Why PDFs?

Motivating questions

1. What does PDF stand for?
2. How many PDFs are in existence? (rough guess)
3. What do you think of when you think PDF?
    - List the attributes and characteristics of PDFs
4. What do you do with PDFs in your work?

Answers (No Peeking, please!)

What does PDF stand for?

- PDF stands for Portable Document Format
- Adobe video
- file format
- goal: faithfully reproduce documents across HW/SW/OS

Number of PDFs in existence

- Nobody knows for sure
- The most widely cited estimate: 2.5 Trillion
  - Conservative back-of-the-envelope calculation
- Source: Phil Ydens, Fellow and VP @ Adobe
  - 2015 keynote talk, starting at 19:40

Other PDF factoids from Ydens (2015)

- 1.6 billion PDF documents on the web
- 60% of non-image Outlook attachments are PDFs
- Cloud storage providers
  - PDF is the most popular format
  - In many cases, PDF > 50% of files stored
  - 18 billion PDF documents in DropBox
  - 73 million new PDF documents each day in Google Drive and Mail
- 1/3 of PDFs are scanned documents!

Think of PDFs as (Ben's list)

- 21st-century paper
- Used for final versions, statements & legal contracts
- Document fidelity: preserves the original layout, formatting, and fonts
- Cross-platform compatibility, the document looks the same everywhere
- The most common digital format for archived text documents
- Adobe $\rightarrow$ ISO
- Data extraction challenges

Things I do with PDFs (Ben's List)

- Read them, like everybody else!
- Extract metadata
- Extract text
- Identify and redact PII
  - PII = personally identifiable information
- Organize in a database