Role of PDF and Open Data

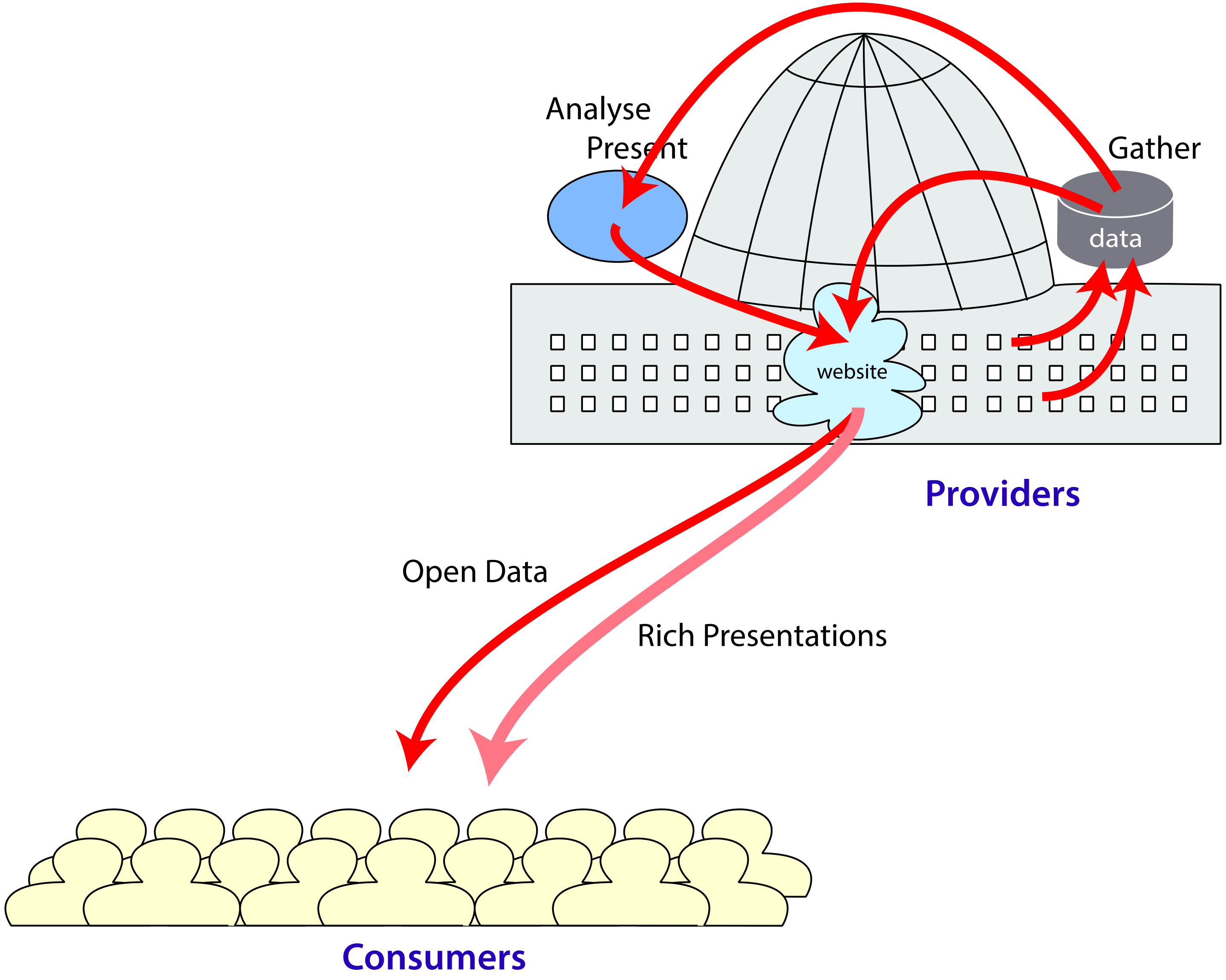
James C. King | Senior Principal Scientist

