Extracting data from PDFs

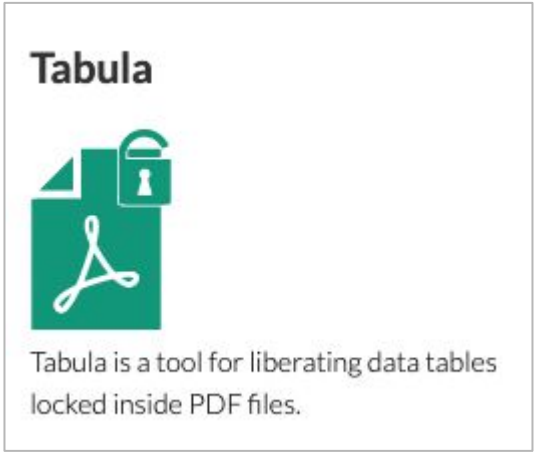
github.com/scottpham/nicar23-pdfs

Who am I

- Data reporter, formerly at BuzzFeed News
- Teach at the City College of New York (CUNY)
- I've worked on both solo investigations and large collaborations like the FinCEN Files
- I do data analysis in Python/Pandas but started with a lot of tools you're seeing today

What tools we're using

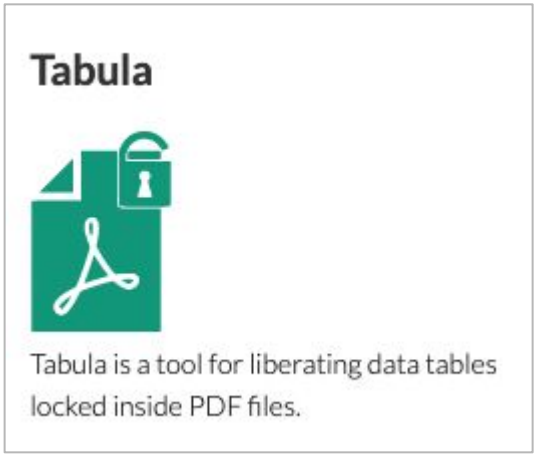
- Tabula - a free and open-source downloadable app for extracting tables
- Ocrmypdf - a free and open-source command-line application for optical character recognition



github.com/scottpham/nicar23-pdfs

What tools we're using

- Tabula - a free and open-source downloadable app for extracting tables
- Ocrmypdf - a free and open-source command-line application for optical character recognition



D/L instructions here!

github.com/scottpham/nicar23-pdfs

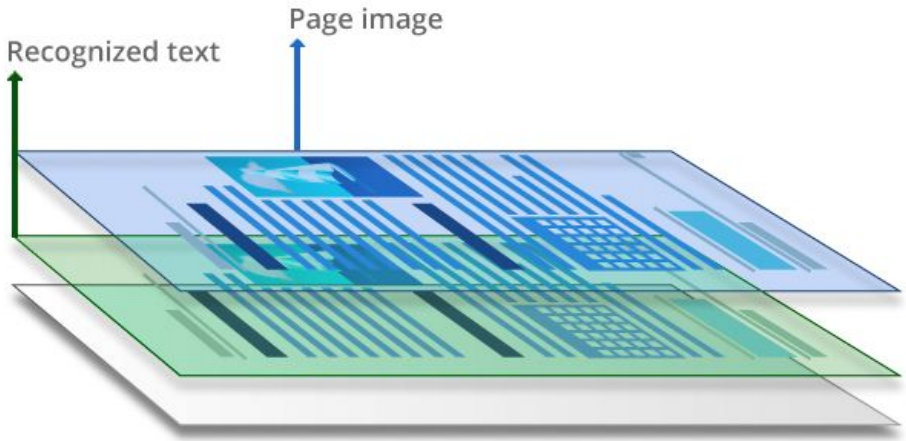
Why is this so damn hard?

- PDFs are for *seeing* not for *reading*
- Designed to look the same everywhere

github.com/scottpham/nicar23-pdfs

The two types of PDFs

- “Computer-generated PDFs”
  - Usually converted from a text document like Microsoft Word
- “Image-only” PDFs
  - Usually a digitally scanned version of print documents.
- OCR’d PDFs
  - Image-only PDFs that have been “augmented” with a searchable text layer



github.com/scottpham/nicar23-pdfs

But...why?

Some reasons why you might make an image-only PDF:

- Convenience/laziness
- Fear of manipulation
- Fear of redaction mistakes
- Actual malice

When to extract data from PDFs

Only when you must.

Downsides:

- Accuracy is hit-or-miss and potentially hard to fact check
- Sometimes it can take you longer than the alternatives

Example

Economic Report of the President

First google link:

Administration Priorities The

ECONOMIC REPORT OF THE PRESIDENT

CEA

The Economic Report of the President (ERP) is an annual report produced by the Council of Economic Advisers. An important vehicle for presenting the Administration's domestic and international economic policies, it provides an overview of the nation's economic progress with text and extensive data appendices.

See below to view or download the full reports from 2010–2022.

