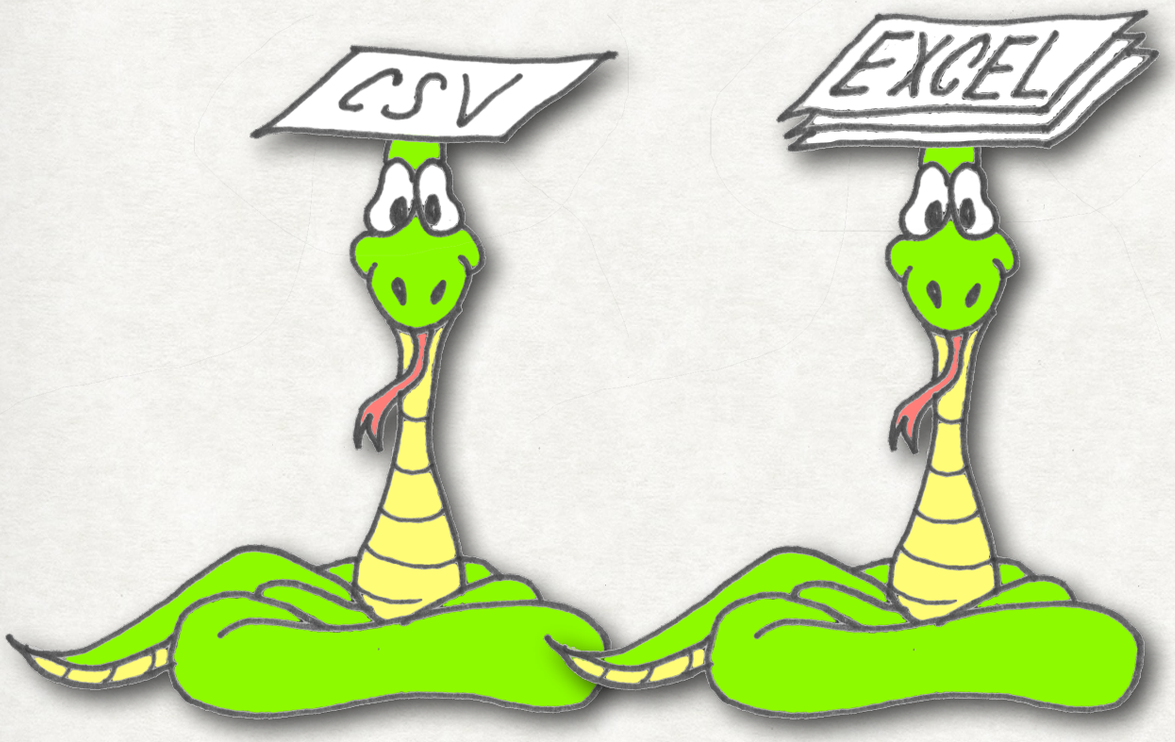
