Information Governance and Artificial Intelligence: Possibilities and Realities

Dr Anthea Seles, MAS, PhD
Project Delivery Manager
Artefactual Systems

31 August 2023

Overview

- Definitions
- Background
- AI and Information Governance
  - Initial Assessment
  - Funnel Approach
  - Duplication and ROT Analysis
  - Machine-Learning In Practice
  - Searching, Classification and Categorization
  - Human Review
  - Understanding Emails
- Machine-Learning Vs Human Review
- Conclusion

Definitions

DATA:

- Structured data: Information, more often numerical information, put in tabular form to enable quantitative analysis.
- Unstructured data: Information consisting of word processing documents, power point presentations, videos, sound records, photographs etc.

ENVIRONNEMENT

- Structured record-keeping environments: Environments where documents and data are placed in an ordered fashion to allow for retrieval. Ex: Information management system or shared drives with a unified classification scheme.
- Non-structured record-keeping environments: An environment where documents and information are not organized and can be comprised of a running sequence of documents or a shared drive with no unified classification scheme.

Definitions

What is Artificial Intelligence? The theory and development of computer systems able to perform tasks normally requiring human intelligence (Oxford online dictionary)

Types of AI

- Supervised
- Unsupervised

Specific Applications

- Reinforcement
- Neural Networks
- Deep Learning

Background

- Information management systems are not always easy to use meaning that users try and find other, easier ways, to file their information.
- They use shared drives in parallel with information management systems, resulting in incomplete folders and duplication
- In the UK, we carried out a study to assess the state of record-keeping in government departments and understand the amount of ‘legacy data’ they held. The Digital Landscape in Government 2014-2015
  - **1 TB: ~25 TB**
  - **1.5 PB = approximately 1.5 billion Word documents (unstructured records, emails and datasets)**
  - Once we accounted for the totality of the information holdings which includes email servers and data sets it added up to over 1.5 petabytes of data that needed to be appraised and selected
  - Information management teams did not know what was contained in legacy data holdings and did not know what documents or data needed to be preserved
  - This information could also have differing levels of contextual information and limited metadata. Metadata could also be compromised because of previous migrations.

Information Governance

- Having an understanding and control over the information an organization creates is essential. For corporate governance, decision-making, risk management etc.
- This is complicated due to:
  - Volume
  - Shared drives, email servers and datasets
  - Corporate/Institutional Memory Loss
- Automation is no longer a choice, especially in unstructured record-keeping environments
- *The Application of Technology Assisted Review to Born-Digital Records Transfers, Inquiries and Beyond*.

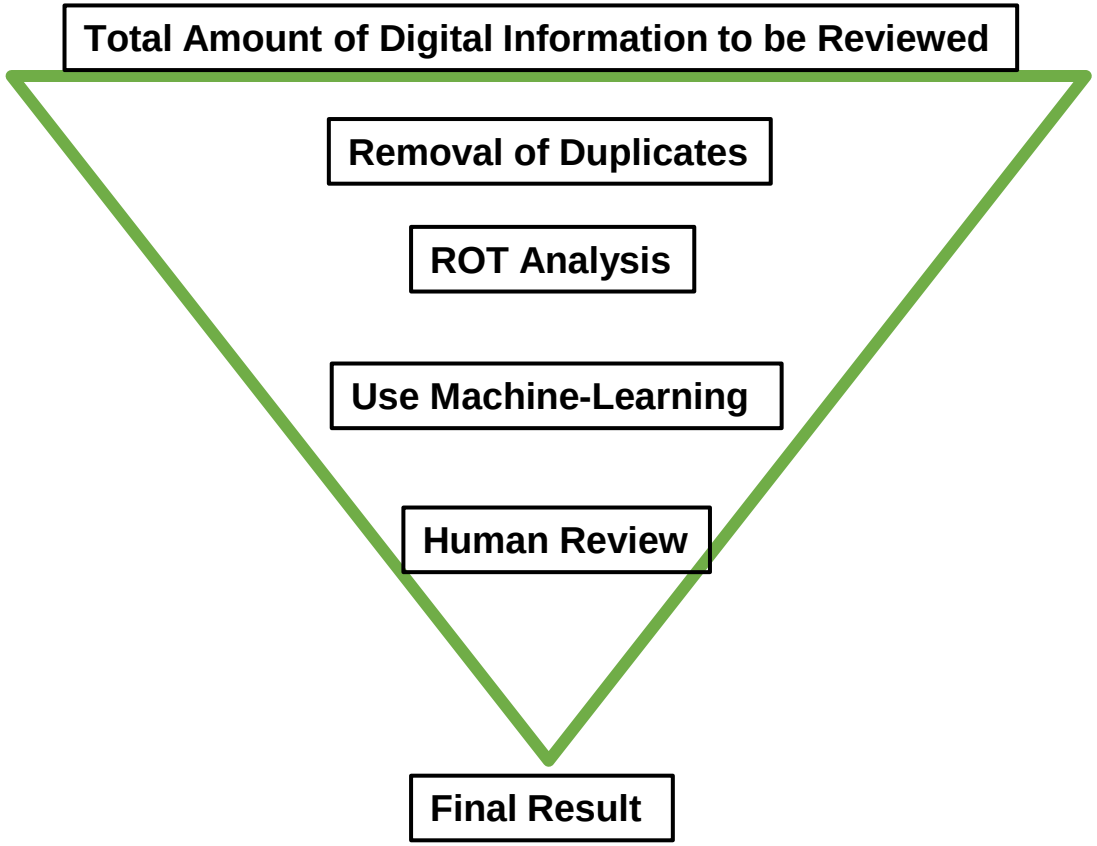
So what do we do?

