

PDF statistics – the universe of electronic documents

Duff Johnson,
Executive Director
PDF Association

Duff Johnson
Executive Director of the PDF Association



Introduction

- PDF is:
 - Recognized worldwide
 - Broadly adopted
 - Accepted everywhere
 - Supported by a rich ecosystem
 - Etc...

- But what do we **know** about PDF adoption, utilization, etc.?



What would it take to compete with PDF?

- Self-contained (portable)
- Capable of being static (at least)
- Create from any source
- Generic (browser) support
- Easy-as-printing to create
- Published specification



Examples of *de facto* standards

“...well-known and illustrative examples of *de facto* standards are... :

- HTML (computer file format) started as "*de facto*" (1993-1995) and became the "*de jure*" standard (1995–present day).
- PDF (computer file format) was first created in 1993 by Adobe.... PDF eventually became the *de facto* standard for printable documents. (...) In 2005, PDF/A became a *de jure* standard as ISO 19005-1:2005. In 2008 Adobe's PDF 1.7 became ISO 32000-1:2008.”

https://en.wikipedia.org/wiki/De_facto_standard



Over time, more people think about PDF more often

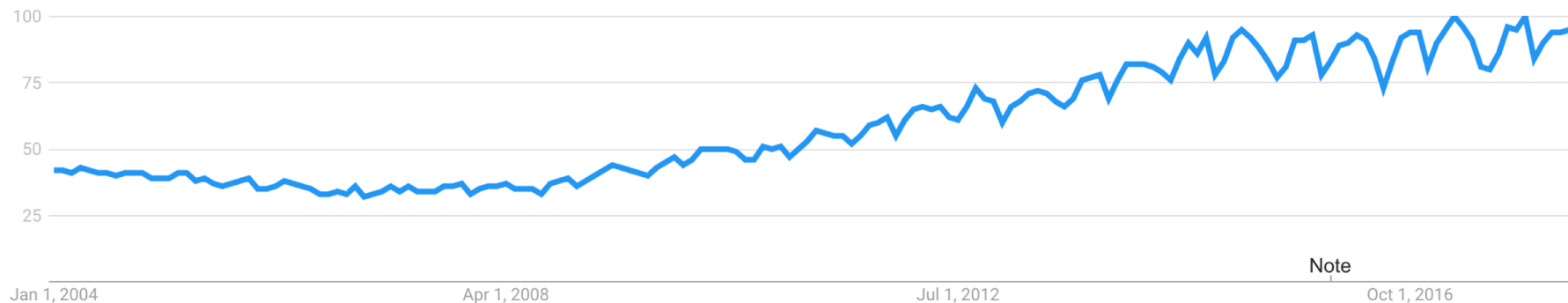
www.pdfa.org



Google Trends



pdf
Search term



Duff Johnson,
Executive Director
PDF Association



...compared to what?