2022 Economic Report of the President

github.com/scottpham/nicar23-pdfs

Example

A 400 + page PDF.

Searchable, but filled with data tables.

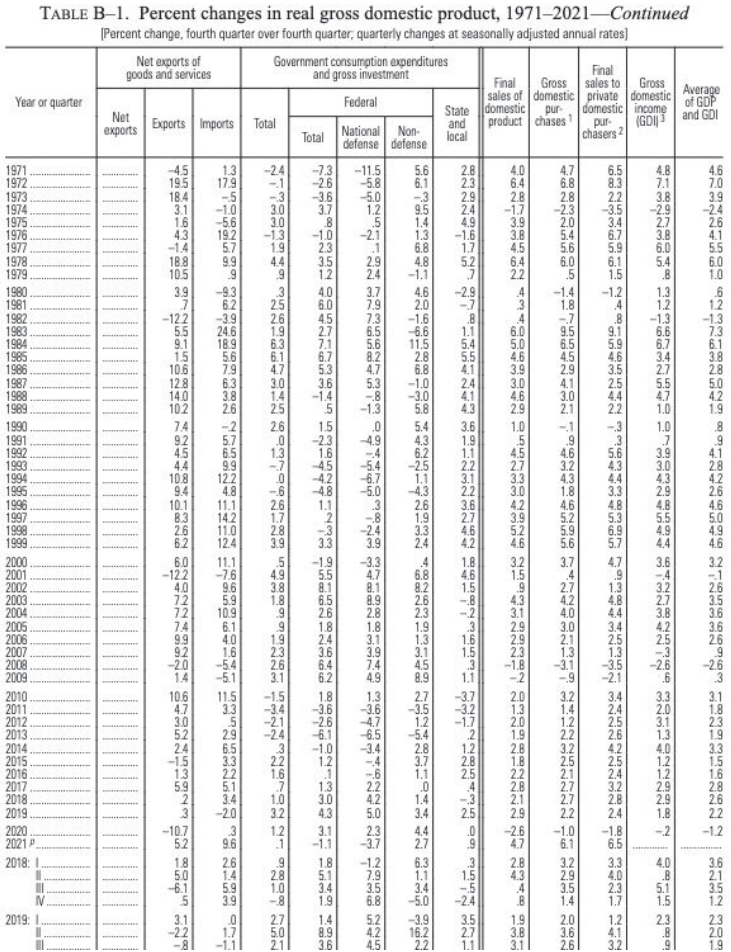


TABLE B–1. Percent changes in real gross domestic product, 1971–2021—*Continued*

[Percent change, fourth quarter over fourth quarter; quarterly changes at seasonally adjusted annual rates]