Initial Assessment

- Unstructured record-keeping environments are unknown territory
- Collect all the data you can on record-keeping practices, migrations and if you have someone in the organization that was there when this records store was being used speak to them/interview them
- This assessment will inform the application of machine-learning
- Potential results from the initial assessment:
  - Metadata has been altered due to previous migrations or other actions
  - Major events
  - Significant dates

Funnel Approach

- Funnel approach: Progressively reducing the amount of digital information that will potentially need manual reviewed

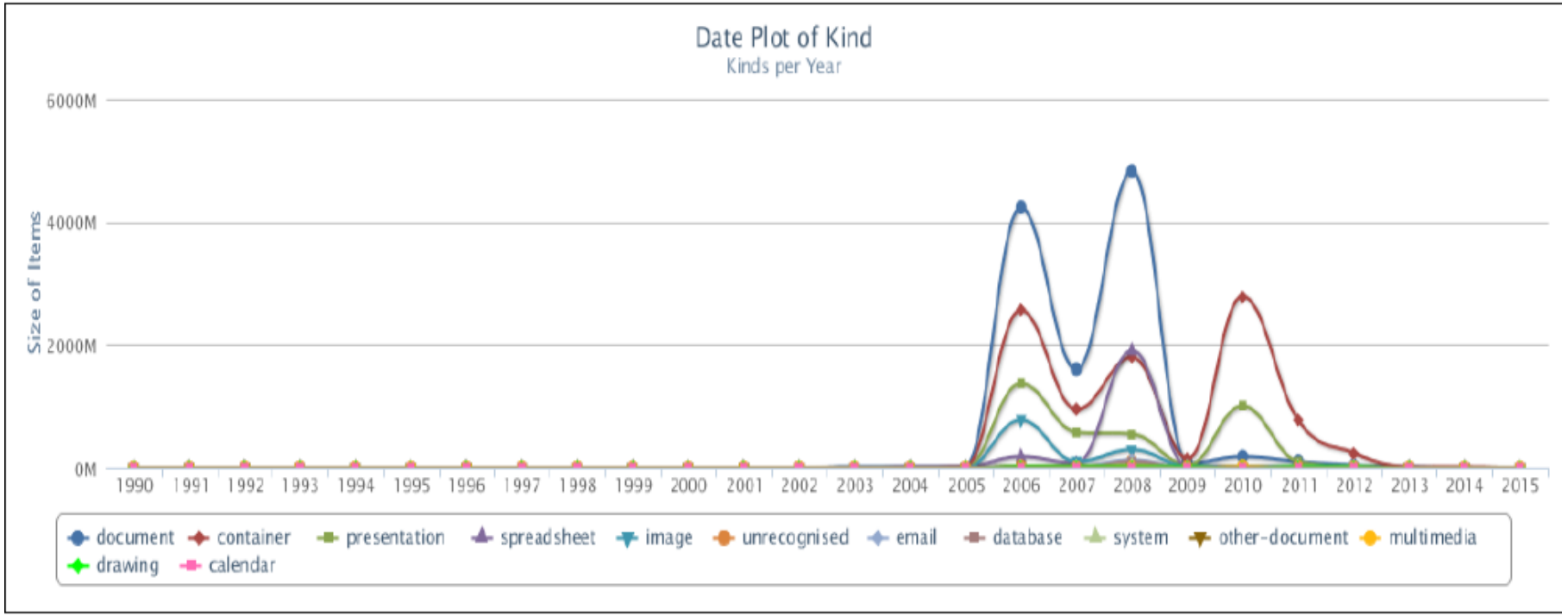


Duplicates and ROT Analysis

- Redundant, Obsolete and Trivial (ROT) Analysis
  - Date
  - Format Type
  - Size
- 50% of unstructured digital recordkeeping environment will be duplicates
  - Meaningful duplicates
  - Unmeaningful duplicates

