



Government
Digital & Data

Managing the digital heap - Intuitions for data scientists

James Lappin

Glow workshop, Manchester

July 2025

About these intuitions

Professional intuitions have many uses. They can help explain complex concepts in a simple way. They can help to generate options for interventions, and they can help when comparing those options.

Intuitions are subjective. Every professional will have their own set of intuitions.

The eleven intuitions set out in this presentation are my best hunches and rules of thumb - I hope you find them useful and I hope it encourages you to articulate your own set of intuitions.

The background is a solid yellow color. In the top-left corner, there is a green L-shaped geometric element. In the top-right corner, there is a pink L-shaped geometric element. Both the green and pink elements have rounded corners and a stepped, architectural appearance.

Data science intuitions

Data science intuitions

- Statistical models can assign any item (or grouping) to a category (on the basis of their features) with a given degree of confidence
- Statistical models can group together items (or grouping) to the most similar other items/groupings on the basis of their features
- Statistical models can assess whether an item/grouping is relevant to a particular topic on the basis of the presence or absence of words that are commonly used within discourse about that topic
- Connections can be made between individuals, the people they communicate with, the items they create, and the topics they discuss
- The names of entities (people, places, organisations) mentioned in an item can be extracted from content and this gives clues as to the content and sensitivities of an item
- Statistical models can understand the meaning of words in a particular context and can respond to questions about content or a given collection of documents

Data science techniques

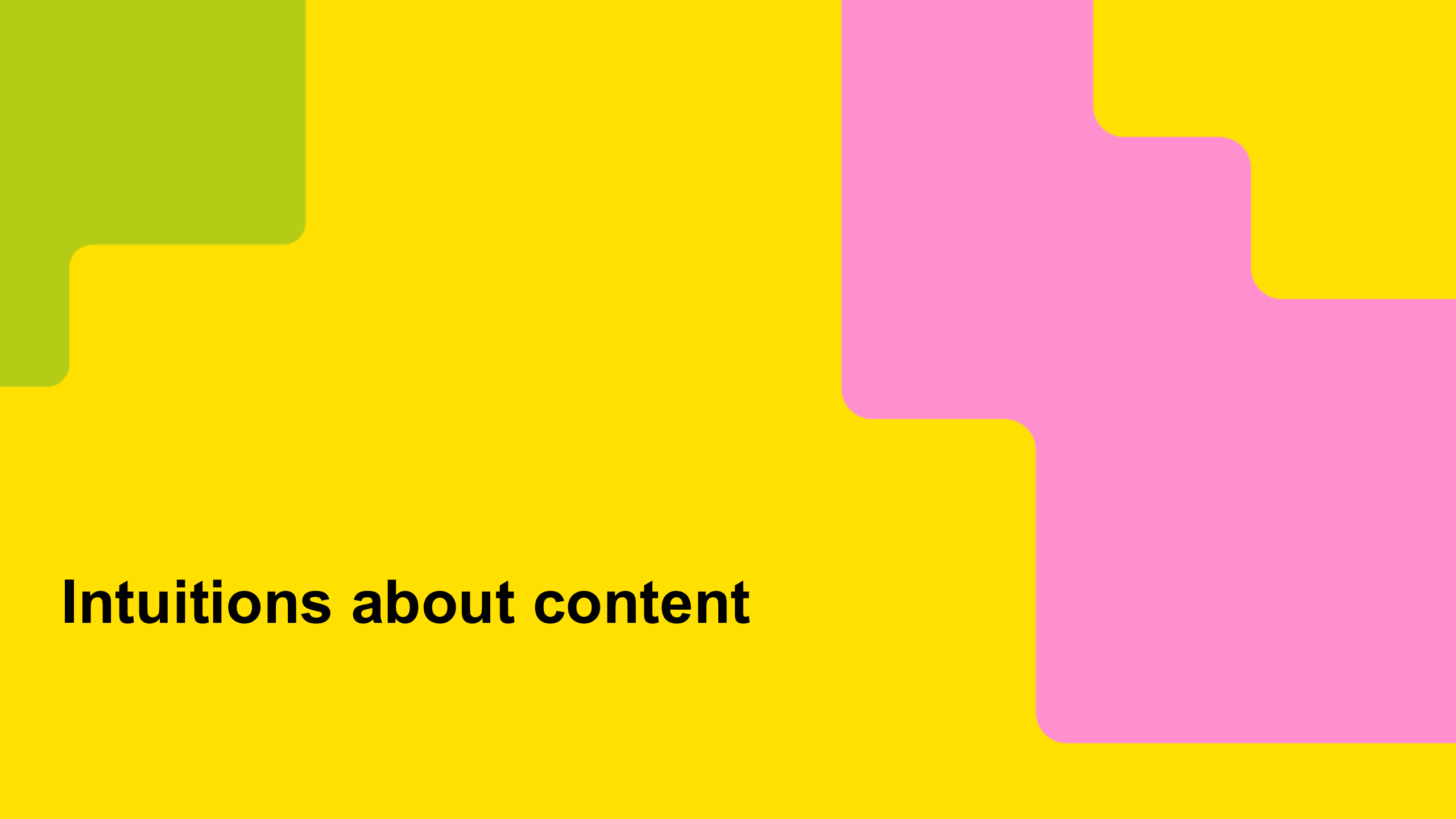
- Classification (logistic regression)
- Clustering
- Topic modelling
- Network analysis and graph building
- Entity extraction
- Large language models

The background is a solid yellow color. In the top-left corner, there is a green L-shaped geometric shape. In the top-right corner, there is a pink L-shaped geometric shape. The text "Archival science intuitions" is positioned in the lower-left area of the slide.