www.pdfa.org



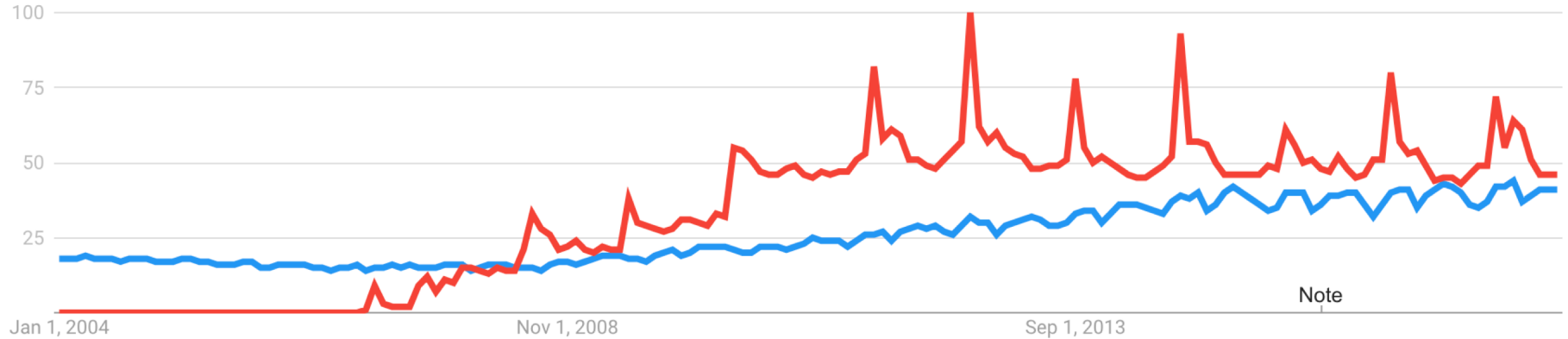
pdf

Search term



iphone

Search term



Duff Johnson,
Executive Director
PDF Association





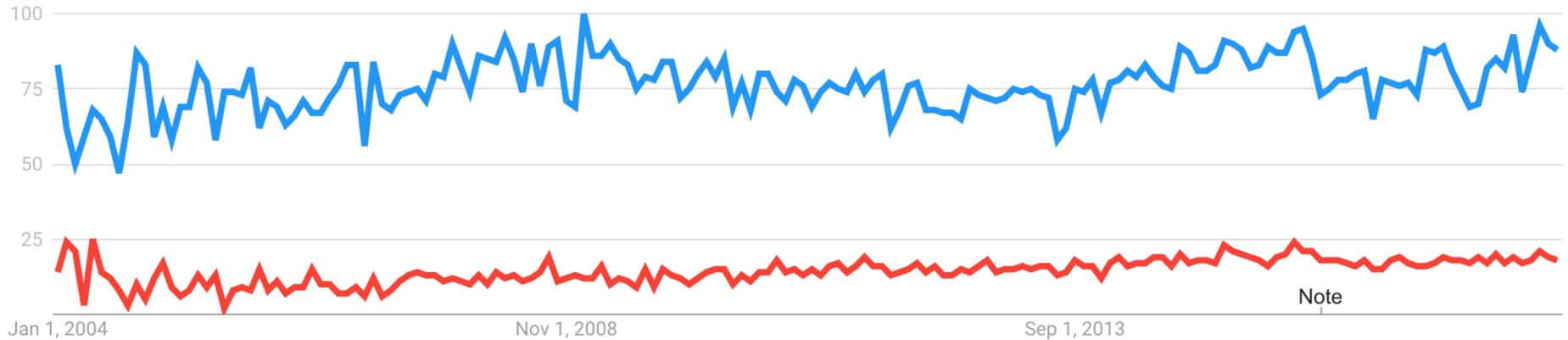
"convert to PDF"

Search term



"convert from PDF"