Outline

- Open Data Paradigm
  - Who is here and why
- PDF
- Role of PDF
  - PDF in the wild
  - PDF purpose-built
    - Structured Data
    - PDF envelopes

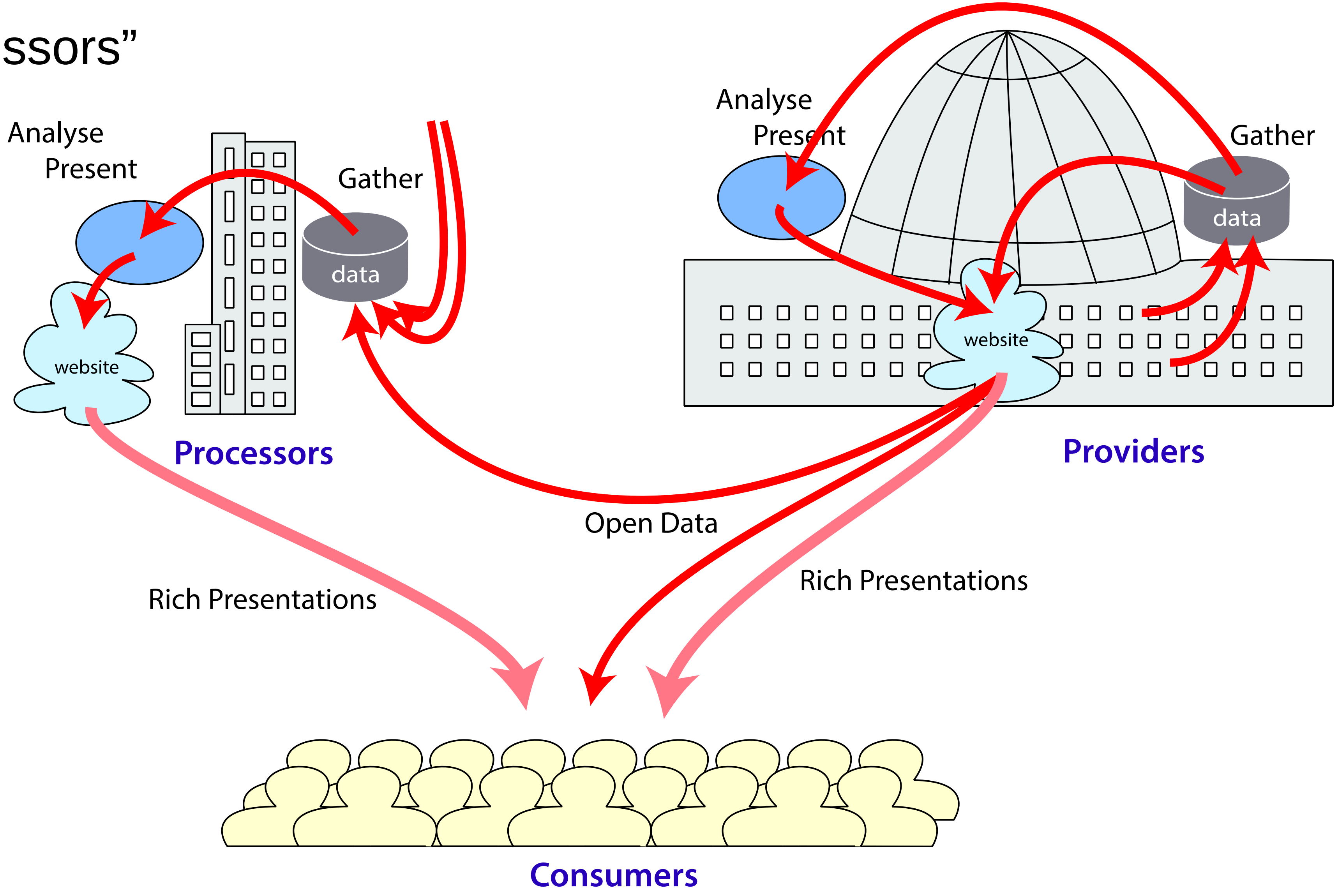
Open Data Paradigm

Providing Open Data



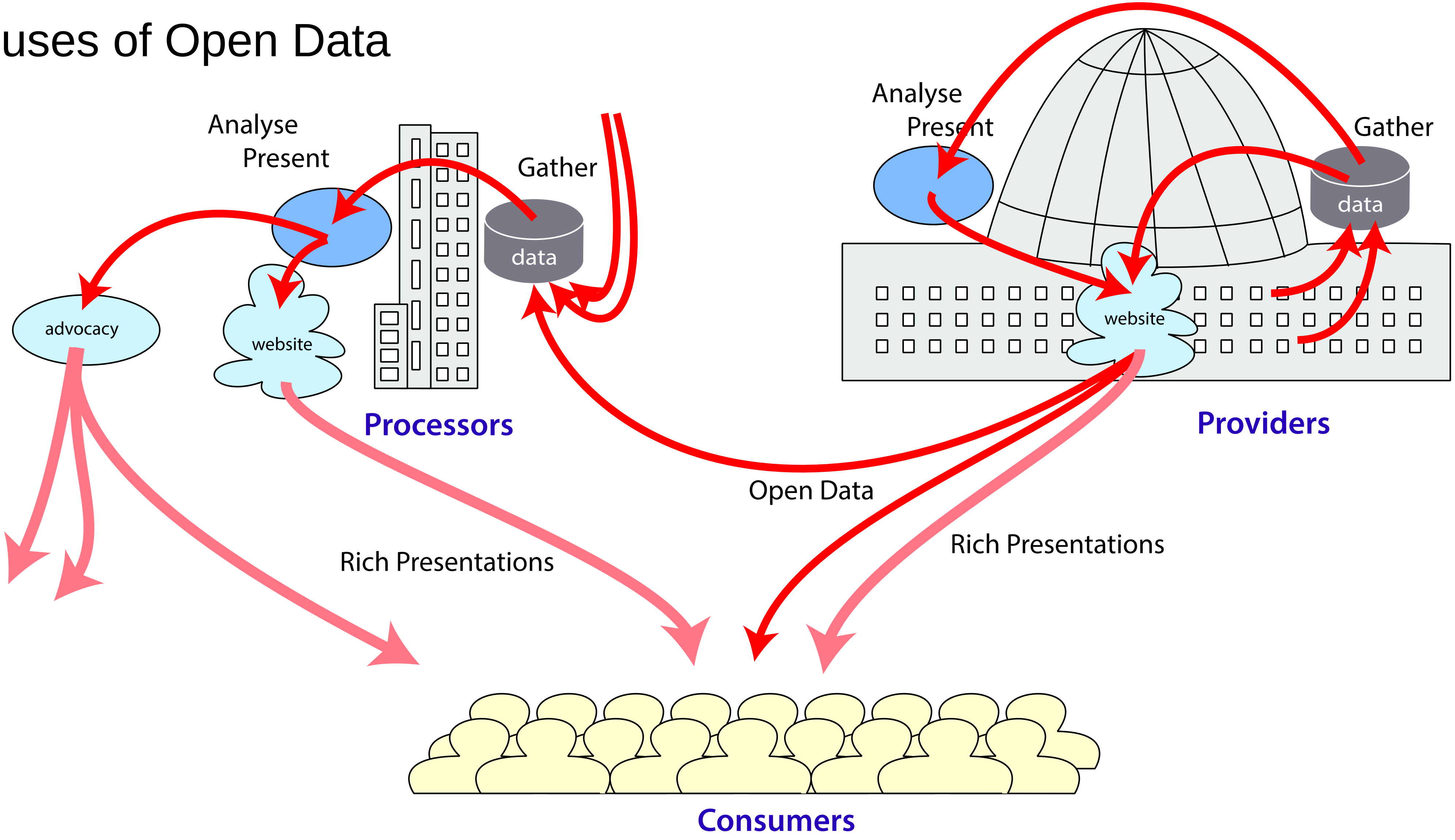
Open Data Paradigm

3rd Party “Processors”