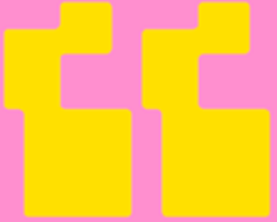
Archival science intuitions

Archival science intuitions

- Records arise when people conduct activities
- The retention rule that should be applied to content depends upon the nature of the work from which that content arose
- Preserve the context as well as the content
- Respect the original order in which content was created
- Act at the highest practical level of content grouping

The background is a solid yellow color. In the top-left corner, there is a green shape with a rounded bottom-right corner. In the top-right corner, there is a pink shape with a rounded bottom-left corner. The text "Intuitions about content" is located in the lower-left area of the slide.

Intuitions about content

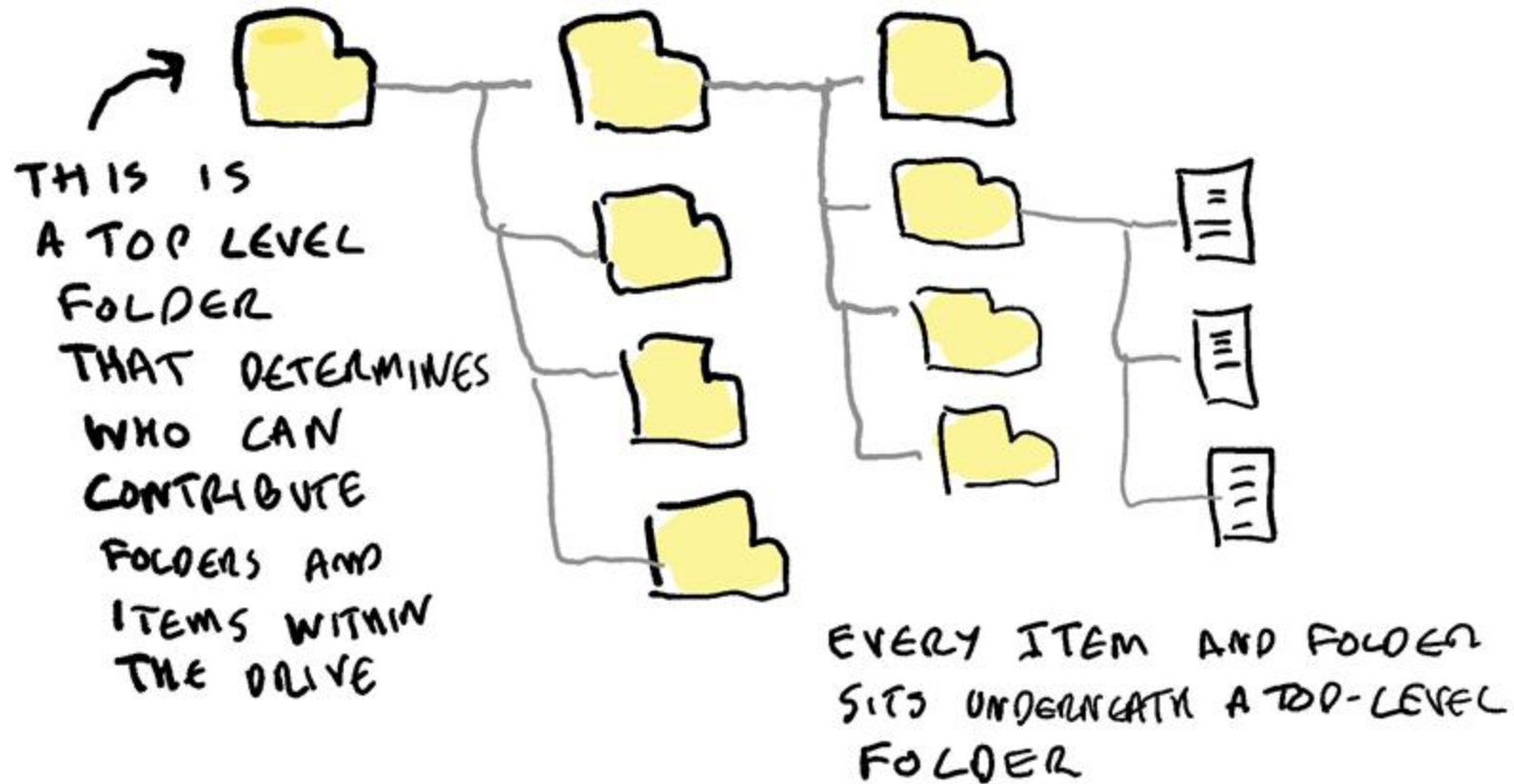


Content in corporate digital systems tends to be partitioned into containers within which a defined individual, team or work group can contribute content

Intuition 1



A team could not save content into an on-premise shared drive (fileshare) unless a top-level folder had been provisioned for them

