

# SEVEN WAYS TEXT ANALYTICS ENGINES TAME BIG DATA

John Felahi, Chief Strategy Officer

Steven Toole, VP Marketing

Content Analyst Company



11720 Sunrise Valley Drive  
Reston, VA 20191

[www.contentanalyst.com](http://www.contentanalyst.com)



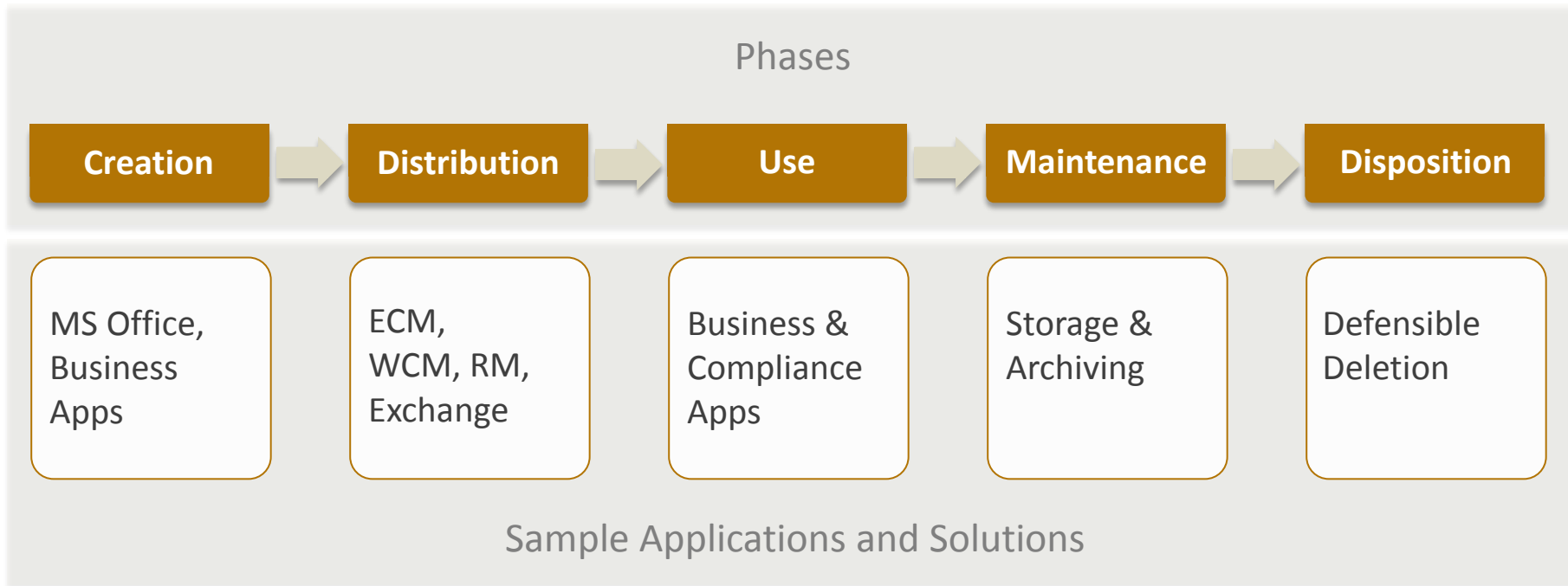
# Big Data – Value and Risk

- Information is either a strategic asset or risk
  - **Strategic** if classified so all applications know how to act on/manage it
  - **Risk (and a Burden)** if it sits as “dark data” driving up storage costs and potentially being a litigation nightmare
- The underlying issues
  - Ability to *easily* create and maintain a classification scheme
  - Have it applied conceptually so all documents that match the category are “tagged” no matter what terminology was used when document was created
- The goal - It is all about “findability” and being able to act on a classification
  - This is true for small or extreme collections of unstructured information
  - With bigger collections is there usually is more unknown data
    - Need to identify what is “junk” so you can get rid of the useless part

***Its all about good information management***

# Information Lifecycle

## All Phases Need to be Addressed



***Consistent classification for consistent findability***


# CAAT Understands Concepts Like Humans Do

- “Super analyst” reads each document, understands relationship between each, organizes, finds more similar, uncovers synonyms, dupes & near dupes, etc.
- Example docs “train” CAAT to learn from what you feed it; continues learning as you refine it
- Patented, deterministic mathematical algorithms define relationships between words and documents
- CAAT responds with conceptually relevant documents and presents them in an organized way
- Language agnostic; no dictionaries required

SEVEN WAYS TEXT  
ANALYTICS ENGINES  
TAMES BIG DATA

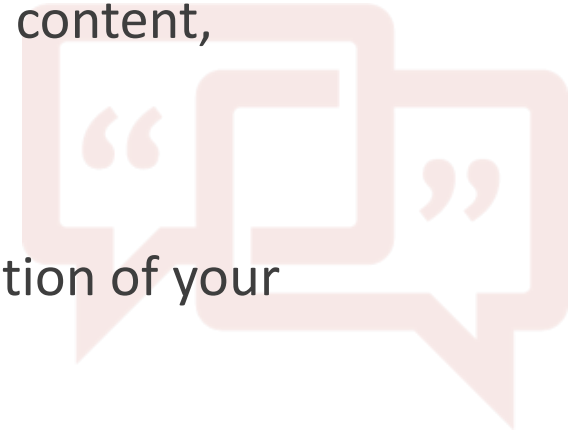
- Big Data Challenge:
  - **Enable your customers to DEFENSIBLY dispose of redundant, outdated and trivial (ROT) documents and e-mails.**
- Solution
  - **Concept aware categorization** reads content the way humans do, understanding concepts
  - **Instant context** uses known keywords to find conceptually similar terms based on their usage in the data set
  - No dictionaries, thesauri, ontologies needed with example-based approaches
- Results:
  - Precisely, Consistently, Systematically and Defensibly Deletes Junk