Year or quarter	Net exports of goods and services			Government consumption expenditures and gross investment					Final sales of domestic product	Gross domestic purchases [1]	Final sales to private domestic purchasers [2]	Gross domestic income (GDI) [1]	Average of GDP and GDI
	Net exports	Exports	Imports	Total	Federal			State and local					
					Total	National defense	Non-defense						
1971		−4.5	1.3	−2.4	−7.3	−11.5	5.6	2.8	4.0	4.7	6.5	4.8	4.6
1972		19.5	17.9	−.1	−2.6	−5.8	6.1	2.3	6.4	6.8	8.3	7.1	7.0
1973		18.4	−.5	−.3	−3.6	−5.0	−.3	2.9	2.8	2.8	2.2	3.8	3.9
1974		3.1	−1.0	3.0	3.7	1.2	9.5	2.4	−1.7	−2.3	−3.5	−2.9	−2.4
1975		1.6	−5.6	3.0	.8	.5	1.4	4.9	3.9	2.0	3.4	2.7	2.6
1976		4.3	19.2	−1.3	−1.0	−2.1	1.3	−1.6	3.8	5.4	6.7	3.8	4.1
1977		−1.4	5.7	1.9	2.3	.1	6.8	1.7	4.5	5.6	5.9	6.0	5.5
1978		18.8	9.9	4.4	3.5	2.9	4.8	5.2	6.4	6.0	6.1	5.4	6.0
1979		10.5	.9	.9	1.2	2.4	−1.1	.7	2.2	.5	1.5	.8	1.0
1980		3.9	−9.3	.3	4.0	3.7	4.6	−2.9	.4	−1.4	−1.2	1.3	.6
1981		.7	6.2	2.5	6.0	7.9	2.0	−.7	.3	1.8	.4	1.2	1.2
1982		−12.2	−3.9	2.6	4.5	7.3	−1.6	.8	.4	−.7	.8	−1.3	−1.3
1983		5.5	24.6	1.9	2.7	6.5	−6.6	1.1	6.0	9.5	9.1	6.6	7.3
1984		9.1	18.9	6.3	7.1	5.6	11.5	5.4	5.0	6.5	5.9	6.7	6.1
1985		1.5	5.6	6.1	6.7	8.2	2.8	5.5	4.6	4.5	4.6	3.4	3.8
1986		10.6	7.9	4.7	5.3	4.7	6.8	4.1	3.9	2.9	3.5	2.7	2.8
1987		12.8	6.3	3.0	3.6	5.3	−1.0	2.4	3.0	4.1	2.5	5.5	5.0
1988		14.0	3.8	1.4	−1.4	−.8	−3.0	4.1	4.6	3.0	4.4	4.7	4.2
1989		10.2	2.6	2.5	.5	−1.3	5.8	4.3	2.9	2.1	2.2	1.0	1.9
1990		7.4	−.2	2.6	1.5	.0	5.4	3.6	1.0	−.1	−.3	1.0	.8
1991		9.2	5.7	.0	−2.3	−4.9	4.3	1.9	.5	.9	.3	.7	.9
1992		4.5	6.5	1.3	1.6	−.4	6.2	1.1	4.5	4.6	5.6	3.9	4.1
1993		4.4	9.9	−.7	−4.5	−5.4	−2.5	2.2	2.7	3.2	4.3	3.0	2.8
1994		10.8	12.2	.0	−4.2	−6.7	1.1	3.1	3.3	4.3	4.4	4.3	4.2
1995		9.4	4.8	−.6	−4.8	−5.0	−4.3	2.2	3.0	1.8	3.3	2.9	2.6
1996		10.1	11.1	2.6	1.1	.3	2.6	3.6	4.2	4.6	4.8	4.8	4.6
1997		8.3	14.2	1.7	.2	−.8	1.9	2.7	3.9	5.2	5.3	5.5	5.0
1998		2.6	11.0	2.8	−.3	−2.4	3.3	4.6	5.2	5.9	6.9	4.9	4.9
1999		6.2	12.4	3.9	3.3	3.9	2.4	4.2	4.6	5.6	5.7	4.4	4.6
2000		6.0	11.1	.5	−1.9	−3.3	.4	1.8	3.2	3.7	4.7	3.6	3.2
2001		−12.2	−7.6	4.9	5.5	4.7	6.8	4.6	1.5	.4	.9	−.4	−.1
2002		4.0	9.6	3.8	8.1	8.1	8.2	1.5	.9	2.7	1.3	3.2	2.6
2003		7.2	5.9	1.8	6.5	8.9	2.6	−.8	4.3	4.2	4.8	2.7	3.5
2004		7.2	10.9	.9	2.6	2.8	2.3	−.2	3.1	4.0	4.4	3.8	3.6
2005		7.4	6.1	.9	1.8	1.8	1.9	.3	2.9	3.0	3.4	4.2	3.6
2006		9.9	4.0	1.9	2.4	3.1	1.3	1.6	2.9	2.1	2.5	2.5	2.6
2007		9.2	1.6	2.3	3.6	3.9	3.1	1.5	2.3	1.3	1.3	−.3	.9
2008		−2.0	−5.4	2.6	6.4	7.4	4.5	.3	−1.8	−3.1	−3.5	−2.6	−2.6
2009		1.4	−5.1	3.1	6.2	4.9	8.9	1.1	−.2	−.9	−2.1	.6	.3
2010		10.6	11.5	−1.5	1.8	1.3	2.7	−3.7	2.0	3.2	3.4	3.3	3.1
2011		4.7	3.3	−3.4	−3.6	−3.6	−3.5	−3.2	1.3	1.4	2.4	2.0	1.8
2012		3.0	.5	−2.1	−2.6	−4.7	1.2	−1.7	2.0	1.2	2.5	3.1	2.3
2013		5.2	2.9	−2.4	−6.1	−6.5	−5.4	.2	1.9	2.2	2.6	1.3	1.9
2014		2.4	6.5	.3	−1.0	−3.4	2.8	1.2	2.8	3.2	4.2	4.0	3.3
2015		−1.5	3.3	2.2	1.2	−.4	3.7	2.8	1.8	2.5	2.5	1.2	1.5
2016		1.3	2.2	1.6	.1	−.6	1.1	2.5	2.2	2.1	2.4	1.2	1.6
2017		5.9	5.1	.7	1.3	2.2	.0	.4	2.8	2.7	3.2	2.9	2.8
2018		.2	3.4	1.0	3.0	4.2	1.4	−.3	2.1	2.7	2.8	2.9	2.6
2019		.3	−2.0	3.2	4.3	5.0	3.4	2.5	2.9	2.2	2.4	1.8	2.2
2020		−10.7	.3	1.2	3.1	2.3	4.4	.0	−2.6	−1.0	−1.8	−.2	−1.2
2021 [p]		5.2	9.6	.1	−1.1	−3.7	2.7	.9	4.7	6.1	6.5		
2018: I		1.8	2.6	.9	1.8	−1.2	6.3	.3	2.8	3.2	3.3	4.0	3.6
II		5.0	1.4	2.8	5.1	7.9	1.1	1.5	4.3	2.9	4.0	.8	2.1
III		−6.1	5.9	1.0	3.4	3.5	3.4	−.5	.4	3.5	2.3	5.1	3.5
IV		.5	3.9	−.8	1.9	6.8	−5.0	−2.4	.8	1.4	1.7	1.5	1.2
2019: I		3.1	.0	2.7	1.4	5.2	−3.9	3.5	1.9	2.0	1.2	2.3	2.3
II		−2.2	1.7	5.0	8.9	4.2	16.2	2.7	3.8	3.6	4.1	.8	2.0
III		−.8	−1.1	2.1	3.6	4.5	2.2	1.1	3.1	2.6	3.2	.9	1.9

github.com/scottpham/nicar23-pdfs

Example

Second google link:

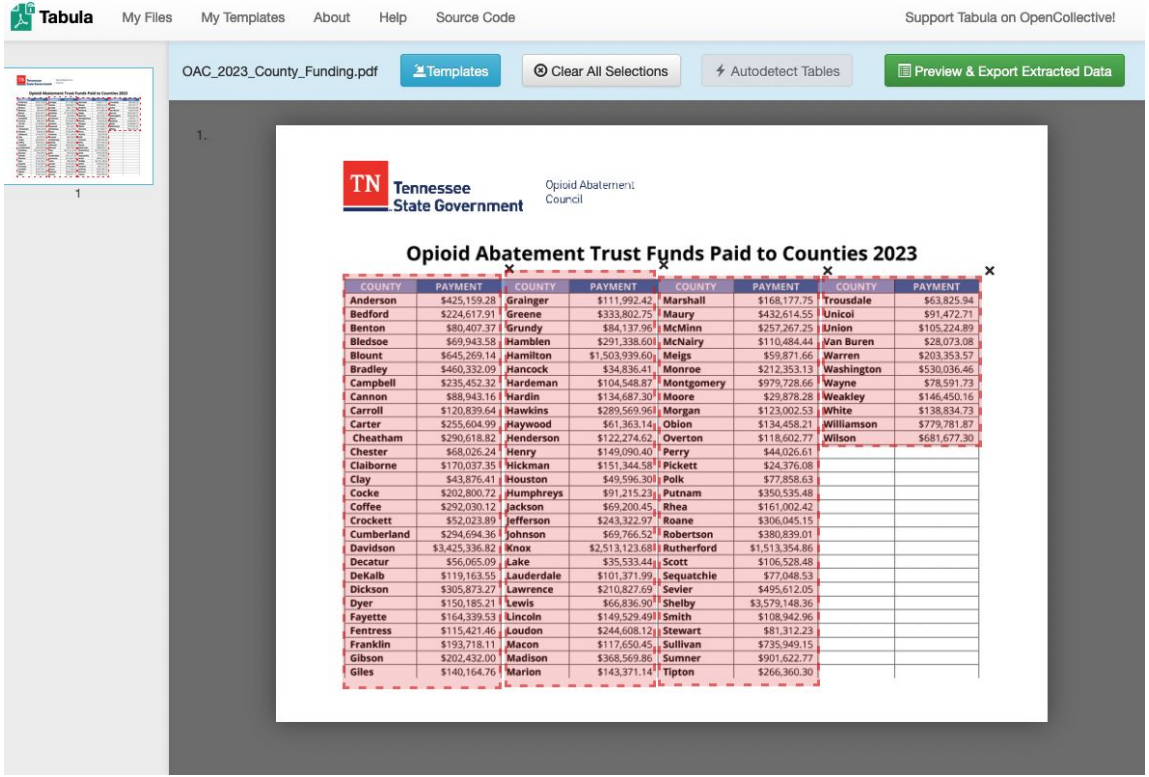
GovInfo site with XLS files for every table

github.com/scottpham/nicar23-pdfs

Working with computer-generated-PDFs

- Tabula
- Drag and drop sites - easy, but may require payment
  - CleverPDF
  - PDF to Excel
- Code
  - Pdfplumber
  - Tabula-py