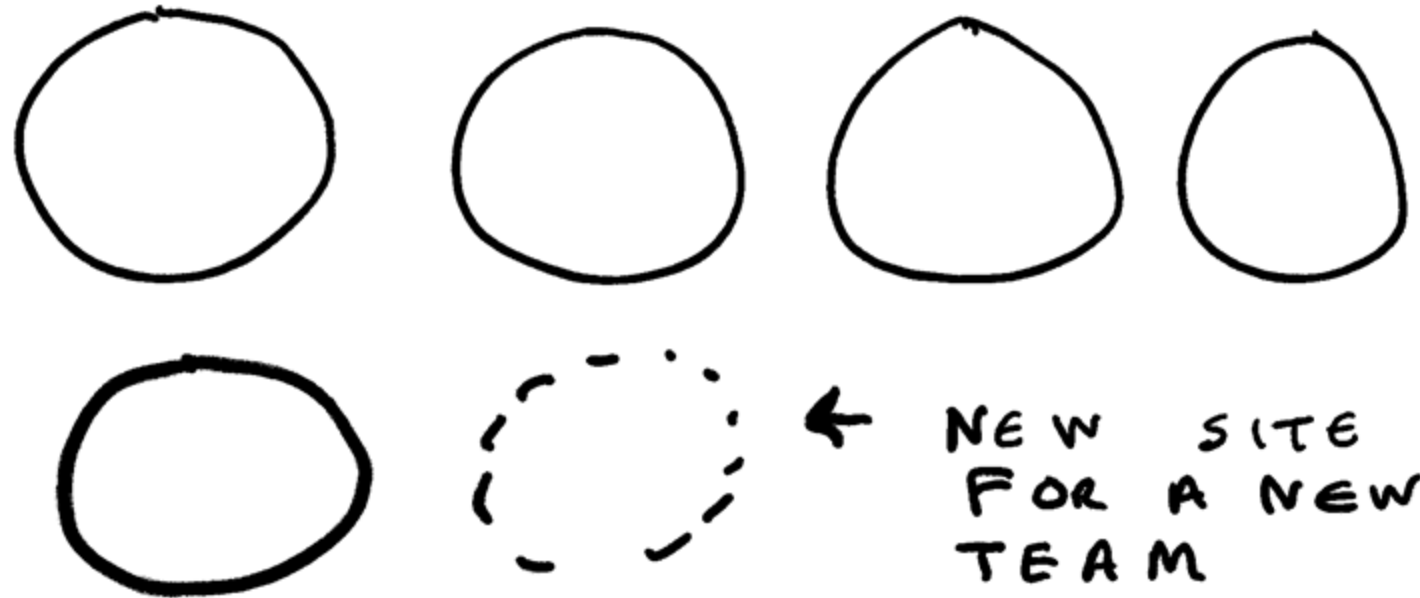


An individual could not use an email system unless an email account had been provisioned for them

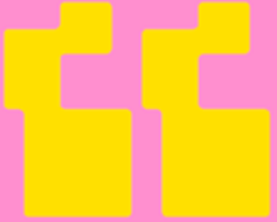


Every message on an email system sits within an email account

A team can could not save content into SharePoint unless a site had been provisioned for them



EVERY DOCUMENT IN SHAREPOINT SITS WITHIN A SITE



Control at the point of provisioning is
the simplest way of avoiding the
formation of digital heaps

Intuition 2



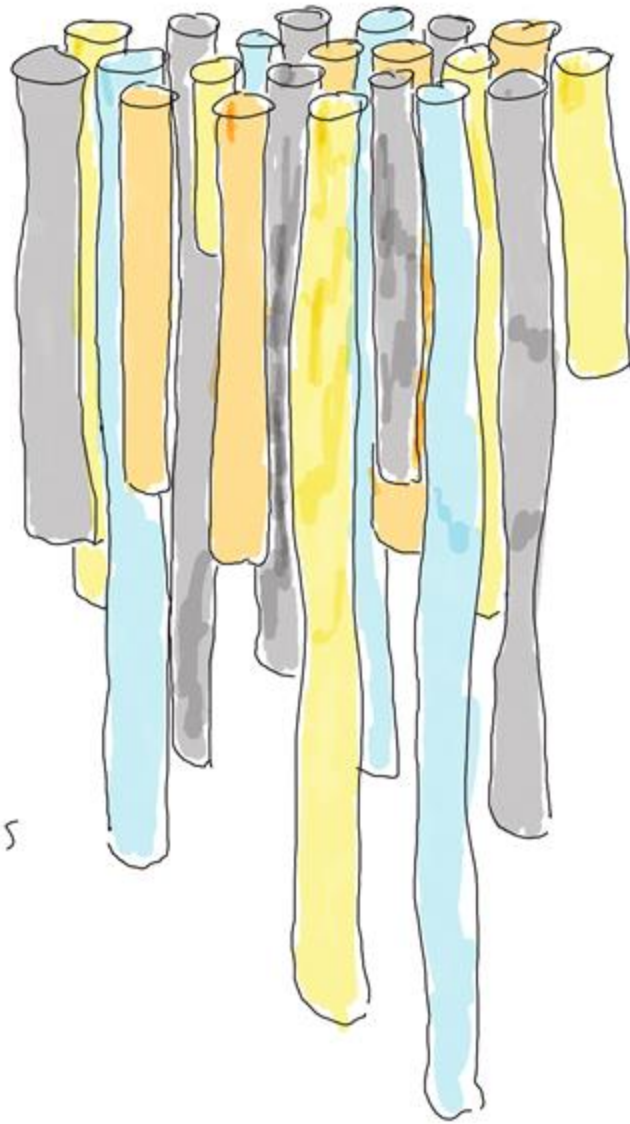
CONTROL AT THE POINT OF PROVISIONING

0 YEARS

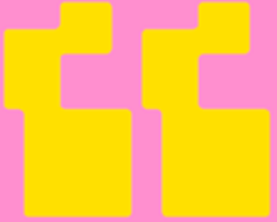
5 YEARS

10 YEARS

20 YEARS



SET ACCESS AND
RETENTION RULES
AT THE POINT AT
WHICH A CONTAINER
IS PROVIDED TO A
TEAM OR INDIVIDUAL

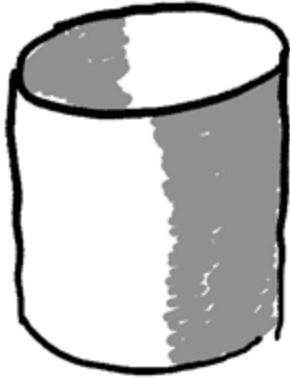


Any container to which a default access rule is applied can also receive a default retention rule