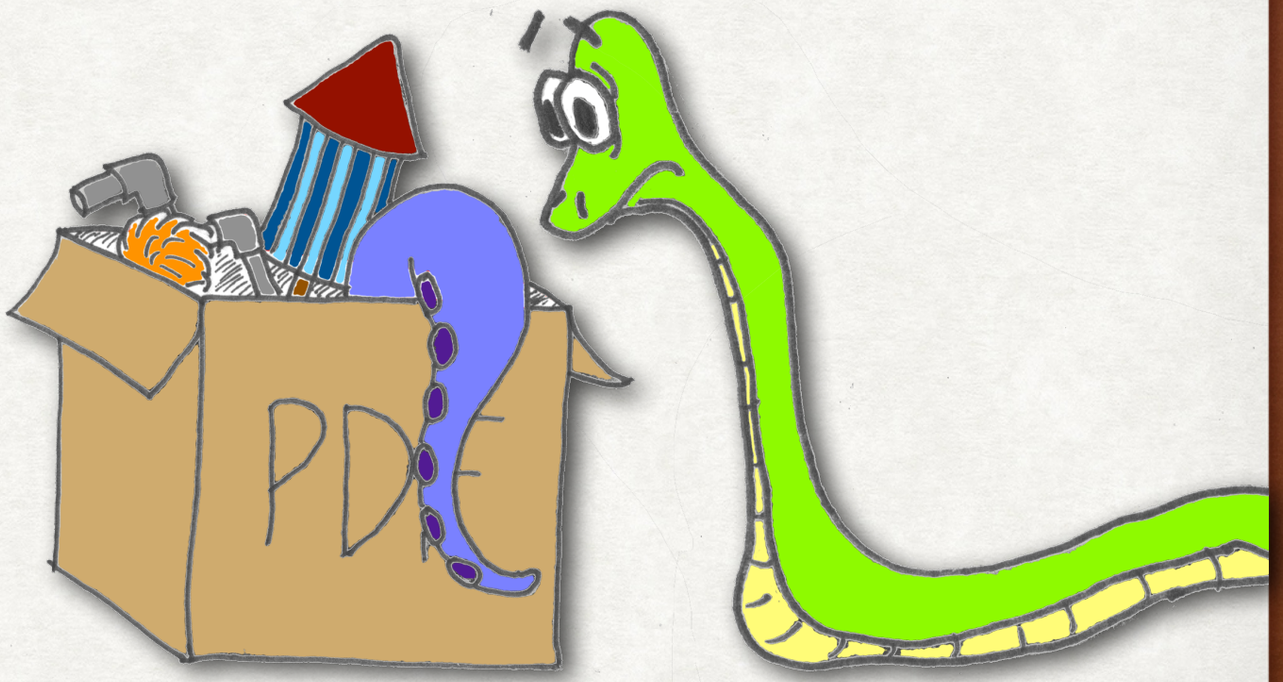
PICKING PDFS APART WITH PYTHON