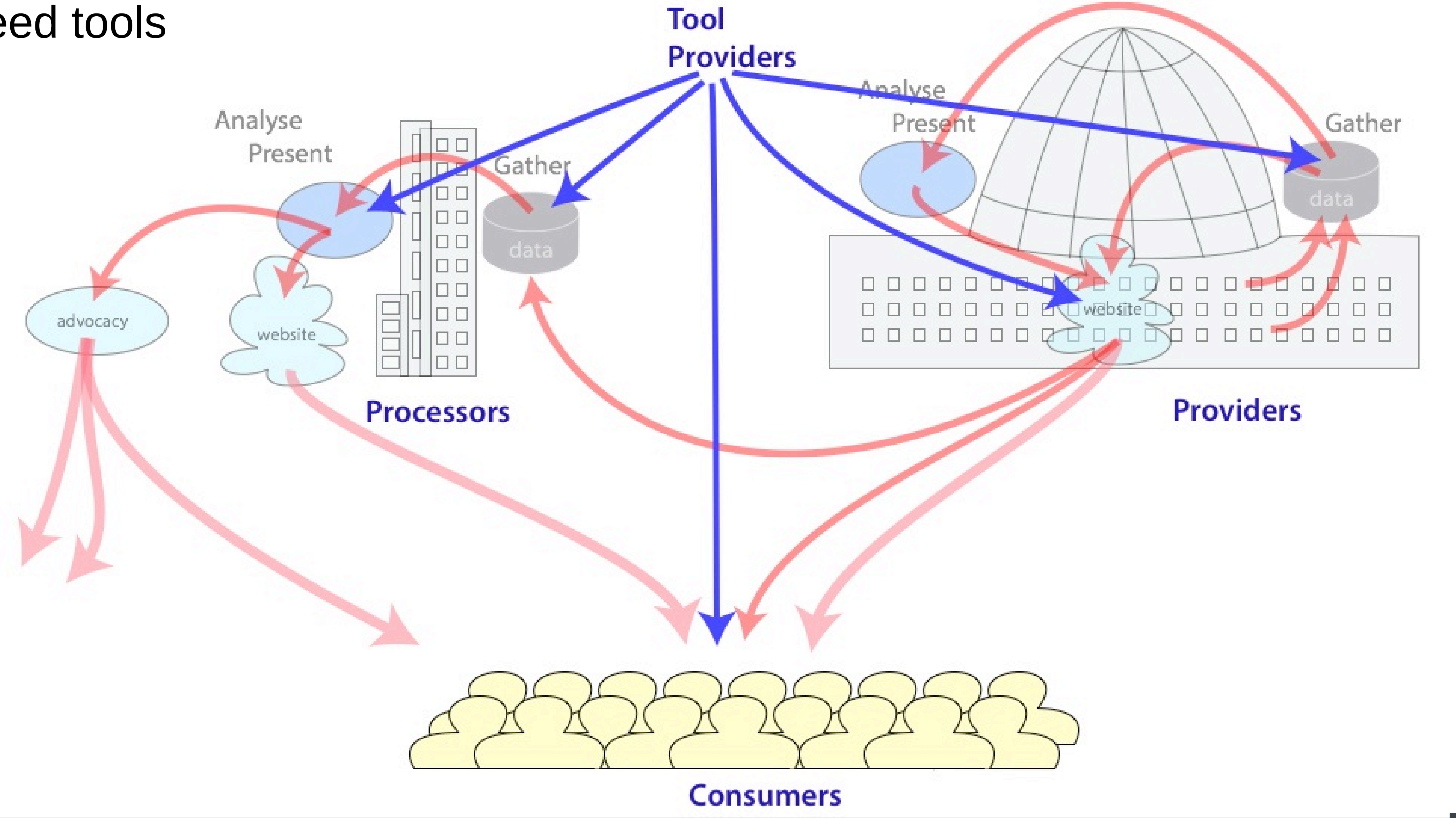
Open Data Paradigm

Other uses of Open Data



Open Data Paradigm

All need tools



Open Data Roles