Machine Learning In Practice

- After carrying out the initial assessment and then the ROT, this would be the time to begin using the features of a data analytics/machine-learning tool
- Carry out a very high-level analysis to see what is left to assess. See an example below:



Representation of a digital records collection by date and format

Machine Learning- Searching, Classification and Categorization

Searching

- Most machine-learning software will support Boolean searching, keyword searching, fuzzy searching and sometimes Natural Language Processing
- These would have been identified during the initial assessment

Classifying and Categorizing

- **Classifying:** Classification is the organization of data into mutually exclusive categories where a one to one relationship exists between the information and the category they are sorted into. ex: sensitivity (high, medium, low)

- **Categorizing:** Categorization is the organization of data but the data can belong to more than one category ex: subject or keyword classification

Think about patterns! Trends in the data.

Machine Learning- Searching, Classification and Categorization

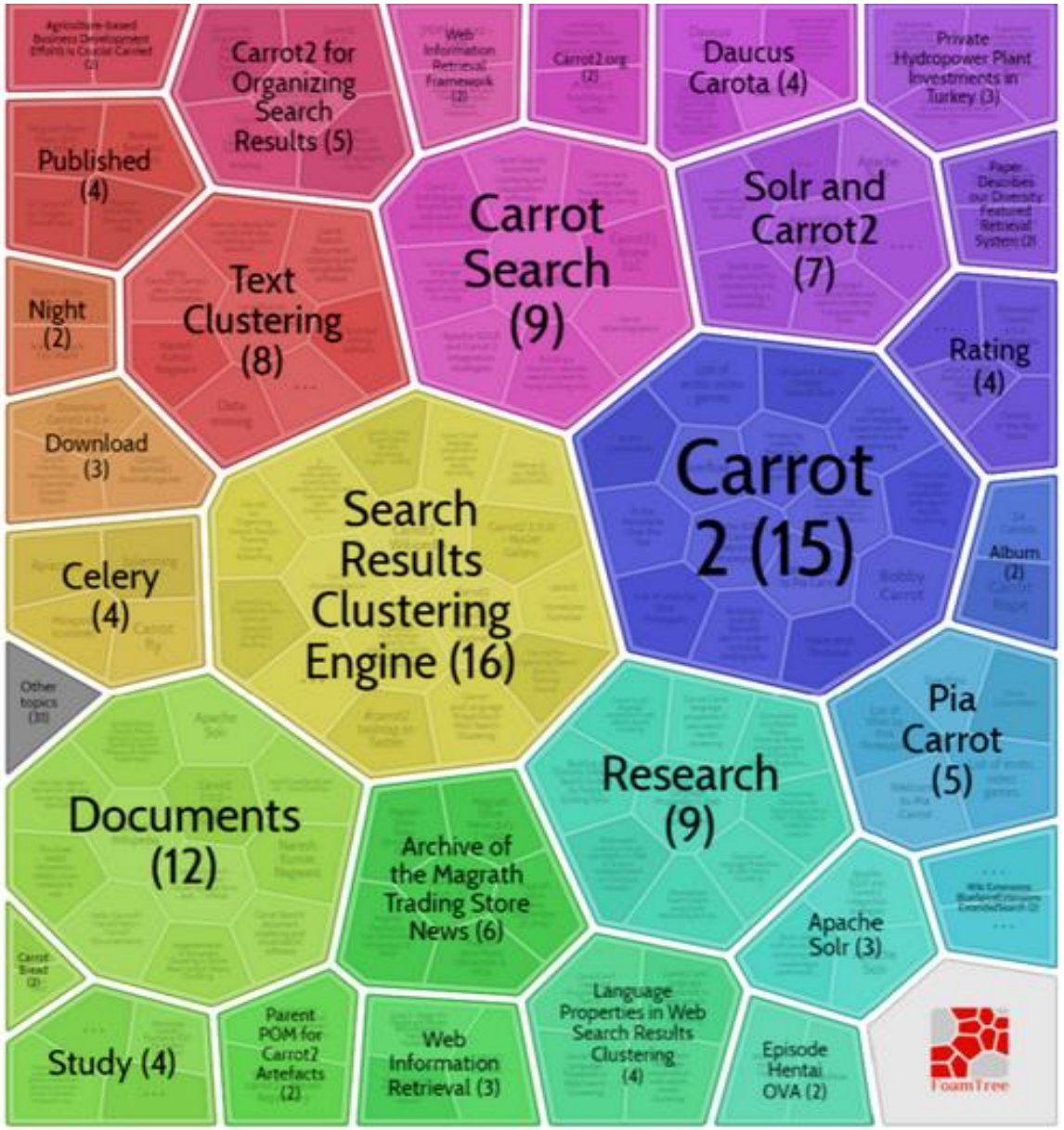


Image is taken from Wikipedia describing Carrot² which is an open source search results clustering engine. The visualization was created by clustering web search results using Carrot² It was created by Dawidweiss (4 January 2021) See: https://en.wikipedia.org/wiki/Carrot2

Machine Learning- Categorization

- When looking at this or any other visualization it worth considering the following questions:

  - How did the machine produce this visualization? What do certain categories mean? How were these categories created?