Shannon DeArmond, GISP

GIS Supervisor, Far Western Anthropological Research Group, Inc.

PDFs are Complicated





pdfrw PDFTron reportlab slate pdfquery pdf-table-extractor pypdf2 pdftk Code2pdf FetchPDF fdfgen pdfcomparator pdfmerge pdfpages qpdf pdf-table-extractor pdfminer ghostscript pdfsplit pdf2text XPDF pdffonts pdf-link-checker pdfbookmark pdfquery pdf-page-counter pdfplumber concatPDF geopdf pdf_extractor

*random sampling from https://pypi.python.org/pypi?%3Aaction=search&term=PDF&submit=search

Make PDFs

pdfrw PDFTron reportlab slate pdfquery pdf-table-

extractor pypdf2 pdftk Code2pdf FetchPDF fdfgen

Manipulate PDFs

pdfcomparator pdfmerge pdfpages qpdf pdf-table-extractor

pdfminer ghostscript pdfsplit pdf2text XPDF pdffonts

Extract Text from PDFs

pdf-link-checker pdfbookmark pdfquery pdf-page-counter

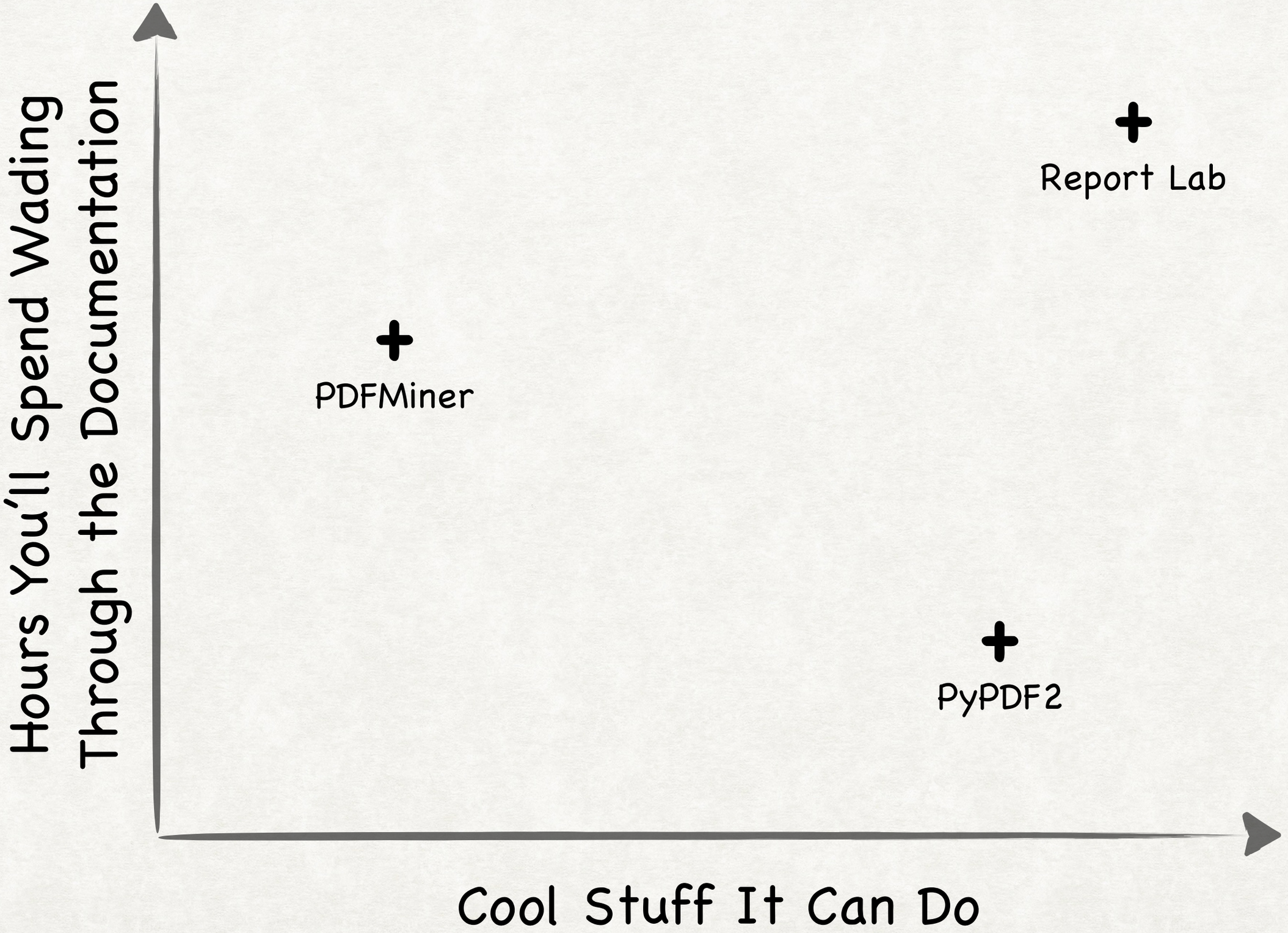
pdfplumber concatPDF geopdf pdf_extractor

*random sampling from https://pypi.python.org/pypi?%3Aaction=search&term=PDF&submit=search

PDFS AND PYTHON

THE OPTIONS...

	PyPDF2	PDFMiner (PDFMiner3k)	ReportLab (open source version)
Creates PDFs	(sort of)		X
Gets text	(mostly)	X	
Assign a password	X		X
Adds bookmarks	X		X
Merges PDFs together	X		(kinda)
Adds objects to a PDF page	(depends)		X


Hours You'll Spend Wading Through the Documentation
Cool Stuff It Can Do
PDFMiner
Report Lab
PyPDF2

MANIPULATING A PDF

with PyPDF2

Three main classes:

- The PdfFileReader Class - reads existing document
  - *decrypt()*
  - *getNumPages()*
  - *getOutlines(node=None, outlines=None)*
  - *documentInfo*
- The PdfFileMerger Class - concatenates, slices, appends
  - *append(fileobj, bookmark=None, pages=None, import_bookmarks=True)*
  - *merge(position, fileobj, bookmark=None, pages=None, import_bookmarks=True)*
- The PdfFileWriter Class - writes out a new pdf
  - *appendPagesFromReader(reader, after_page_append=None)*
  - *addBlankPage(width=None, height=None)*
  - *addBookmark(title, pagenum, parent=NoneFit)*
  - *addLink(pagenum, pagedest, rect)*
  - *encrypt(user_pwd, owner_pwd=None, use_128bit=True)*

MANIPULATING A PDF

with PyPDF2

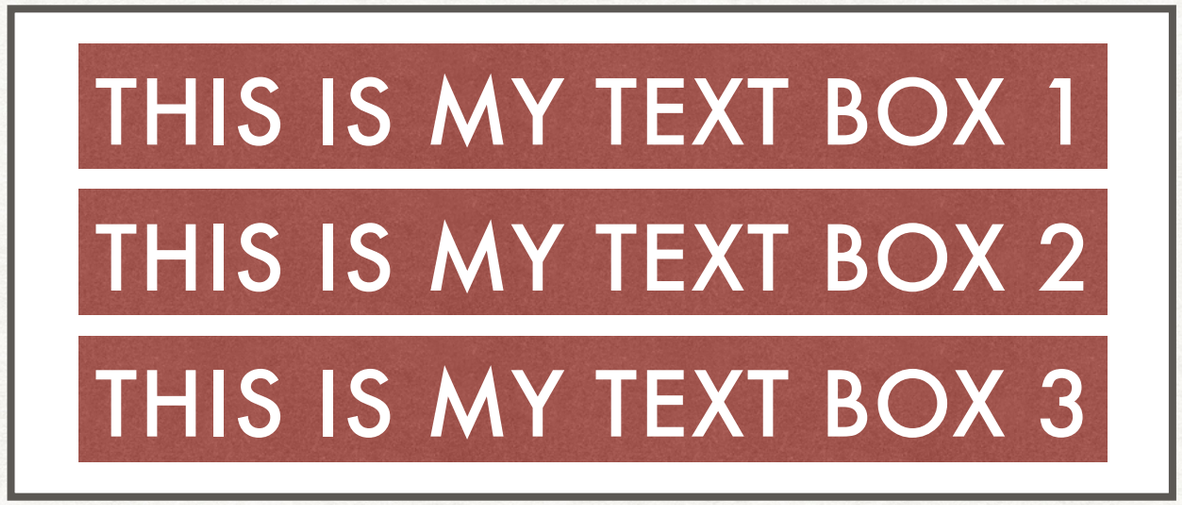
http://localhost:8888/notebooks/Documents/Presentations/SacPy_PDFs_201710/pdf-repo/Combine%20PDFs%20Together.ipynb

EXTRACTING TEXT FROM A PDF

PyPDF2 vs. PDFMiner

http://localhost:8888/notebooks/Documents/Presentations/SacPy_PDFs_201710/pdf-repo/Extracting%20Text%20from%20a%20PDF.ipynb