Intuition 3



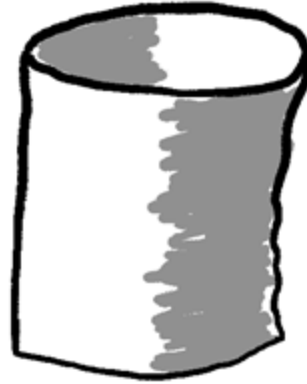
A CONTAINER COULD BE SCOPED ON :



A PIECE OF
WORK



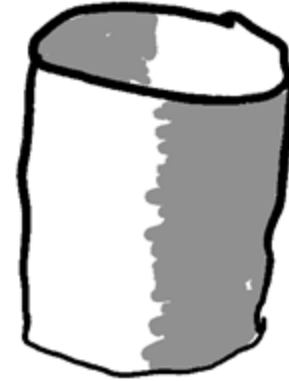
SOME PIECES OF
WORK ARE MORE
IMPACTFUL THAN
OTHERS



A TEAM



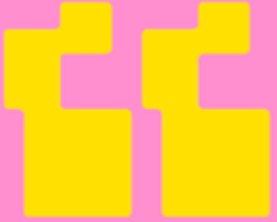
THE ROLE OF
SOME TEAMS IS
MORE IMPACTFUL
THAN OTHERS



AN INDIVIDUAL



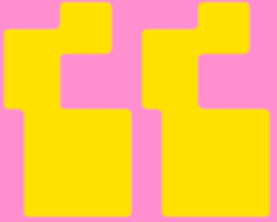
THE ROLE OF
SOME INDIVIDUALS
IS MORE IMPACTFUL
THAN OTHERS



The precision by which retention rules
can be applied to a container depends
on the scope of the container