  - When looking specifically at classification or categorization what base training data was used to help the machine carry out this work? Could it influence the outcome of what I am seeing? If so, how? Was the training data biased?

  - If there are issues with the categories in the data visualization how will it change any methods or approaches I use in carrying out digital preservation?

Human Review

- We need to remember that these systems are tools, human play an important role throughout the process
- There are still decisions that only humans can make.
- Machine-learning is not sophisticated enough to do complex multi-variable assessments outside the data it is provided. At least, not yet.
- Therefore there may still be a small amount of materials that may need to be reviewed by humans
- But we are trying to limit the volume to be reviewed
    - Tiring
    - Ergonomics

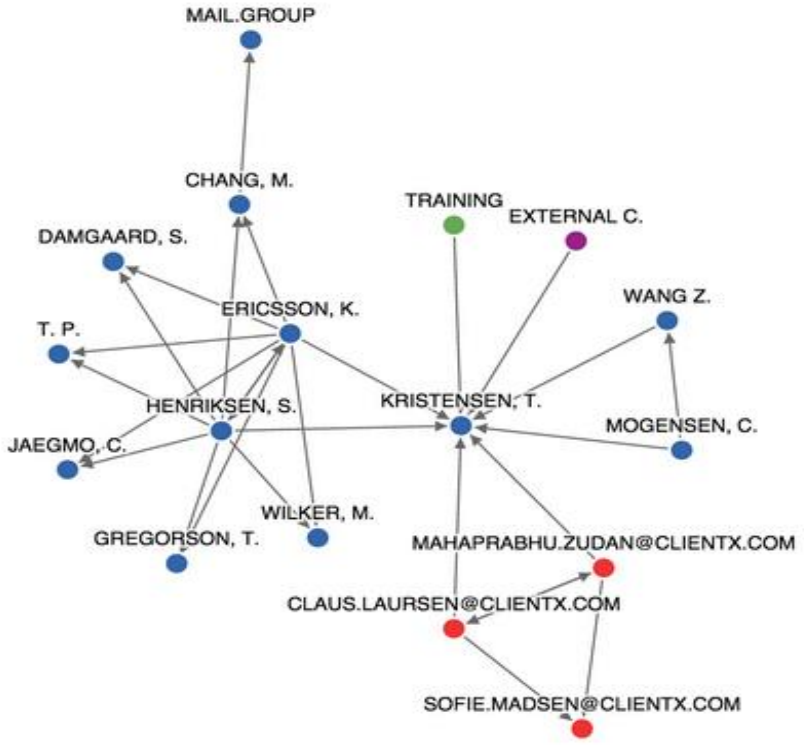
Human Review

Nielsen. J., (2006) 'F-Shaped Pattern For Reading Web Content' https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

- What Nielsen found was that the more information that people had to review, the less attention they began to pay to the details of the material they were reading and they began to skim
- This can lead to oversights in decisions because information has been missed
- This is why we apply the funnel approach

Reviewing Email Collections

- Email review can be complex, but data analytics/machine-learning tools can help
- They can thread emails so you can see how an email chain has evolved
- They can also create network visualization based on email exchanges



Reviewing Email Collections

- There have been some studies looking at how to manage and preserve email such as The Future of Email (August 2018) and there was a pilot project at the University of Illinois at Urbana-Champaign that was looking to test the recommendations of this report.

- ePadd: Open source tool designed and tested a Standford University for email appraisal, selection and preservation.

- Capstone Approach

Strengths and Weaknesses

What can machines do well?

- Boolean and keyword searches
- Regular expressions
- Process at scale

What can humans do well?

- Understand and infer context
- Handwriting analysis

Documenting Decisions

- Information managers and archivists have to be accountable for the decisions made using machine-learning
- The entire process should be documented from initial assessment to final disposal decisions
- Documenting: searches, classification, categorization, deduplication and ROT analysis results
- List all digital information slated for destruction and what will be kept
- **It won’t be perfect, but nothing is! Machine-learning is NOT a silver bullet. Manage your risks**

Conclusion and Considerations

- Automation is not an option
- AI is not a silver bullet. We need to engage intelligently and pragmatically with the technology
- It wasn’t perfect with paper, it won’t be perfect with AI
- Do some legwork and find the patterns in the data
- It’s about balancing the human and the machine

Thank you

Questions?