Search term





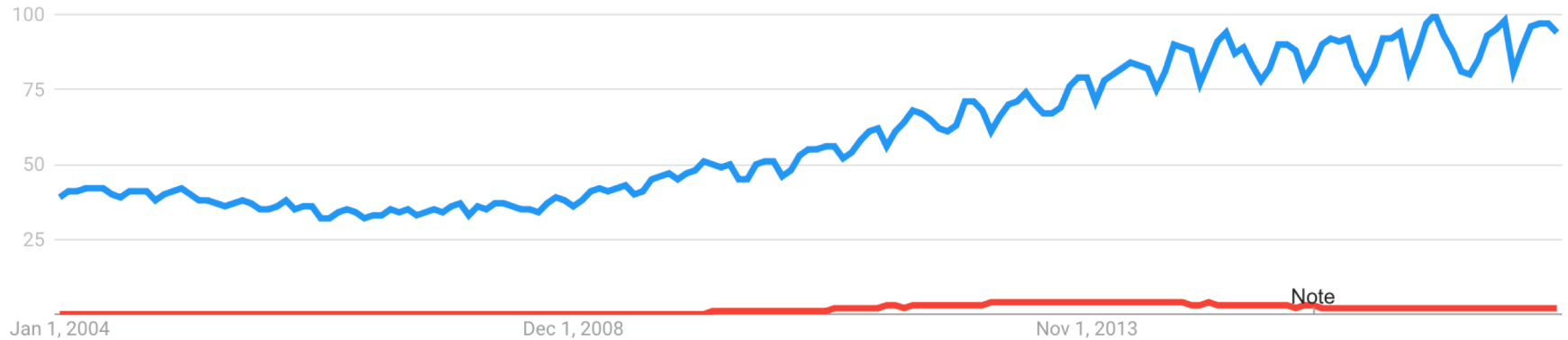
pdf

Search term



epub

Search term





Google Trends



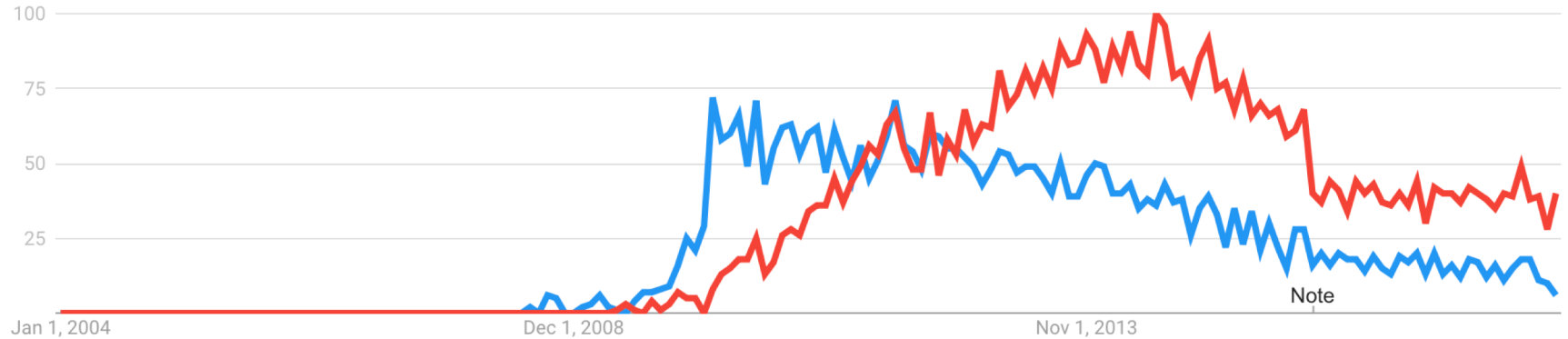
"convert pdf to epub"

Search term



"convert epub to PDF"