Intuition 4

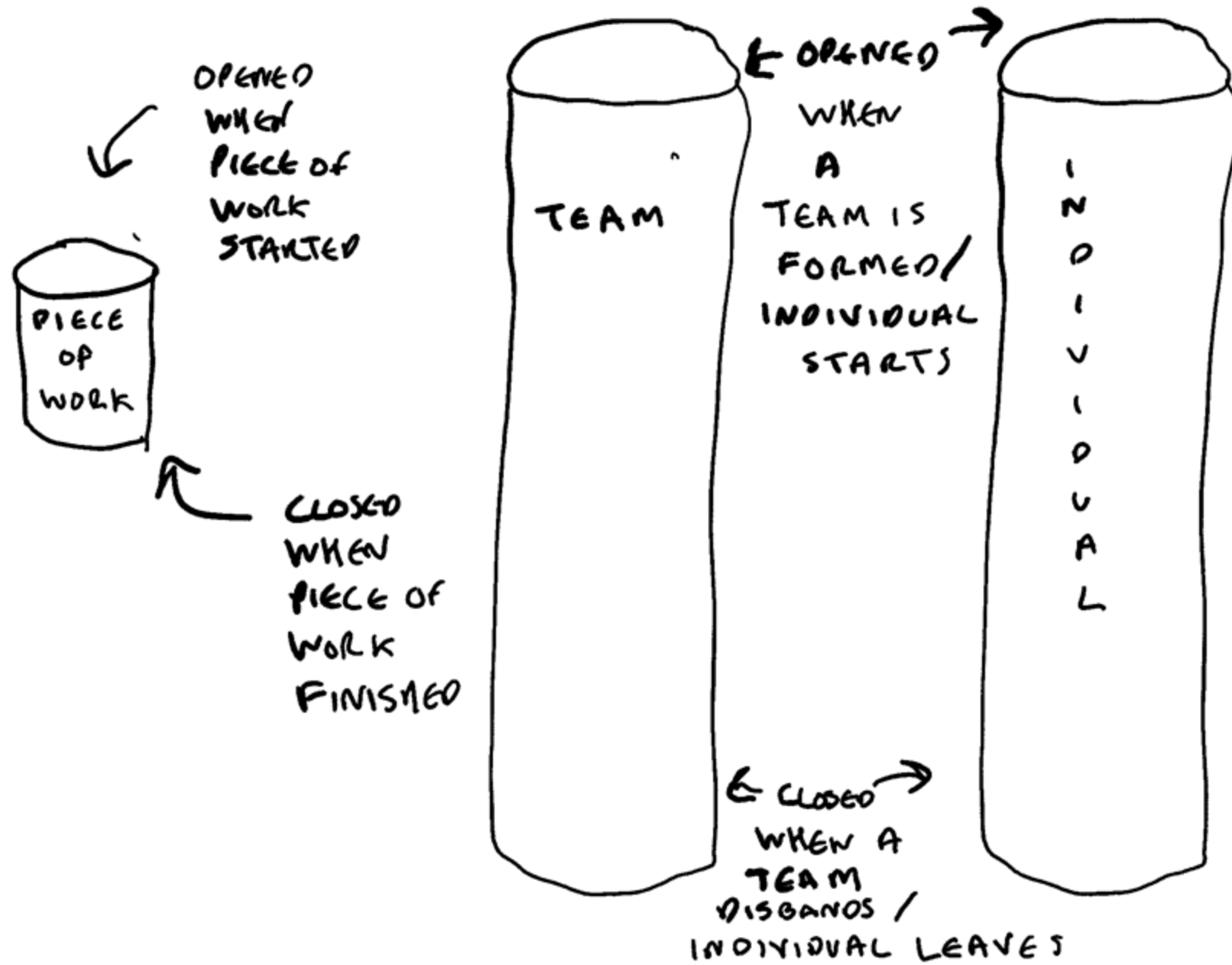


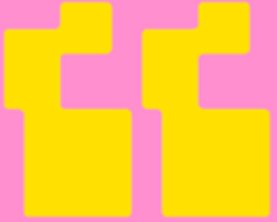


If a container is scoped on a specific piece of work then a retention rule can be applied with optimum precision

Intuition 5







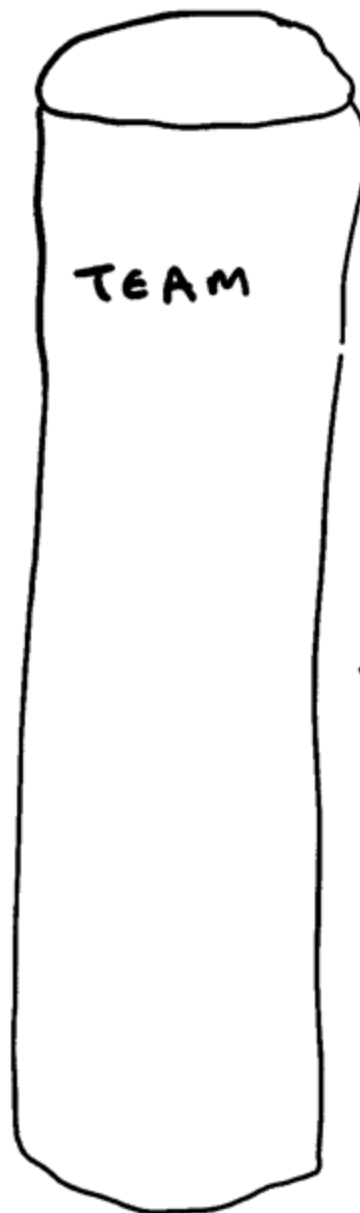
If a container is scoped on a team or an individual then it will contain content from a range of different pieces of work. The retention rule that can be applied will be less precise.

Intuition 6

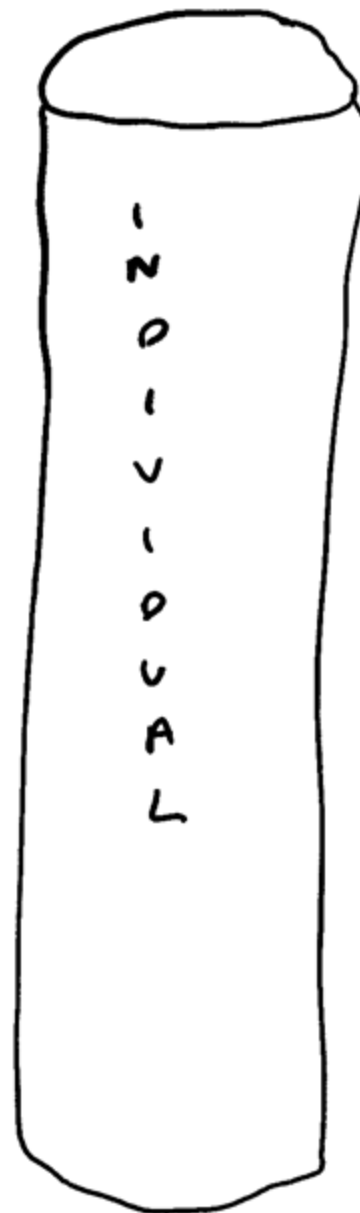




COVERS
ONE
INSTANCE
OF
ONE
TYPE OF
WORK

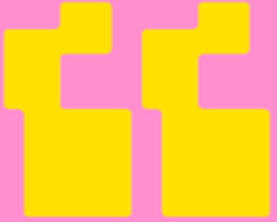


COVERS
MULTIPLE
INSTANCES
OF
A
RANGE
OF
TYPES
OF
WORK



The background is a solid yellow color. On the left side, there is a green geometric shape that looks like a large 'L' or a corner piece. On the right side, there is a pink geometric shape that also looks like a large 'L' or a corner piece, mirroring the green one. The text is positioned in the lower-left area of the yellow background.