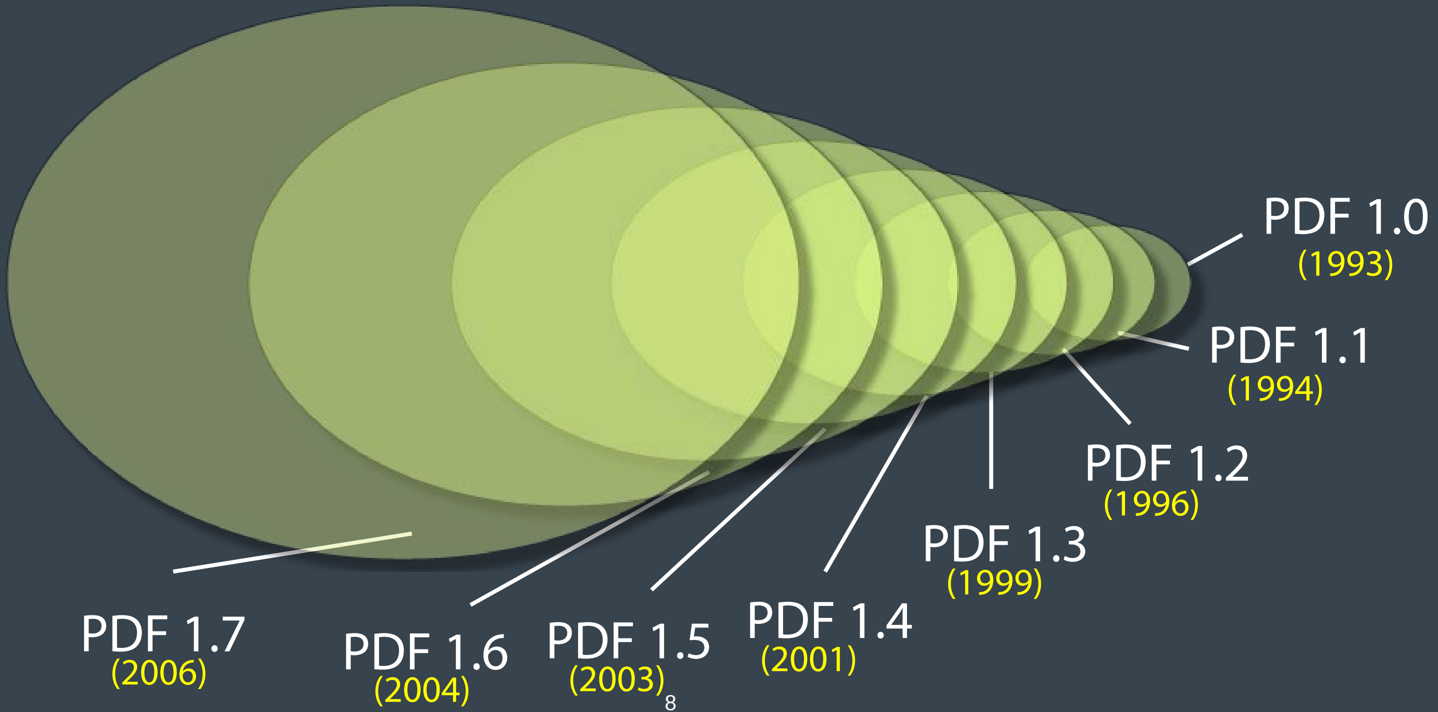
- Which is your role(s)?
  - Providers
  - Consumers
  - Processors
  - Tool Providers

- Did I miss some roles?