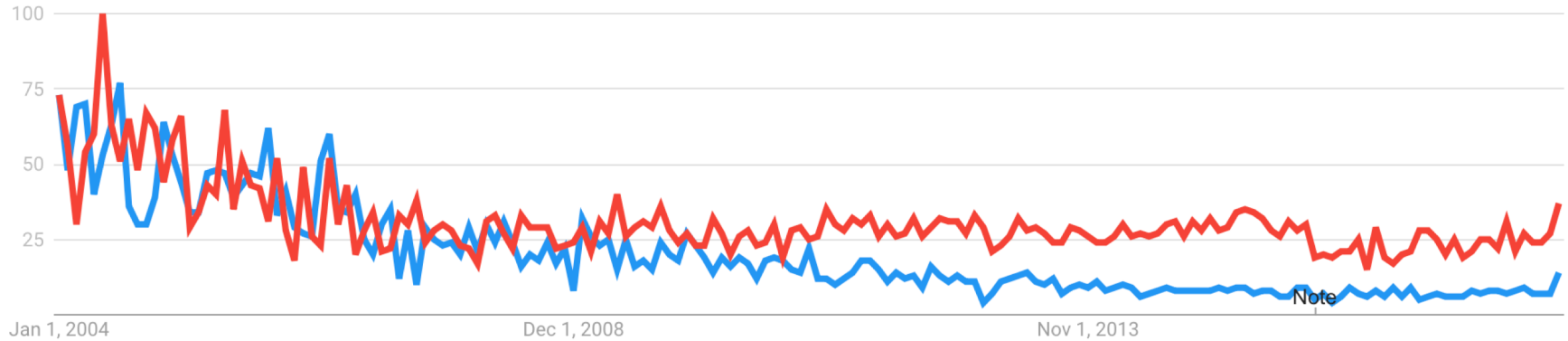
Search term





● "convert pdf to html"
Search term

● "convert html to PDF"
Search term



Conversions are especially popular...



Christiaan Triebert 

@trbrtc

Follow

PDF overtakes Christianity and becomes second most popular religion.



how do i convert to



how do i convert to **islam**

how do i convert to **pdf**

how do i convert to **christianity**

how do i convert to **number in excel**

how do i convert to **mp3 in itunes**



Importance of e-book features to academic librarians

✓
n/a
X
✓
n/a



US 2016

Source: Statista

<https://www.statista.com/statistics/715005/share-of-e-book-feature-importance-according-to-academic-librarians-us/>



Importance of e-book features to academic librarians



Inclusion of color/images/interactive content

77%



Print on demand

76%



Accessibility (text-to-speech, etc.)