Intuitions about acting on content



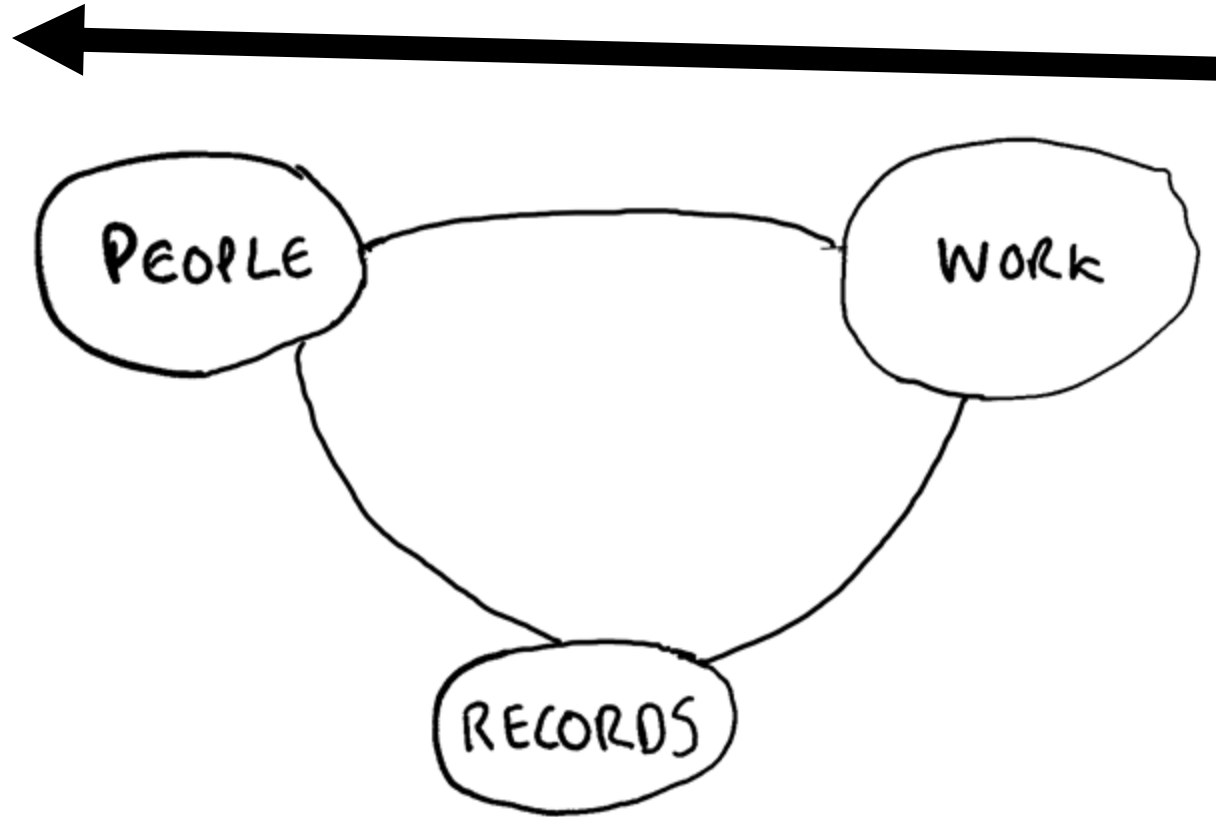
Start by looking at the context, not the
content

Intuition 7



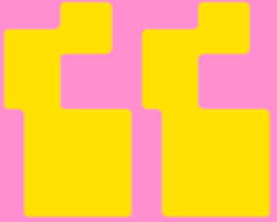
What teams and individuals were responsible for those areas of work?

Which teams and individuals were responsible for the strategic direction of the organisation?



What were the most important activities, initiatives, responsibilities and programs of the organisation?

What containers did those teams and individuals use?



Act on containers before acting on
items

Intuition 8

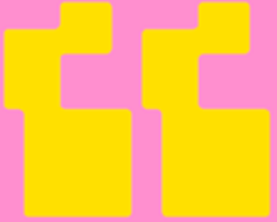


Imagine a shared drive with circa **10 million items** underneath circa **1,000 top level folders**

Imagine an email system with circa **10 million messages** in circa **1,000 accounts**

Imagine a SharePoint system with circa **5 million items** in circa **1,000 sites**