# #1: DEFENSIBLE DELETION

- Big Data Challenge:
    - **Enable your customers to maintain ARCHIVING REGULATORY COMPLIANCE**
  - Solution:
    - Automatically group conceptually similar content together
    - “Find more like these” identifies records for retention
    - **Text Near Dupe** and **Conceptual Near Dupe** finds nearly identical content based on the words, or the concepts
  - Results:
    - Content-aware archival for better compliance and reduced big data burden
- 

## #2: MAINTAIN COMPLIANCE

- Big Data Challenge:
  - **Improve cross-functional, divisional, and external content sharing and collaboration.**
- Solution:
  - Concept-aware auto categorization makes unstructured content much easier to find
  - Applies categories to documents, email, web content, database content
- Results:
  - Improved collaboration, sharing, and syndication of your customers' valuable content.



## #3: COLLABORATION

- Big Data Challenge:
  - **Maintain relevant taxonomies as new terms evolve**
- Solution:
  - Concept-aware text analytics understands concepts and context of terms
  - Automatically retags historical content with new terms and categories
- Results:
  - Content remains categorized with any new terms



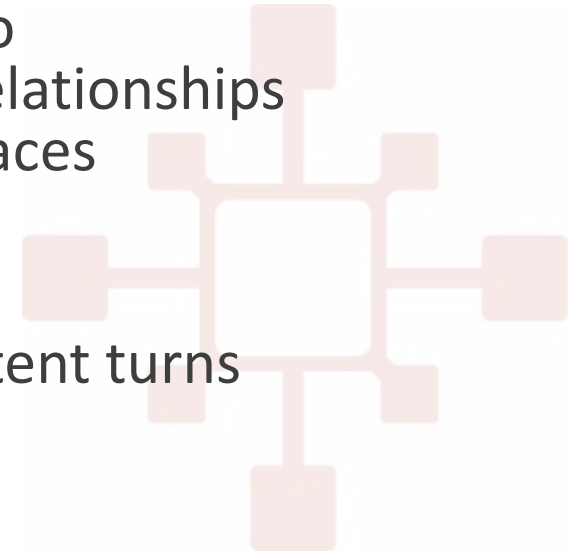
## #4: TERM DRIFT

- Big Data Challenge:
  - **Obsolete content could be an unnecessary liability and increase the cost burden to cull through in any future litigations.**
- Solution:
  - Concept-aware auto categorization can reduce risks by enabling your customers to identify and defensibly dispose of these materials
  - Email threading identifies who knew what, when and who they forwarded it to internally or externally
- Results:
  - Improved security, reduced cost, mitigated risk

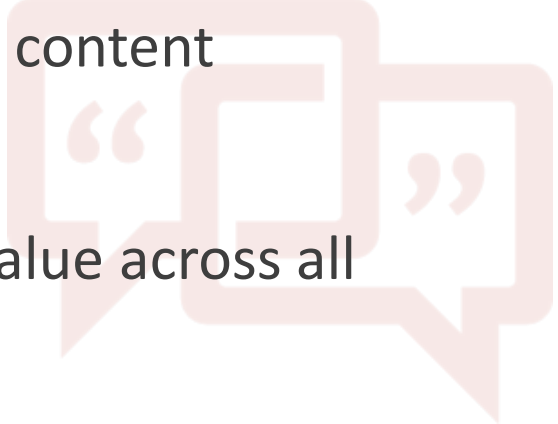


## #5: RISK MITIGATION

- Big Data Challenge:
  - Identify trends and correlations across unstructured big data
- Solution:
  - Unassisted **clustering** groups content into conceptually similar buckets, revealing relationships between concepts, terms, people and places
- Results:
  - Powerful insights from unstructured content turns big data from a burden to an asset



## #6: TREND ANALYSIS

- Big Data Challenge:
    - Applying all of the previous six approaches across any language
  - Solution:
    - Concept aware text analytics is language-agnostic, enabling concept-aware auto categorization in any language
    - No native speakers required to categorize content
  - Results:
    - Reduced big data burden and increased value across all languages
- 

## #7: LANGUAGE

# Summary

- Proven advanced analytics leader
- Highly effective engine
- OEM model to enhance partners' solutions in the shortest time
- Program and methodology for your short and long term success

***Better Analytics for Smarter Solutions***

# Company Credentials

- 13 patents awarded/pending
- Dozens of OEM partners
- 10,000+ users via partners
- Named in Gartner's "Who's Who in Text Analytics"
- KM World Top 100 (5 Years)
- KM World's Trend Setting Product Company (5 Years)

# THANK YOU

John Felahi, Chief Strategy Officer – [Jfelahi@ContentAnalyst.com](mailto:Jfelahi@ContentAnalyst.com)

Steven Toole, VP Marketing – [smttoole@ContentAnalyst.com](mailto:smttoole@ContentAnalyst.com)

Content Analyst Company



11720 Sunrise Valley Drive  
Reston, VA 20191

[www.contentanalyst.com](http://www.contentanalyst.com)