75%

n/a

Ability to download articles in multiple formats

75%



Ability for student to annotate/highlight text

73%

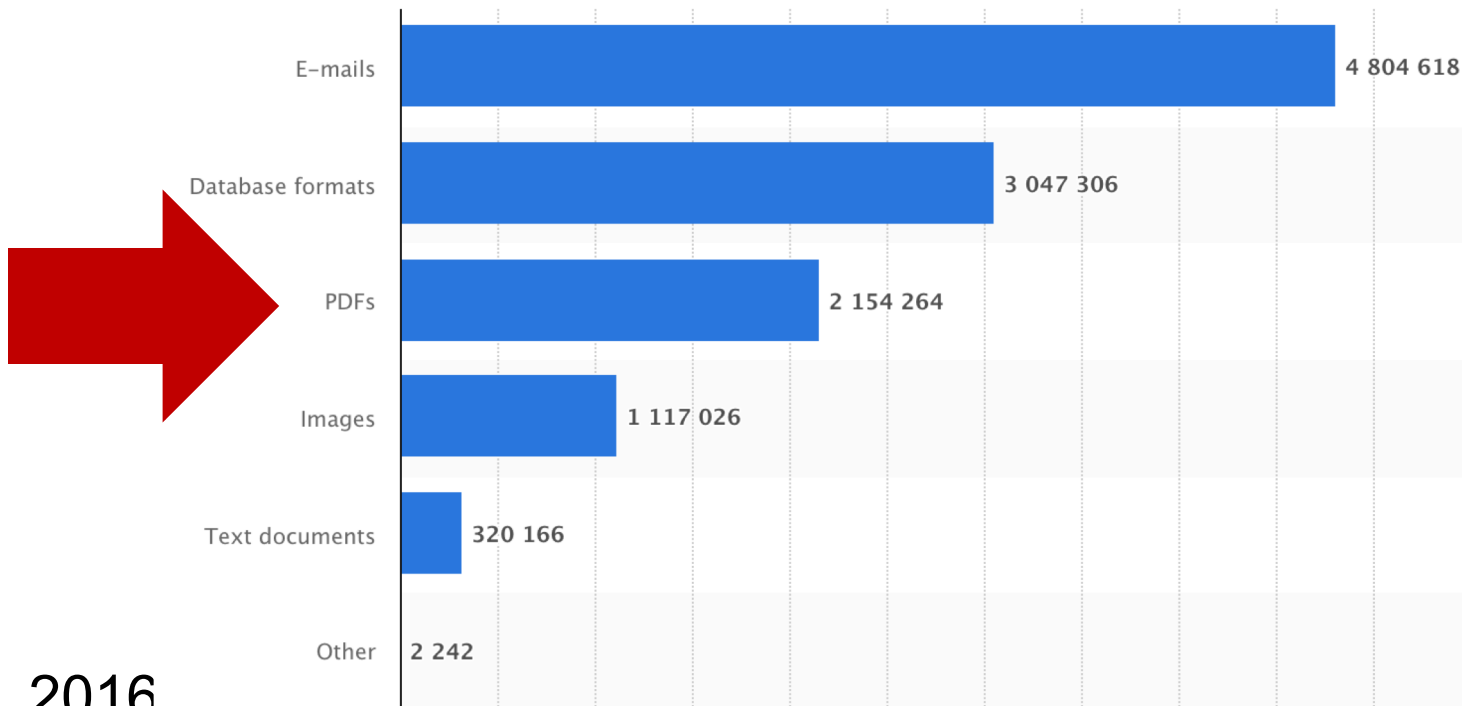
US 2016

Source: Statista

<https://www.statista.com/statistics/715005/share-of-e-book-feature-importance-according-to-academic-librarians-us/>



Type of files revealed in Panama Papers data leak



2016

Duff Johnson,
Executive Director
PDF Association

Source: Statista
<https://www.statista.com/statistics/531286/panama-papers-data-type/>



How many PDF files?



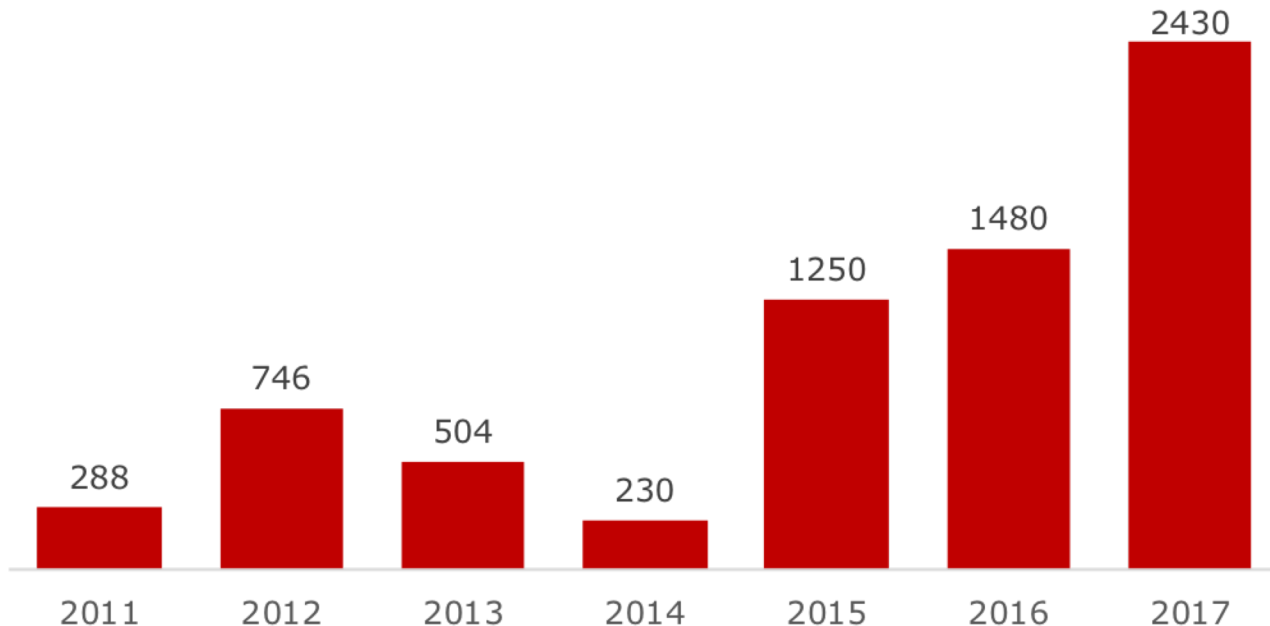
In April, 2016...