PDF

PDF introduced by Adobe in June 1993

PDF 1.7 became an ISO Standard in July 2008



ISO Work on PDF is ongoing

Role of PDF and Open Data

- PDF in the wild
- PDF purpose-built

Pre-existing PDFs (PDF in the wild)

- PDFs abound containing useful content
  - but, PDF is a document format not a data format

- If pages contain graphics – can extract those graphics
  - Vector graphics: use Adobe Illustrator
  - Images: see http://blogs.adobe.com/vikrant/2010/12/extract-images-from-a-pdf/

- If pages are textual (including tables) – can extract that text/tables
  - see, Wikipedia entry for “List of PDF Software”

- If pages are images – must turn to OCR technology
  - see, Wikipedia entry for “Comparison of optical character recognition software”

Purpose Built PDFs – Structured PDFs

- ISO Standard allows for optional structural information to be added to PDFs for
  - reading order
  - tagging information (headings, footnotes, figures, math)

- Content extraction tools can make use of this structure while extracting content

- Structure best obtained from authoring tool (e.g., document processing tools)

- Can be added after-the-fact

Purpose Built PDFs – PDF Attachments

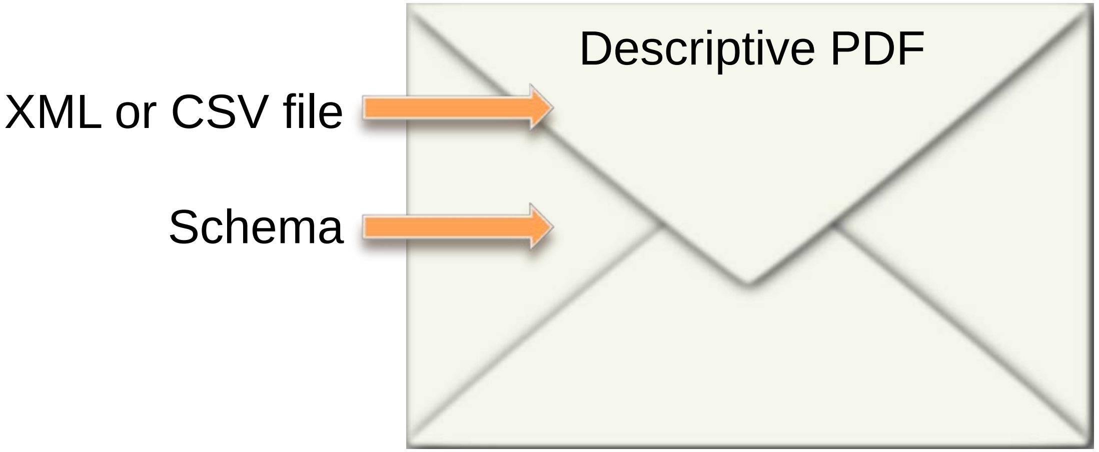
- ISO Standard defines attachments to PDF files
  - attachments get compressed using same lossless technology as ZIP and PNG

- Attach icon to page to select attachment

- Here is a sample of using attachments for datasets used in a presentation

PDF Enveloping

- Raw data needs defining information
  1. documentation for source, ownership, semantics
  2. schema for syntax
  3. proof of authenticity



- PDF can provide 1. and include data and schema as attachment
  - typical XML file gets reduced by an order of magnitude
  - PDF document features cover the attachments (authenticity, signatures, forms)

- Attachments easily extracted from mother PDF

- see an example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf

References to more about PDF

- PDF attachment example: http://www.w3.org/2013/04/odw/EducationalAttainment.pdf
  - Derived from http://www.census.gov/hhes/socdemo/education/data/cps/historical/index.html
  - Using the Acrobat web capture feature to convert HTML to PDF (14 pages)
  - All of the 8 dataset files were downloaded and added to this PDF as attachments

- PDF package example: http://blogs.adobe.com/insidepdf/files/2010/11/LeadershipPacs_2010.pdf

- My PDF blog: http://blogs.adobe.com/insidepdf
- My tutorial on what is inside of a PDF file:

  http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_Look_Inside.pdf
- Other presentations and papers by me:

  http://www.adobe.com/technology/people/san-jose/jim-king.htm

Adobe