EXTRACTING TEXT FROM A PDF



PyPDF2

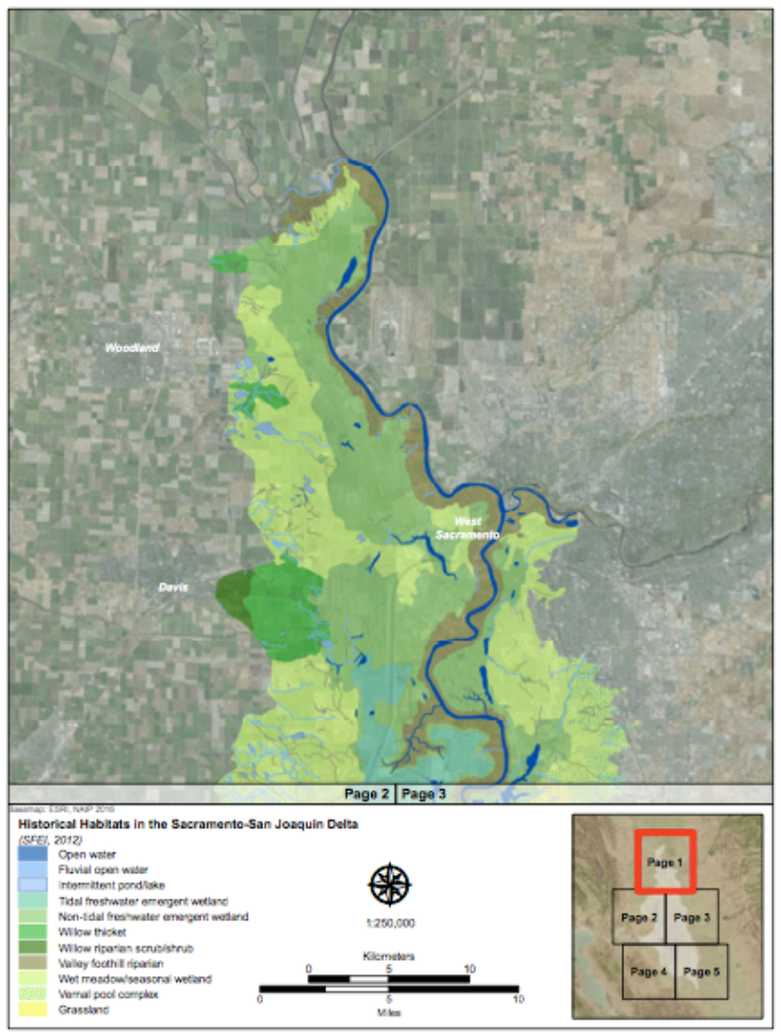
```
"THIS IS MY TEXT BOX 1\n
THIS IS MY TEXT BOX 2\n
THIS IS MY TEXT BOX 3"
```

PdfMiner

```
[["THIS IS MY TEXT BOX 1",
[1,6,8,3]],
["THIS IS MY TEXT BOX 2",
[1,4,8,3]],
["THIS IS MY TEXT BOX 3",
[1,2,8,3]]]
```

REAL WORLD EXAMPLE

USEFUL MATCHLINES IN A MAP BOOK



http://localhost:8888/notebooks/Documents/Presentations/SacPy_PDFs_201710/pdf-repo/Add%20Page%20Name%20Bookmarks%20to%20a%20Map%20Series.ipynb

ArcMap Users beware of...

- *halos*
- *vertical text*
- *text on a curve*
- *and the maplex labeling engine in general*

CREATING A PDF FROM SCRATCH

with ReportLab

http://localhost:8888/notebooks/Documents/Presentations/SacPy_PDFs_201710/pdf-repo/Make%20a%20PDF%20from%20Scratch%20with%20ReportLab.ipynb

My slides and notebooks...

https://github.com/sdearmond/python_and_pdfs

And further reading...

"Manipulating PDFs with Python" by Tim Arnold
https://www.binpress.com/tutorial/manipulating-pdfs-with-python/167

"A Simple Step-by-Step Reportlab Tutorial" by Michael Driscoll
https://www.blog.pythonlibrary.org/2010/03/08/a-simple-step-by-step-reportlab-tutorial/

"Python PDF 2: Writing and Manipulating a PDF with PyPDF2 and ReportLab" by dadruid5
https://dadruid5.com/2014/08/19/python-pdf-2-writing-an-manipulating-a-pdf-with-pypdf2-and-reportlab/

Thank You.

blog:	www.geobug.net
twitter:	@geobugShannon
github:	https://github.com/sdearmond
sac-tech.slack.com:	@geobug

Shannon DeArmond - Far Western Anthropological Research Group, Inc.