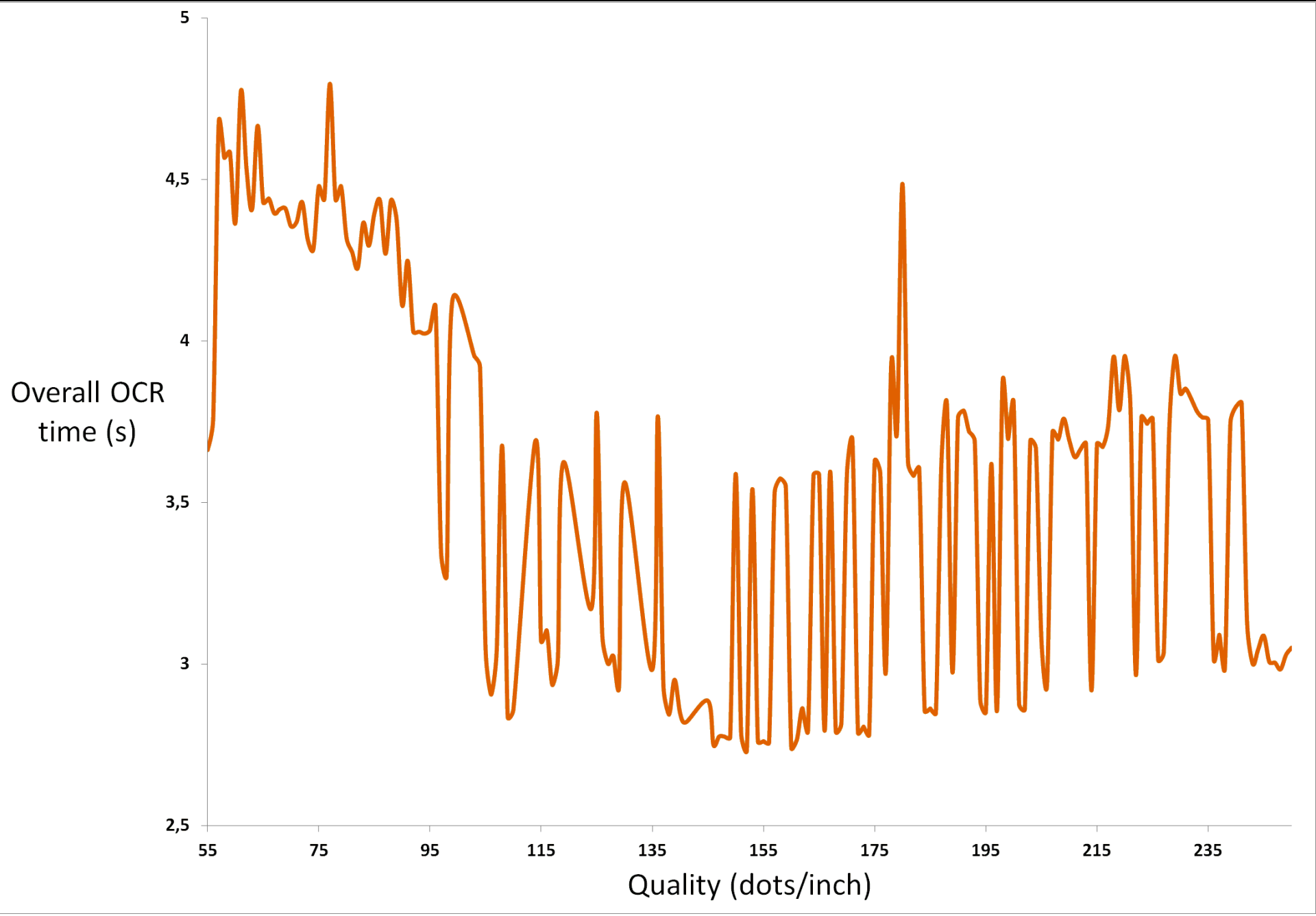


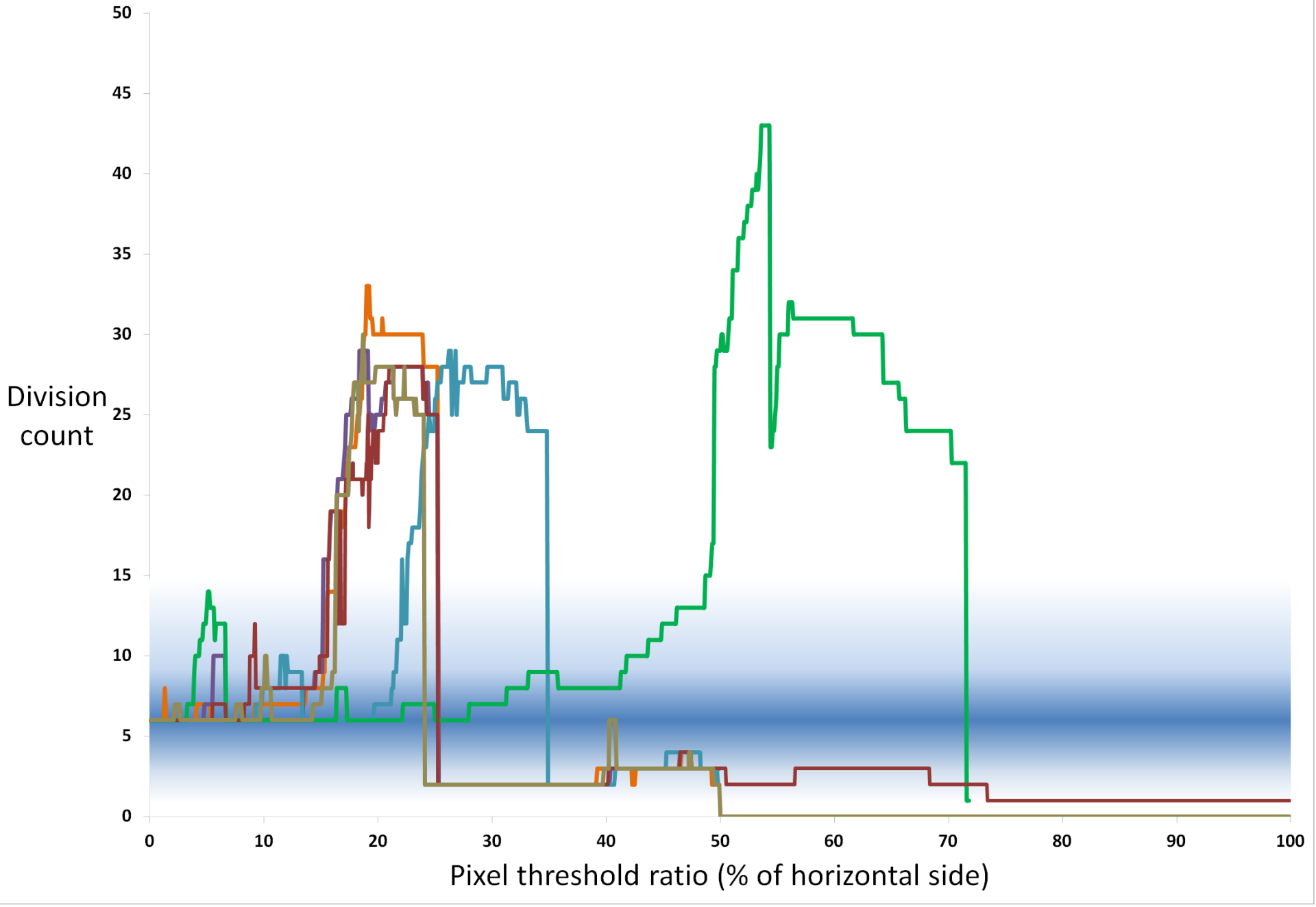
What I decided to try

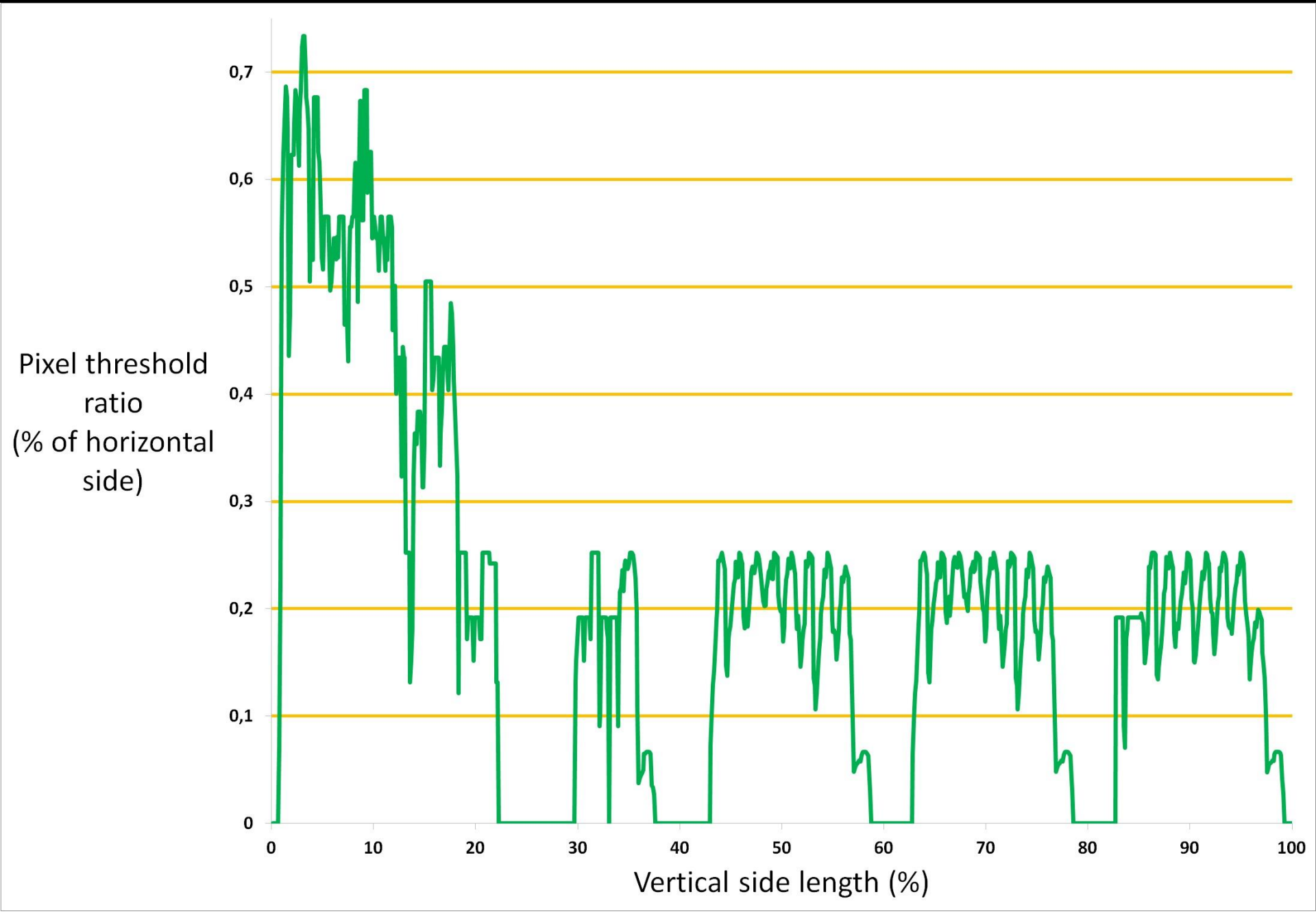
- Used Scrapy to show it is possible to find certain document links
- Used xlrd to extract from excel spreadsheets
- Approached PDFs differently
  - Turned a page into an image
  - Used image processing and heuristics to find table dimensions
  - Used Tesseract OCR with approximate string matching to extract cell contents
- Built a simple user interface