- 2.2 billion PDF files on the (public) web (Google)
- 20 billion PDF files in Dropbox
- Airbus, Boeing and the US Dept. of Justice each have over 1 billion PDFs
- 2 billion PDFs opened every year in Outlook.com
- 73 million new PDF files saved every day in Google Drive & Mail
- 60% of non-image attachments are PDFs in Outlook Exchange Enterprise



In April, 2018

In 2018 Google no longer supports pure "filetype:" searches.



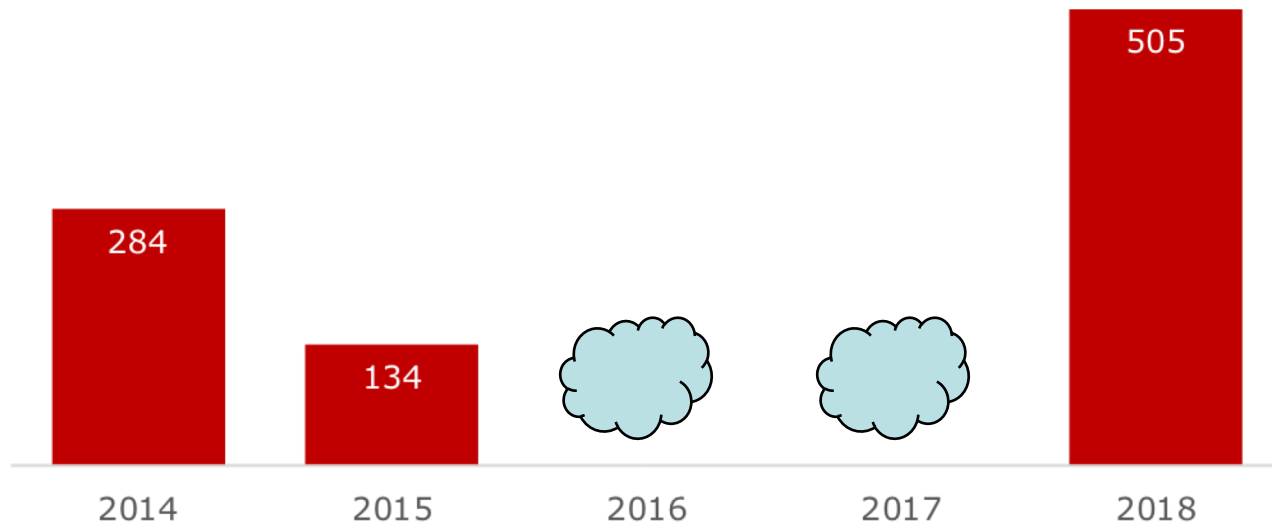
In April, 2018

- Nothing has replaced PDF yet (or even threatened to do so)



In April, 2018

Bing still supports pure "filetype:" searches...



How many PDFs??

- Trillions of PDF files
- Billions more PDF files every day



Some hard data (courtesy of Adobe)

Data was collected anonymously from **desktop** users who agreed to allow their usage to be tracked.



PDF files randomly collected from the web (courtesy of Dual Lab)

These are early results – more to come!