Hands-on: Working with Tabula



COUNTY	PAYMENT	COUNTY	PAYMENT	COUNTY	PAYMENT	COUNTY	PAYMENT
Anderson	$425,159.28	Grainger	$111,992.42	Marshall	$168,177.75	Trousdale	$63,825.94
Bedford	$224,617.91	Greene	$333,802.75	Maury	$432,614.55	Unicoi	$91,472.71
Benton	$80,407.37	Grundy	$84,137.96	McMinn	$257,267.25	Union	$105,224.89
Bledsoe	$69,943.58	Hamblen	$291,338.60	McNairy	$110,484.44	Van Buren	$28,073.08
Blount	$645,269.14	Hamilton	$1,503,939.60	Meigs	$59,871.66	Warren	$203,353.57
Bradley	$460,332.09	Hancock	$34,836.41	Monroe	$212,353.13	Washington	$530,036.46
Campbell	$235,452.32	Hardeman	$104,548.87	Montgomery	$979,728.66	Wayne	$78,591.73
Cannon	$88,943.16	Hardin	$134,687.30	Moore	$29,878.28	Weakley	$146,450.16
Carroll	$120,839.64	Hawkins	$289,569.96	Morgan	$123,002.53	White	$138,834.73
Carter	$255,604.99	Haywood	$61,363.14	Obion	$134,458.21	Williamson	$779,781.87
Cheatham	$290,618.82	Henderson	$122,274.62	Overton	$118,602.77	Wilson	$681,677.30
Chester	$68,026.24	Henry	$149,090.40	Perry	$44,026.61		
Claiborne	$170,037.35	Hickman	$151,344.58	Pickett	$24,376.08		
Clay	$43,876.41	Houston	$49,596.30	Polk	$77,858.63		
Cocke	$202,800.72	Humphreys	$91,215.23	Putnam	$350,535.48		
Coffee	$292,030.12	Jackson	$69,200.45	Rhea	$161,002.42		
Crockett	$52,023.89	Jefferson	$243,322.97	Roane	$306,045.15		
Cumberland	$294,694.36	Johnson	$69,766.52	Robertson	$380,839.01		
Davidson	$3,425,336.82	Knox	$2,513,123.68	Rutherford	$1,513,354.86		
Decatur	$56,065.09	Lake	$35,533.44	Scott	$106,528.48		
DeKalb	$119,163.55	Lauderdale	$101,371.99	Sequatchie	$77,048.53		
Dickson	$305,873.27	Lawrence	$210,827.69	Sevier	$495,612.05		
Dyer	$150,185.21	Lewis	$66,836.90	Shelby	$3,579,148.36		
Fayette	$164,339.53	Lincoln	$149,529.49	Smith	$108,942.96		
Fentress	$115,421.46	Loudon	$244,608.12	Stewart	$81,312.23		
Franklin	$193,718.11	Macon	$117,650.45	Sullivan	$735,949.15		
Gibson	$202,432.00	Madison	$368,569.86	Sumner	$901,622.77		
Giles	$140,164.76	Marion	$143,371.14	Tipton	$266,360.30		

github.com/scottpham/nicar23-pdfs

OCR

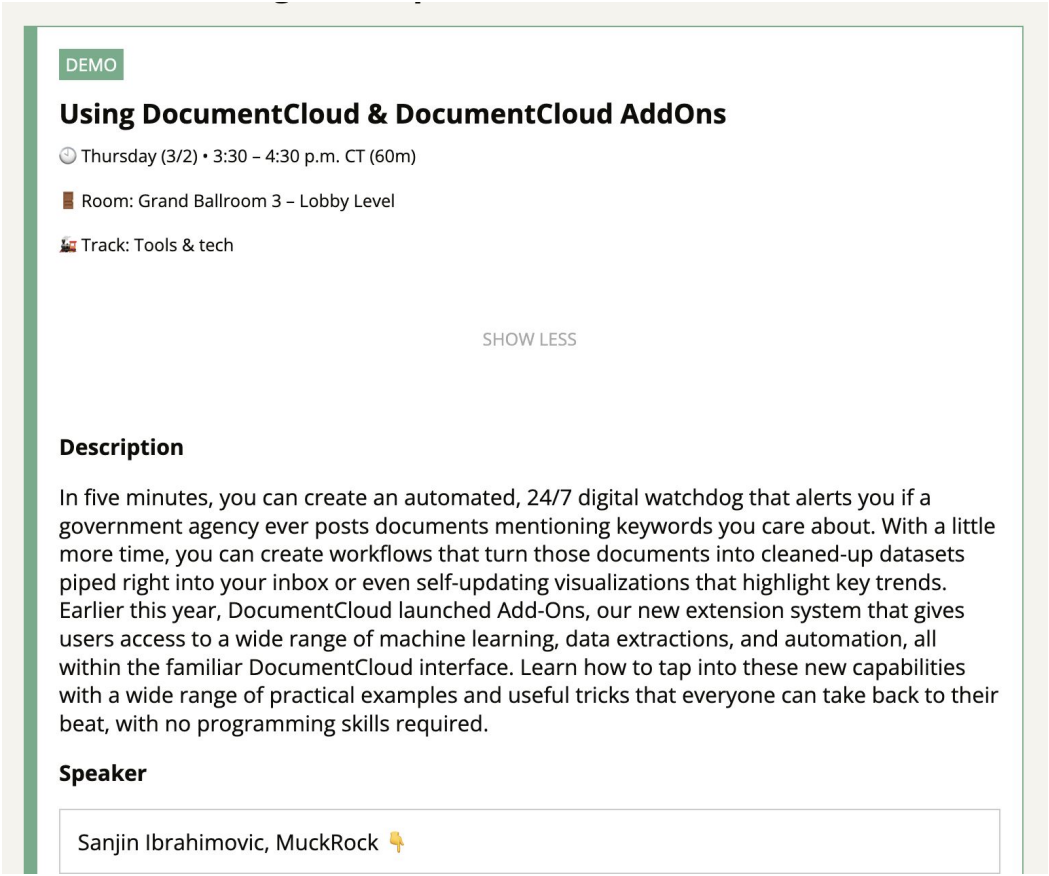
(Optical Character Recognition)

- Commonly needed with court documents, older documents
- Easier with text-only or text-mostly documents
  - Tabular data is troublesome
- Images, handwriting, and redaction can mess you up
  - Some algorithms are better than others

