StreamScape™

Data Classification
with Cognitive AI

Using Semantics and Thematic Analysis to identify and categorize text and documents

Today's organizations create and manage more digital content than ever before. Much of it stored as documents, spreadsheets, text, and media files, frequently containing personal information and sensitive, financial, or healthcare data. Often, embedded in the text are also critical terms and key phrases that can identify the themes, subjects, and significance of a document. Unfortunately, the immense *volume* and *velocity* of such data makes it difficult to separate relevant, high-value information – from noisy, low-quality data. Inability to turn data into knowledge can have a crippling effect on human decision making, often resulting in poor judgement, and misunderstanding of risk.

Sorting through data across an enterprise can be an expensive and cumbersome process. Few organizations are equipped to handle data classification by traditional (manual) methods. Automation tools can help streamline the process, but an enterprise must determine the categories and criteria that will be used to classify data and clearly define the objectives. Using *Cognitive AI* to analyze and classify data at scale can significantly improve knowledge worker productivity and decision making by automatically discovering useful patterns, trends or data relationships based on semantic meaning and context.

Cognitive Analytics AI
Improves Quality
and Consistency
of Human Decision Making
with Machine Guided
Outcomes

Data classification offers a higher level of data management, security, and control, by separating and organizing information and files into relevant groups ("classes") based on shared characteristics ("features"), such as demographics, economic profile, level of data sensitivity, the risks it presents, or the type of legal and regulatory compliance.

Thematic Analysis further improves data management by mimicking human judgement and perception to unlock hidden knowledge in enterprise data. It works by assigning features to data that prescribe how to process, retrieve, and secure each group ("category"). When done correctly, this process will enable employees and third parties involved in storage, transmission, or retrieval of data with the ability to turn data patterns and relationships into actionable knowledge.

Semantics: When Context Matters

One critical feature of *Thematic Analysis* is the ability to identify and classify text or documents by Topics or Semantic Meaning. Analyzing term relationships, themes and similar key words or phrases will often uncover deeper meaning in data and help organize information into similar areas of interest. This type of semantic analysis lets machines simulate human cognition by classifying results based on specialized context or content, in the same way a subject-matter expert can do when analyzing complex information.

Understanding themes and semantics is increasingly important given the enormous amount of text and documents being generated by modern systems. Research indicates that about 30% of all documents read by specialists are mis-categorized. In some area as high as 40% when industry-specific jargon is used. Contextual analysis performed by *Cognitive AI* can greatly reduce or eliminate the manual effort of document review and categorization, providing more accurate results then a typical human reviewer can. Semantic context is established in the following ways:

Synonyms + Antonyms
When information needs to be included or excluded based on multiple terms that have the same meaning

Similarity + Exogeny
When information needs to be identified by contextual meaning or entities that do not belong to a specific domain

Hypernymy
When information needs to be sorted by semantic relation of belonging to a category or super-group (ie. an instance of)

Meronymy
When information needs to be identified as a constituent part of a greater whole (ie. a wheel is part of an automobile)

Pragmatic Inference
When the resulting information is derived by expanding on initial, available facts or evidence (ie. an educated guess)

Combining these language processing techniques, semantic analysis models can simulate the process of human decision making. Most importantly, classification AI learns from a mix of user actions and AI operations, improving productivity and freeing up users to engage in higher value activities.

Why Classify Your Data?

Data classification is part of *predictive analytics*. Classification models are trained to break down text and documents into features (relevant terms and phrases) that can identify content or infer meaning based on domain-specific rules. Feature terms are then associated with their source data and used to define data categories – a technique called clustering or data tagging.

Cognitive AI services can discover, organize, and retrieve information based on similar content, unique features, or semantic meaning. Data classification can also be used to establish intent or sentiment in a body of text and predict features that are unknown or may appear in the future. Classification and Thematic Analysis have broad application in decision support systems, legal due diligence, direct marketing, regulatory compliance or data privacy, insurance fraud detection and medical diagnosis.

Classification typically makes use of the following disciplines:

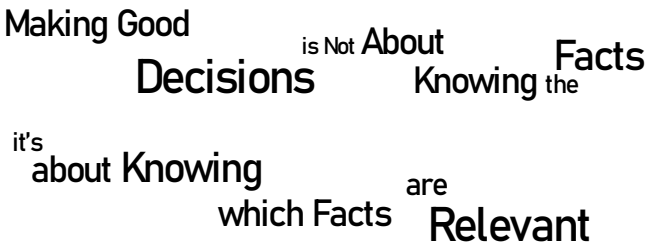
Identification – *Discovery of data that exists across the enterprise*