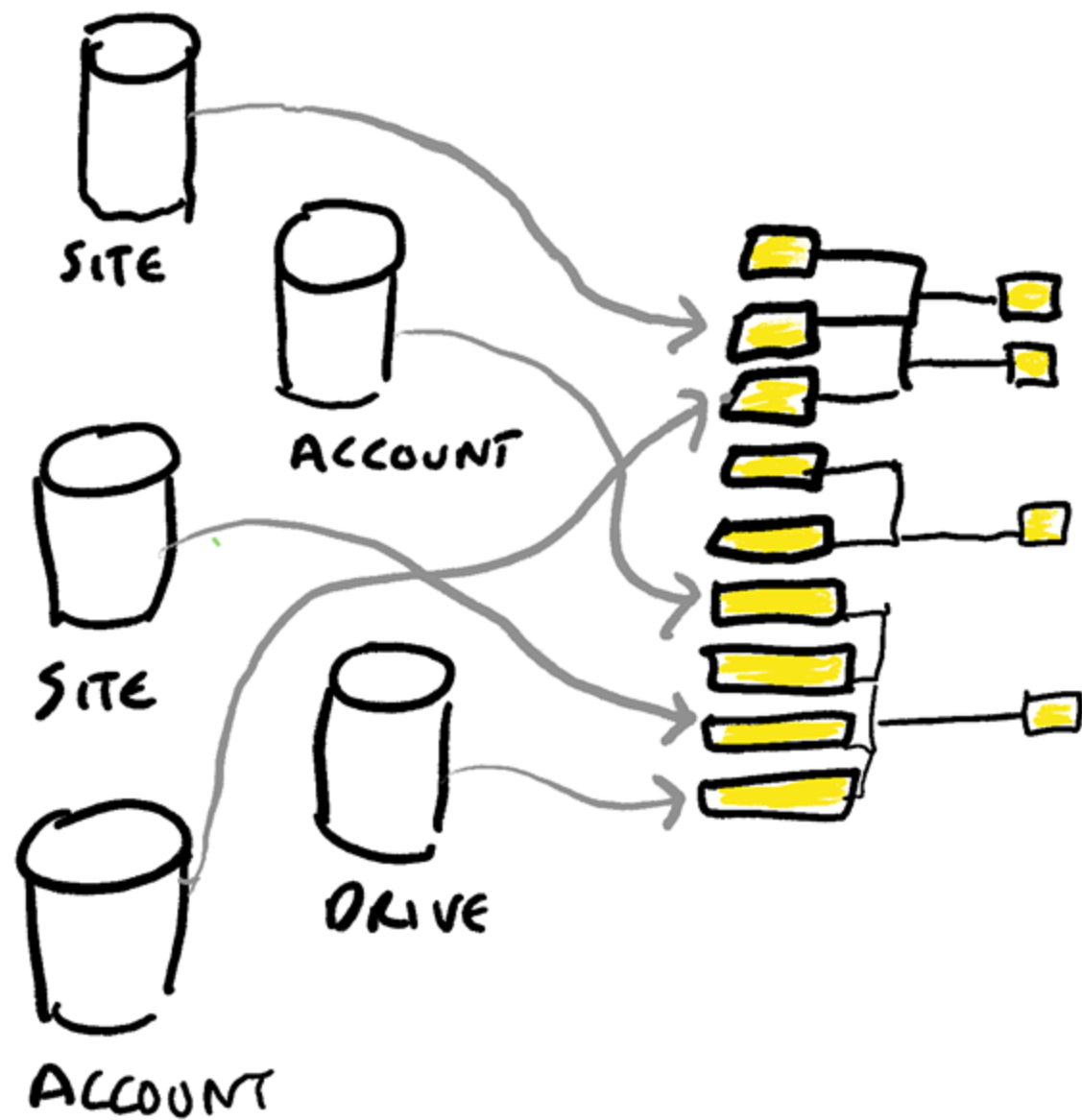
By concentrating on containers rather than items in the first instance, you reduce the number of decisions you have to make by two or three orders of magnitude. You also keep content in its original context.