OCR

Your options:

- DocumentCloud
- Google PinPoint
- Adobe Acrobat Pro
  - Short guide
- Code
  - pytesseract
- Command-line utilities
  - OCRmyPDF



DEMO

Using DocumentCloud & DocumentCloud AddOns

Thursday (3/2) • 3:30 – 4:30 p.m. CT (60m)

Room: Grand Ballroom 3 – Lobby Level

Track: Tools & tech

SHOW LESS

Description

In five minutes, you can create an automated, 24/7 digital watchdog that alerts you if a government agency ever posts documents mentioning keywords you care about. With a little more time, you can create workflows that turn those documents into cleaned-up datasets piped right into your inbox or even self-updating visualizations that highlight key trends. Earlier this year, DocumentCloud launched Add-Ons, our new extension system that gives users access to a wide range of machine learning, data extractions, and automation, all within the familiar DocumentCloud interface. Learn how to tap into these new capabilities with a wide range of practical examples and useful tricks that everyone can take back to their beat, with no programming skills required.

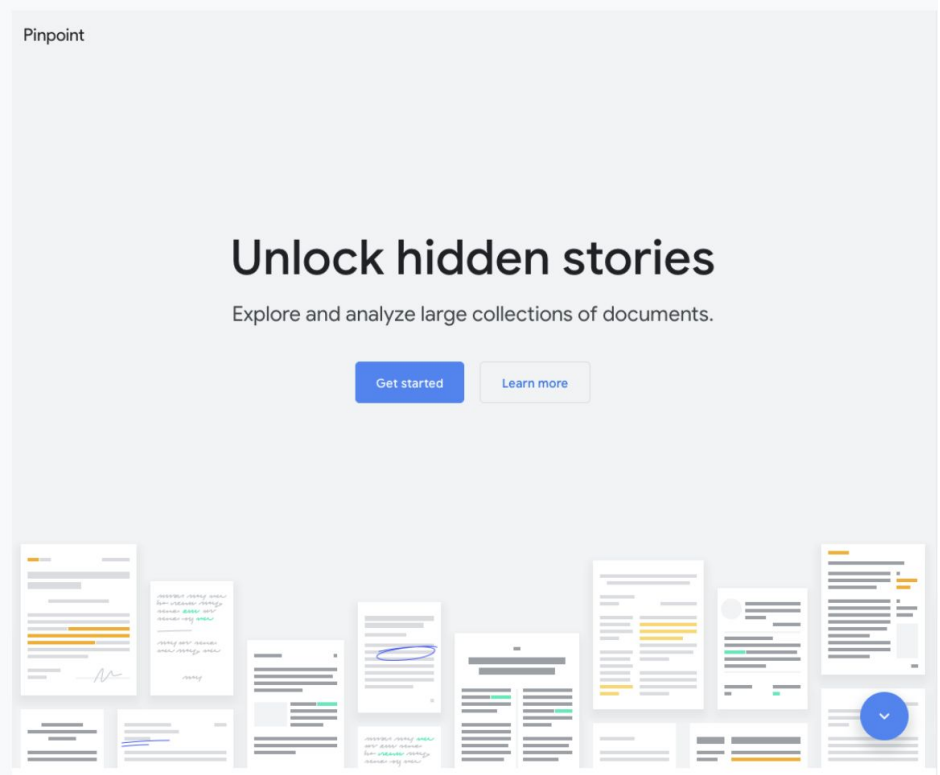
Speaker

Sanjin Ibrahimovic, MuckRock

github.com/scottpham/nicar23-pdfs

Google Pinpoint

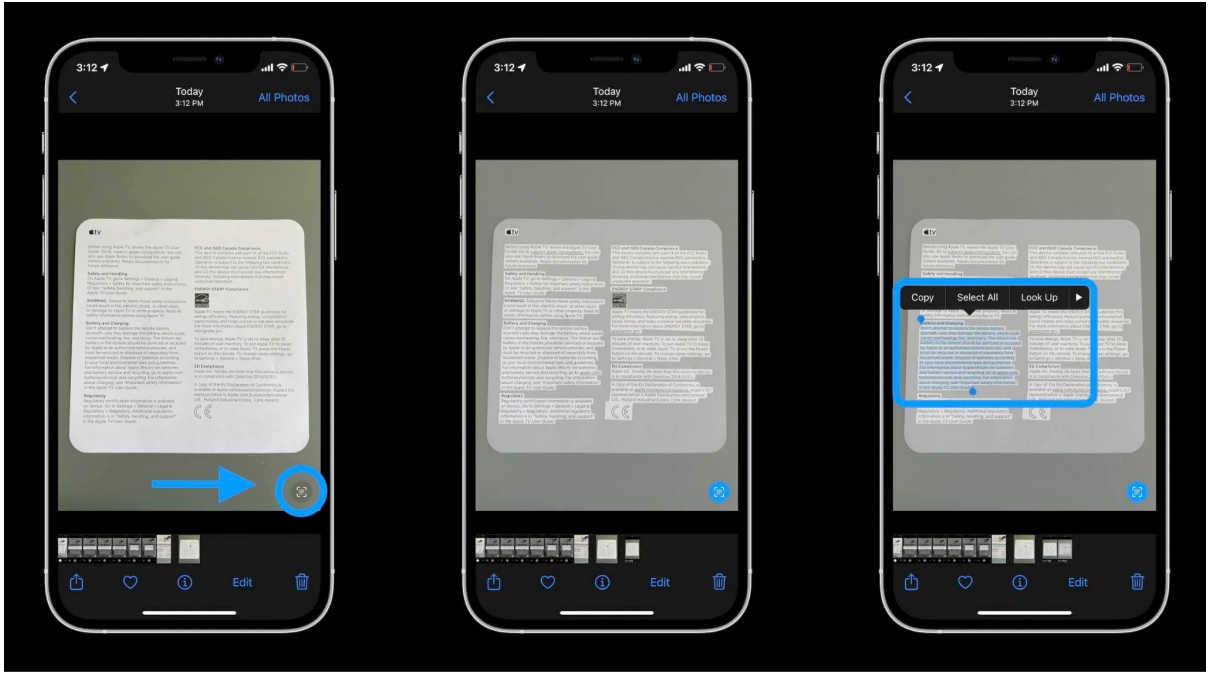
- Part of Google's "Journalist Studio"
- Experimental structured data scraping in beta