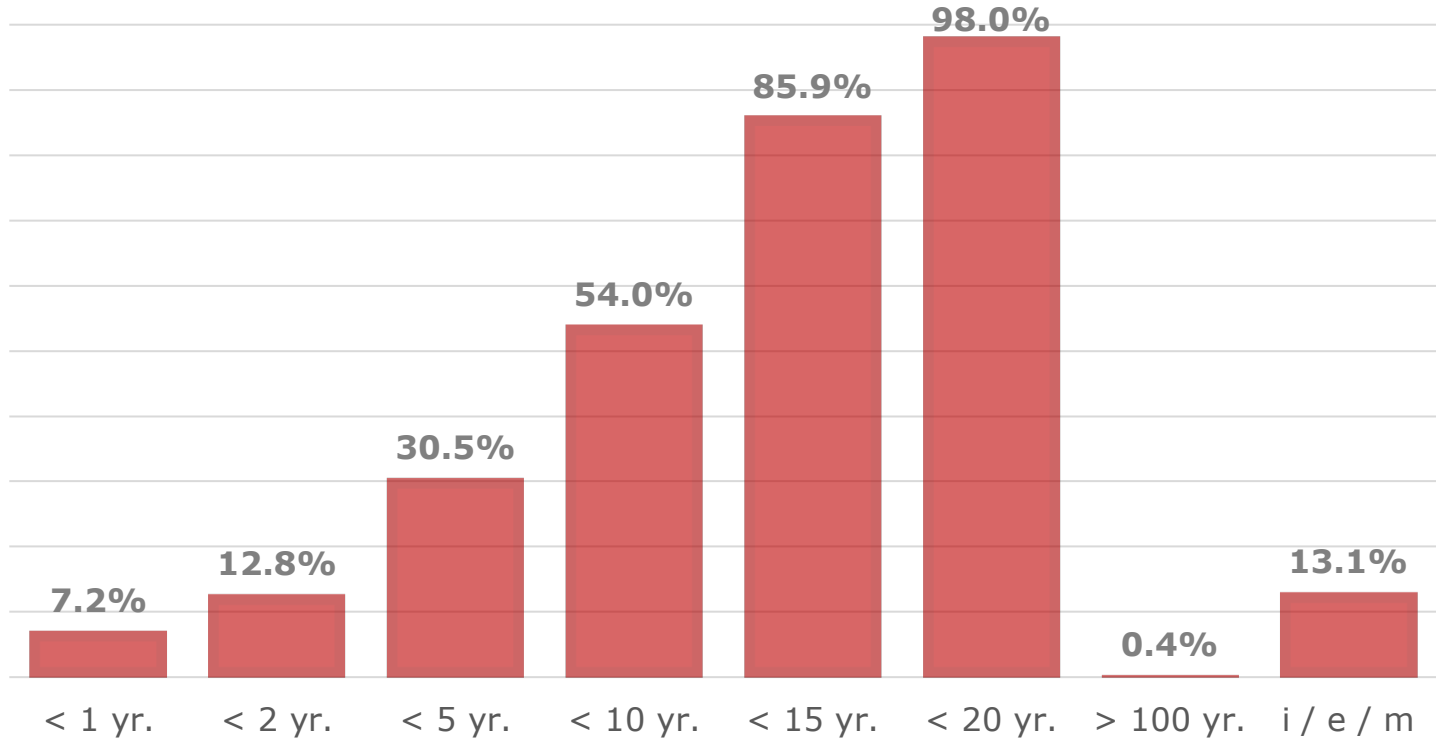


Dataset to-date

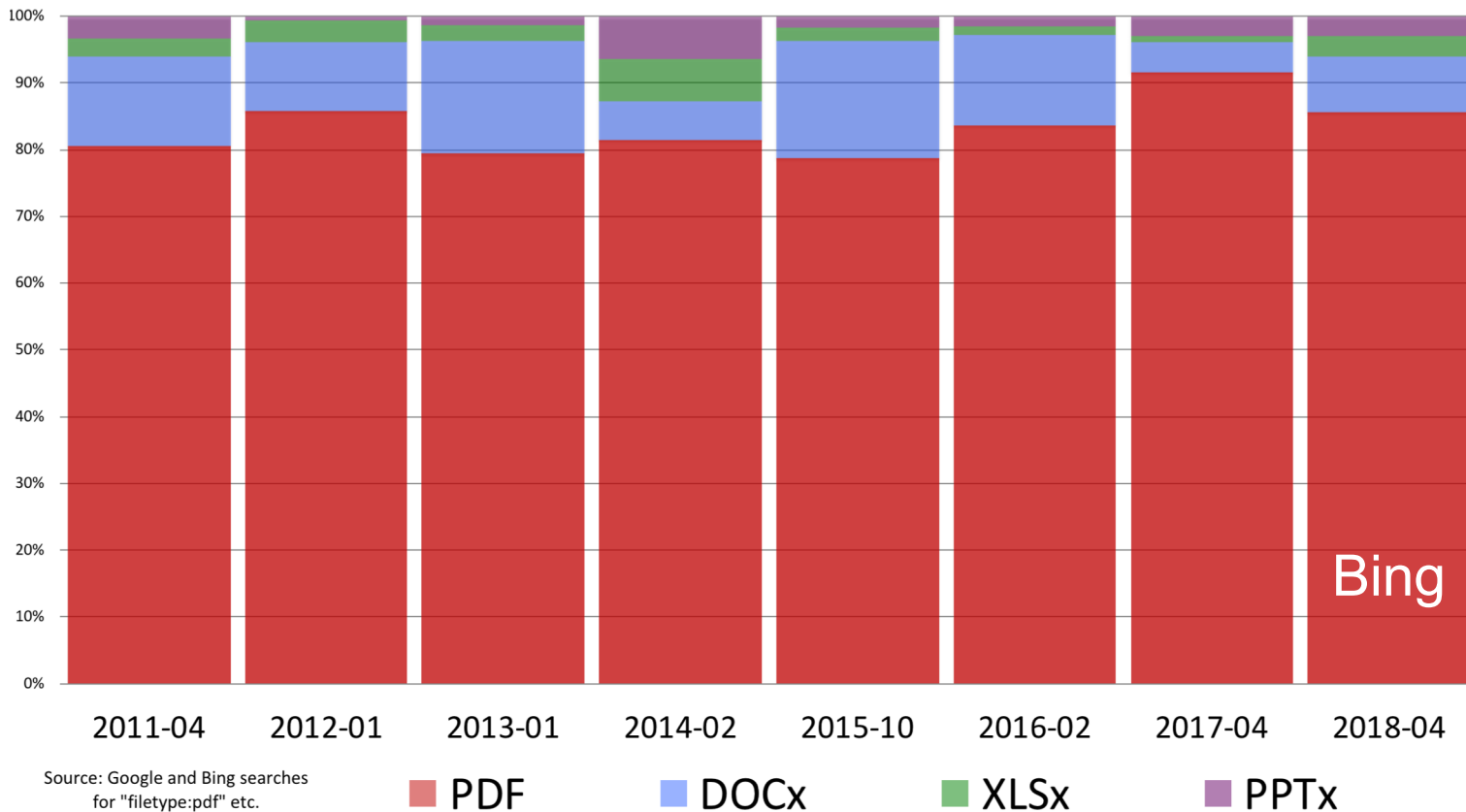
- 548,775 PDF files comprising 14,486,443 pages (26.4 pages/file)
- Method of collection:
 - Chose approx. 2,000-3,000 random words in English, German, French, Spanish, Russian (+ 100 words in Korean)
 - Perform a Bing search for each word + filetype:PDF
Download the first 200 first search results
 - Remove duplicates based on hash



Age of PDF files found online



PDF as a percentage of document formats on the web



2018-April - PDF vs. HTML on the web, according to Bing

- 9,360,000,000 “static” HTML files
Does not include lots of web content!
(Facebook, WordPress archives, etc.)
- 505,000,000 PDFs

1 in 20
static “pages”
are PDF files!



Summary

- There's a lot of PDF documents, and many more everyday
- PDF's niche is pretty large
- PDF has no real competitor in its niche
- PDF continues to gain mindshare (as expressed in searches)
- The industry has a lot of room to grow



Thank you!

Any questions?