Add context to containers - but don't
mix or merge them

Intuition 9

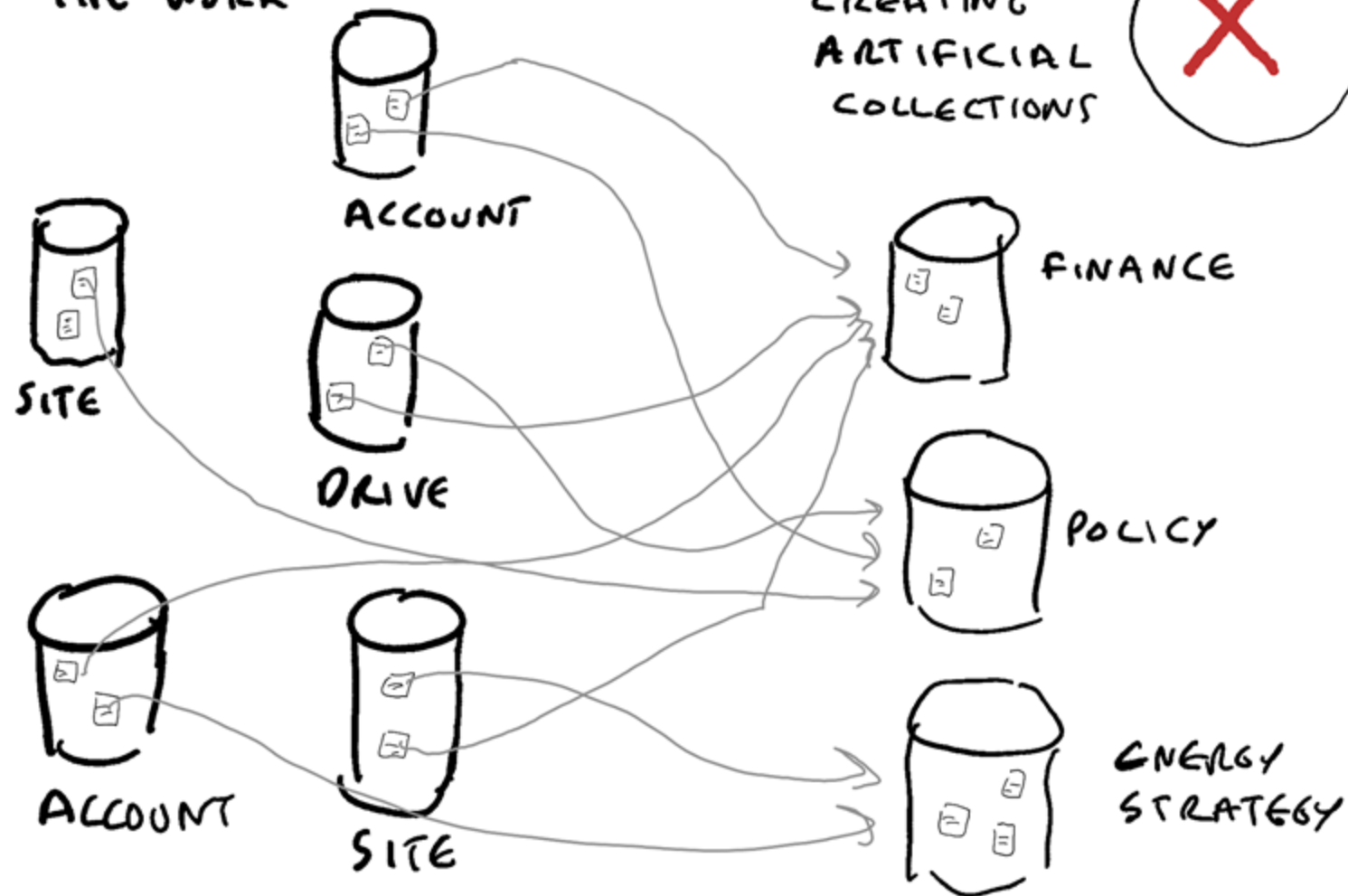


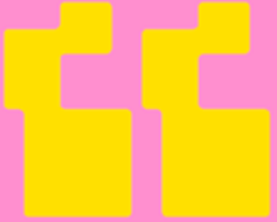


ASSIGNING
CONTAINERS TO
A TAXONOMY



CONTAINERS USED
AT THE TIME OF
THE WORK

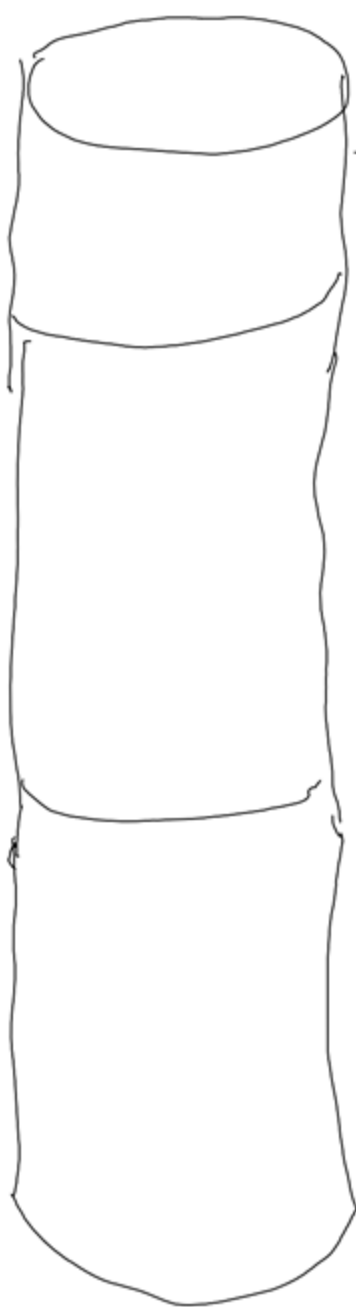




Remediate within those containers that
are worth retaining

Intuition 10



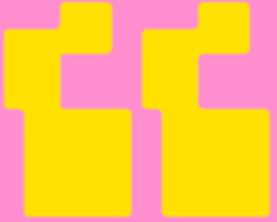


Intrinsic value the landmark documents/messages that were the key outcomes/deliverables of the work of the team or individual

Contextual value the humdrum documents/messages through which work was conducted. On their own each individual item has negligible value, but taken together they enable the landmark documents of the team or individual to be better understood, questioned and defended

No value documents/messages that had no impact or connection to any piece of work of the individual or team. For example:

- uncommunicated drafts
- personal, social and unsolicited correspondence



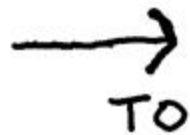
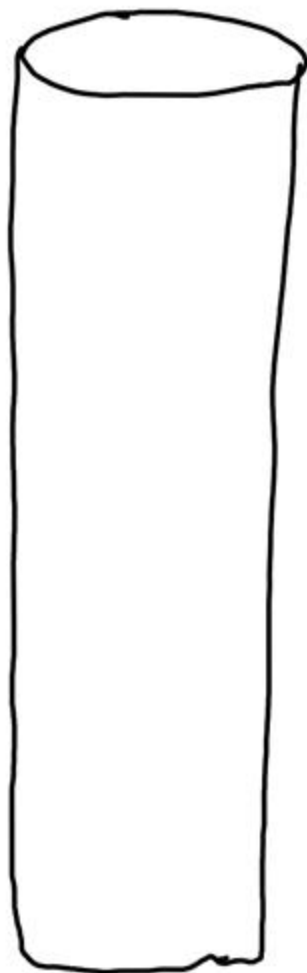
If a container is too big to process,
break it down into smaller chunks

Intuition 11



COULD YOU FIND A WAY TO GO FROM:

AN EMAIL
ACCOUNT
WITH
10,000
MESSAGES



25 CLUSTERS,
EACH OF
CIRCA
400
MESSAGES



Here is one potential way that you might go about breaking a container (such as a large email account) into more manageable chunks.

This approach is included to encourage experimentation and testing

```
CHOOSE A CLUSTERING
ALGORITHM
|
TARGET IT AT THE CONTAINER
|
DEFINE THE NUMBER OF
CLUSTERS YOU NEED
|
ITERATIVELY WALK WITH THE
ALGORITHM TO IMPROVE THE
COHERENCE OF THE CLUSTERS
|
VALIDATE THE CLUSTERS
|
SELECT WHICH CLUSTERS
ARE WORTHY OF PRESERVATION
|
DEFINE AN ORDER FOR
SENSITIVITY REVIEWING
THOSE SELECTED CLUSTERS
```



Government
Digital & Data

James Lappin

james.lappin@dsit.gov.uk