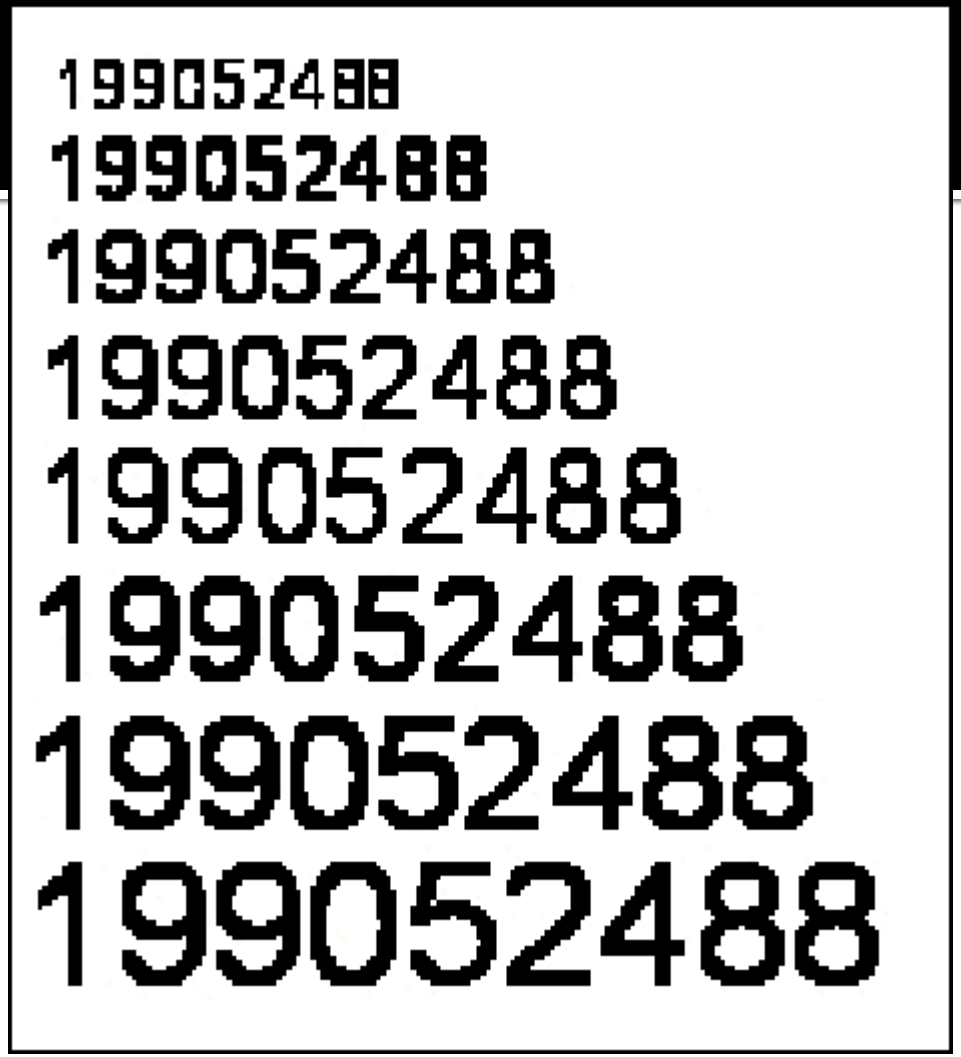

100
80
60
40
20
0
Accuracy
(% correct)
With fuzzy matching
Without fuzzy matching
75
95
115
135
155
175
195
215
235
Quality (dots/inch)


101
100
99
98
97
96
95
Accuracy
(% correct)
155
175
195
215
235
255
Quality (dots/inch)
With fuzzy matching
Without fuzzy matching


Overall OCR time (s)
5
4,5
4
3,5
3
2,5
55
75
95
115
135
155
175
195
215
235
Quality (dots/inch)


Division count
50
45
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
80
90
100
Pixel threshold ratio (% of horizontal side)


Pixel threshold ratio (% of horizontal side)
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0
10
20
30
40
50
60
70
80
90
100
Vertical side length (%)


19905288
19905288
19905288
19905288
19905288
19905288
19905288
19905288

Algorithmic nuances

- Row fixing algorithm
- Dark pixel counts
- OCR tweaks – single characters
- Fuzzy string matching

Findings

- Using OCR isn't always accurate enough
  - The text exists in a readable form
  - Need to develop better technique
- Cell dimension finding needs more robustness – smoothing pixel counts could help
- Accurate automated information extraction is made difficult by the popular PDF
- Dynamic resolution of links is a challenge when scraping

What's next?

- Improving the table dimension finding
  - Possible use of AI algorithms
- Implementing a coordinate to element extraction instead of OCR
- Building a robust user interface
- Moving from proof of concept to development

Questions?