github.com/scottpham/nicar23-pdfs

Your phone (seriously)

- iPhone photos app, notes app
  - Requires iPhone Xs or later (2018+)
- Android: People seem to like “Text Fairy”



github.com/scottpham/nicar23-pdfs

When to use a command-line activity

- Your files are big
- You need it fast
- You need a little more flexibility than the GUI tools

Example

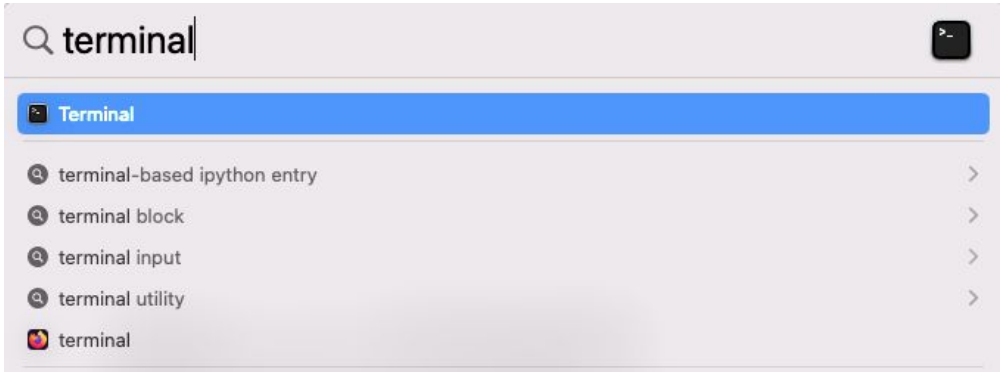
The Mueller Report:

- More than 400 pages
- Intense competition to publish upon release
- Completely unsearchable
- Embarrassment to the “PDF Association”

github.com/scottpham/nicar23-pdfs

Hands-on: OCRmyPDF

1. Download the PDF: `comey_memoranda.pdf` from the Github repo
2. Move it to your computer's Desktop
3. Open up your terminal
    a. Mac users: Cmd + SPACE will open spotlight. Then just type "terminal" and hit Enter