Correlation – *Cataloging of data sources and relationships*

Categorization – *Generation of Features that describe Data Content*

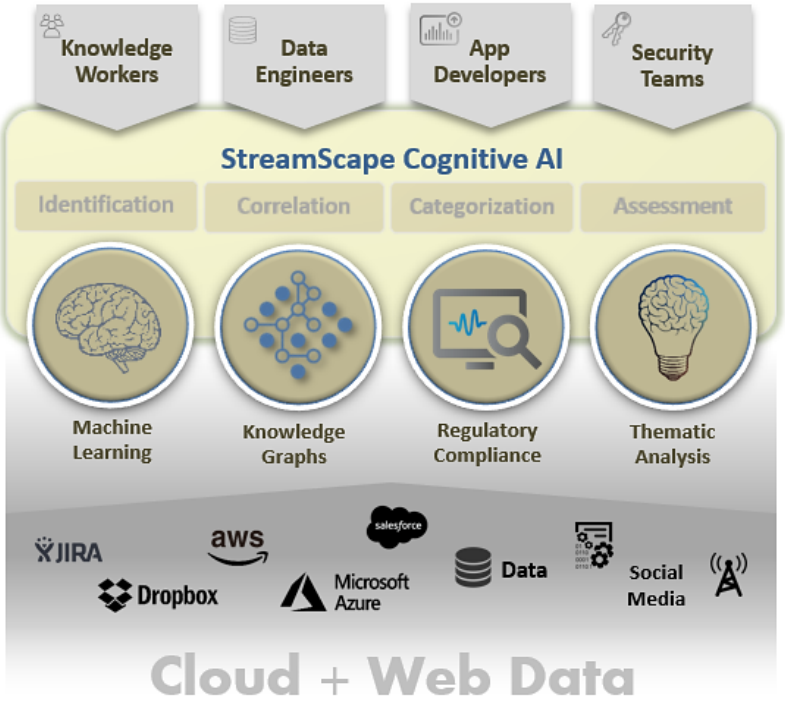
Assessment – *Analysis of content to determine theme and meaning*

Automating the process of discovery and identification makes the data more usable and secure, improving the quality of information used in decision making. According to a recent study by HFS Research, 75% of executives do not have a high level of trust in their data. Despite an overwhelming majority that use data as the main decision-making tool, data quality remains a critical and costly issue. Classification simplifies the task of detecting critical content, extracting relevant aspects, and eliminating duplicate or erroneous information, wherever it resides across an enterprise.



Inference: When Relevance is Critical

One significant challenge of *Classification* and *Thematic Analysis* is the ability to accurately identify relevant content. To solve this problem, *Cognitive AI* makes use of Data Dictionaries and Semantic Graphs to facilitate *inference queries*. Inferences are made when a person (or machine) goes beyond available facts or evidence to draw a conclusion. The result is *inferred knowledge* that expands on initial information, adding new data into the output. A critical part of accurate inference is being able to describe missing information using familiar terms that are already part of a user's general knowledge.

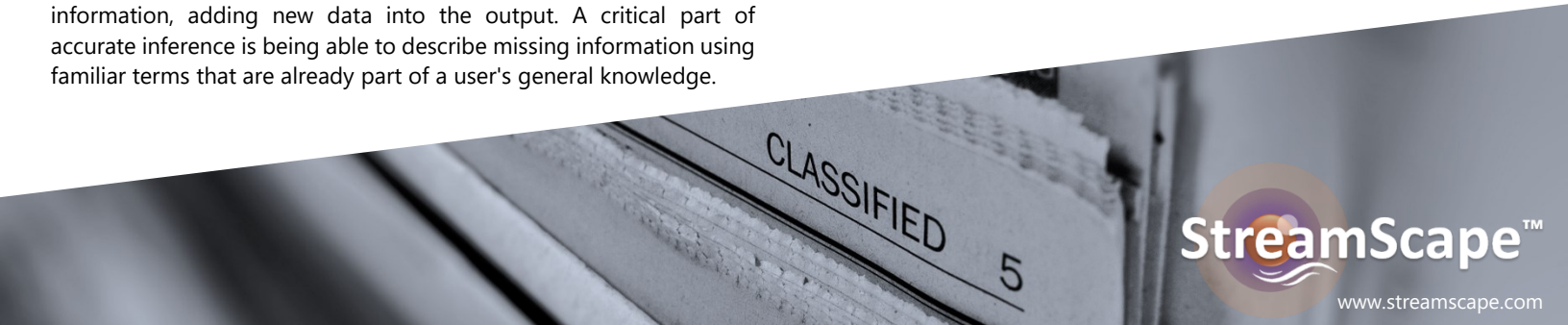


At the heart of our platform is an innovative Dataspace™ technology purpose-built for *Cognitive AI* that offers broad connectivity to SaaS infrastructure and hundreds of Cloud and Web data sources.

Dataspaces make use of Inference Types to annotate source data with additional information, filling the knowledge gap. The annotations may be populated by looking up related terms or concepts in a domain-specific *semantic graph* or resolved using a *data dictionary service*, allowing the original information to be tagged with synonyms or enriched with relevant features. This new knowledge can be used in data fabric queries to Join or Lookup related information, simulating Pragmatic Inference of human decision making.

Key Benefits of Classification

- *Organize Information into Categories to uncover Relationships, Meaning and Relevance – thereby Improving Data Quality*
- *Discover Relationships and Critical Patterns across Enterprise Data Sources to Improve and Automate Decision Making*
- *Identify, Manage and Protect sensitive data to meet Regulatory Requirements and Demonstrate Compliance*
- *Support Data Governance requirements such as Data Protection Act (DPA), European General Data Protection Regulation (GDPR), Sarbanes-Oxley Act, HIPAA and ITAR*
- *Allow Visualization Tools to present Related Data in intuitive ways, easy to navigate – for example via Knowledge Graphs*



StreamScape™

www.streamscape.com