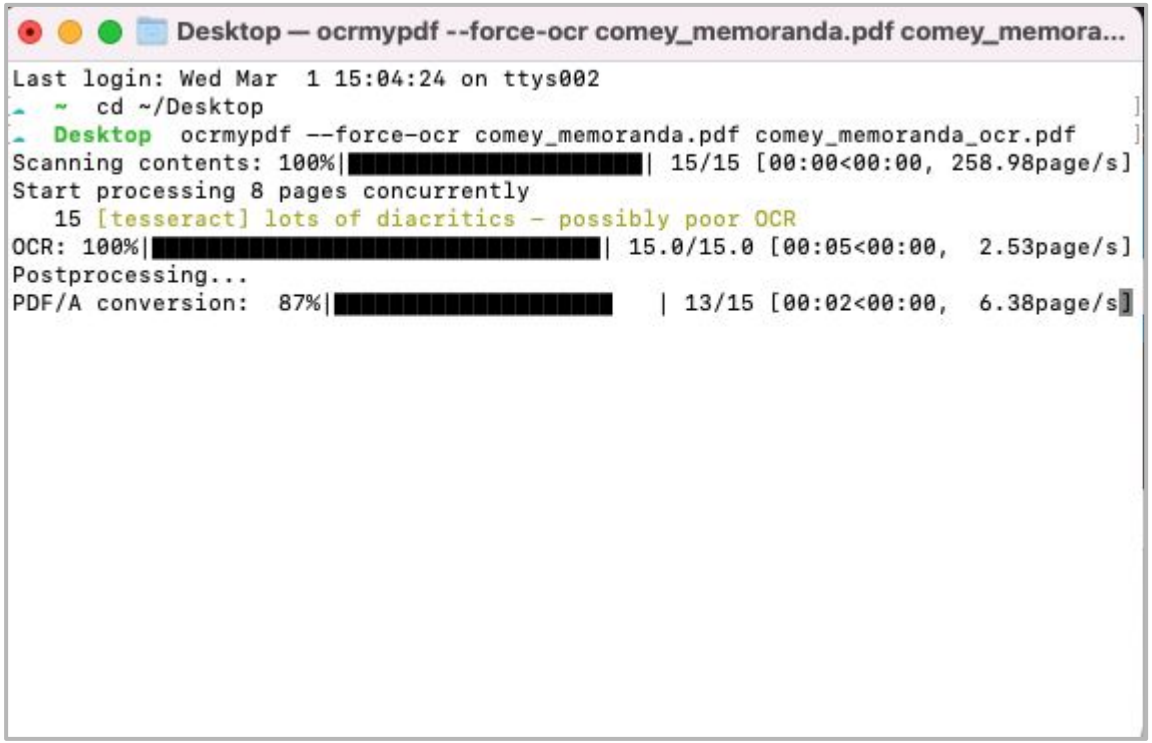


github.com/scottpham/nicar23-pdfs

Hands-on: OCRmyPDF

```
cd ~/Desktop
```

```
ocrmypdf –force-ocr
comey_memoranda.pdf
comey_memoranda_ocr.pdf
```



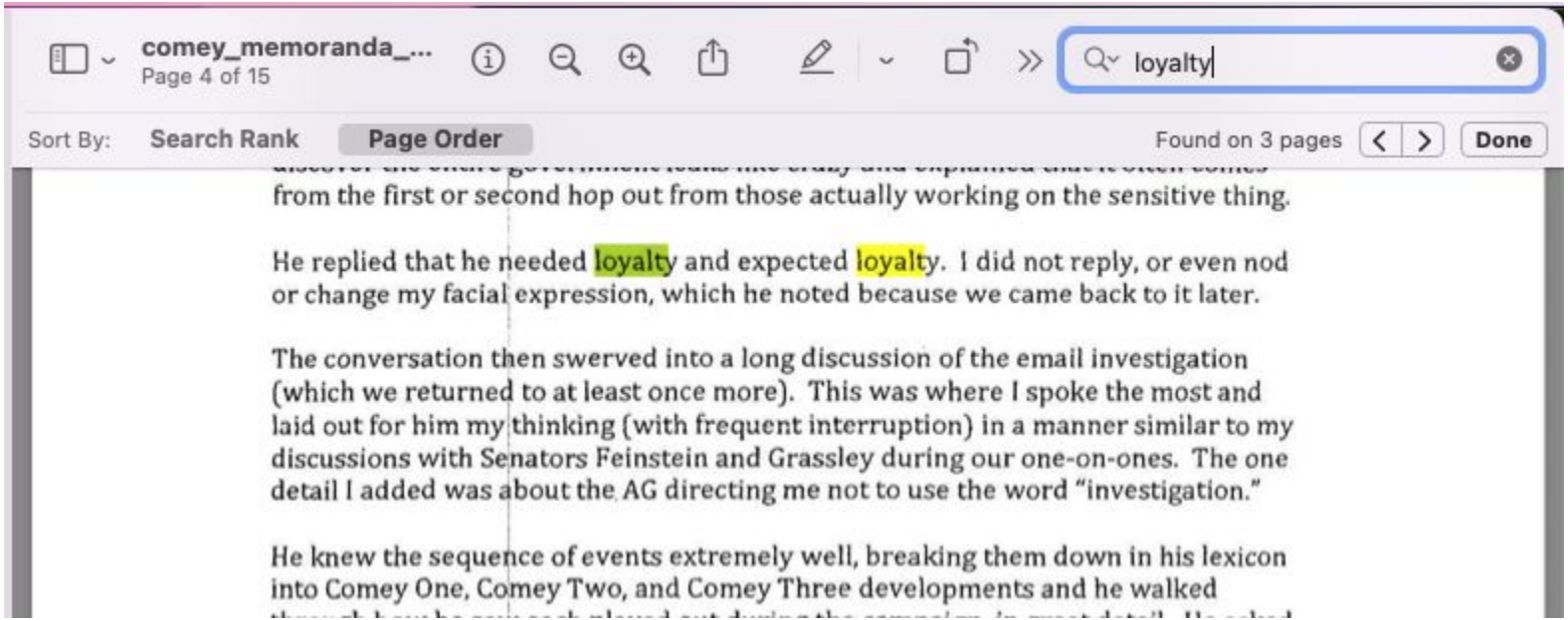
github.com/scottpham/nicar23-pdfs

Hands-on: OCRmyPDF

```
cd ~/Desktop
```

```
ocrmypdf --force-ocr comey_memoranda.pdf comey_memoranda_ocr.pdf
```

Hands-on: OCRmyPDF



github.com/scottpham/nicar23-pdfs