Achieve More With Data

Content Data Science: The Key to Extracting Value from Unstructured Data

Did you enjoy House of Cards on Netflix? The pioneering media company anticipated you would, investing $100 million in the series' first two seasons based almost entirely on data. Netflix identified that fans of the British version also liked Kevin Spacey and David Fincher's work. Blending these "ingredients" into the 2013 remake helped deliver one of the studio's biggest hits. While you may not be in the entertainment industry, a similar approach — using content data science — can guide you in creating content tailored for maximum impact within your organization.

What are the basics you should know about this emerging field? This article outlines what content data science is all about, covering:

- The business goals it supports vs traditional data science
- The technical challenges it addresses
- The key role AI plays in addressing these challenges
- Examples of how traditional and content data science can come together to improve performance across multiple enterprise business functions.

What Is Content Data Science?

Author and entrepreneur Peter Hinssen has recently popularized the term content science, which he defines as "the management of a company's unstructured data paired with the vast potential of LLM platforms." He uses the term 'content' to refer to unstructured documents or 'unstructured data' — such as digital documents, slideshow presentations, emails, images, etc.

This relatively new field, enabled by the rapidly accelerating capabilities of AI, is particularly exciting because the vast majority — a full 90%, according to an IDC white paper — of most organizations' data is unstructured. This data is found in a wide variety of content like marketing materials, internal documentation, instruction manuals, images, and videos, to name just a few examples.

Whereas traditional data science primarily involves working with structured data neatly organized in databases and spreadsheets, similar principles can be applied to unstructured data found in content. For that reason, we prefer to think of this extended application of these principles as content data science.

Why Content Data Science Matters: Business Goals vs. Traditional Data Science

To understand the business value of content data science, it helps to compare its overall purpose to that of traditional data science in an enterprise setting. While both disciplines are focused on deriving actionable insights, the types of insights and their applications differ significantly. Here is a breakdown of these differences:

	Traditional Data Science	Content Data Science
Insights	Quantitative snapshots explaining "what is happening."	Qualitative and quantitative pictures explaining "why something is happening."
Objective	Support business decision-making. For example: • Identifying the optimal time for a product launch based on sales data • Adjusting pricing strategies based on seasonal trends • Optimizing inventory management based on demand forecasting	Improve the overall effectiveness of content. For example: • Understanding audience needs and preferences • Identifying content gaps and opportunities • Enhancing content quality and engagement • Informing content governance and compliance

The Technical Challenges Content Data Science Addresses

Extracting valuable insights from structured data is not easy. Extracting them from unstructured content data is even more difficult because it involves more complexities that traditional data science methodologies often fall short in addressing.
Top challenges include:

- Storing and retrieving unstructured content data, as it is not organized in a predefined manner.
- Handling immense volumes of multi-type data found in text, images, videos, etc.
- Extracting nuanced information like the sentiment behind a customer review.
- Content data science aims to overcome these challenges by employing advanced methodologies such as natural language processing (NLP), sentiment analysis, and topic modeling. Enter AI.

AI: A Core Technical Enabler of Content Data Science

Artificial intelligence plays a crucial role in content data science. By discerning patterns from extremely large datasets of unstructured data, large language models (LLMs) can extract value from such content through:

Analyzing Content and Generating Insights
AI algorithms, particularly those based on NLP, are essential for analyzing unstructured content such as text documents, social media posts, and customer reviews. These algorithms can identify correlations, extract key themes, and understand sentiment, providing valuable insights into customer behavior, market trends, and content performance.

Example
A content data scientist at a CPG company might use AI-powered sentiment analysis tools to understand why a specific ad banner is performing poorly. By analyzing social media comments and online reviews, the AI can identify negative sentiment associated with certain design elements or messaging, helping the company optimize the banner for better results.

Personalizing Content Experiences
AI can analyze user data and behavior to create personalized content experiences. Doing so can involve recommending relevant articles or products, tailoring website content to individual preferences.

Example
Netflix leverages AI to personalize content recommendations for its viewers. By analyzing viewing history and preferences, AI algorithms suggest shows and movies that align with individual tastes, enhancing user engagement and satisfaction.

"Grafting" Internal Content with Knowledge Tools

One of the most promising applications of AI in content data science is the concept of "grafting." This involves combining the power of LLMs with a company's own internal content, creating custom AI tools that are uniquely tailored to the organization's knowledge base.

Example
McKinsey developed Lilli, an internal GenAI tool that combines a commercially available LLM with the company's vast collection of documents and reports. Lilli enables McKinsey consultants to quickly access relevant information and insights, potentially enhancing their efficiency and client service.

It is worth noting that insights gained through AI-enabled content data science can form a natural foundation for developing future content with generative AI (GenAI) tools. For example, the creation of blog articles, social media posts, and other forms of marketing content — even dynamic email campaigns that adapt to user interactions — is one of the most widely recognized categories of use cases for GenAI. Also, AI can help govern content to ensure adherence to brand guidelines, legal requirements, and ethical standards.

Powerful synergies between traditional and content data science

While traditional data science and content data science are distinct fields, they are rarely practiced independently. For example, a common technical goal of content data scientists is to bring structure to unstructured data using techniques such as retrieval-augmented generation (RAG) and knowledge graphs. These tools can help reduce the amount of training needed for LLMs and lower the cost of AI development.

Also, content data scientists may rely on structured data to inform how they will structure or curate unstructured data.

To illustrate the business value potential of these synergies in marketing and advertising, consider a hypothetical example of a CPG company that has launched a digital advertising campaign for a new product:

- Traditional data science methodologies could be used to analyze click-through rates and conversion data from the campaign and identify that one specific ad banner is performing significantly worse than others.
- Content data science methodologies could then be applied to identify the "why" behind the poor performance by analyzing the content (design elements, messaging, call to action, etc.) of the underperforming banner, investigate audience reactions via social media comments, and assess alignment with the overall campaign strategy — perhaps the banner was designed for a different stage of the customer journey, for example.

Marketing and advertising are, however, far from being the only business functions where synergies between traditional and content data science can add value. Here are a few more hypothetical examples:

Customer Service

Using traditional data science methodologies, analysis of a telecommunications company's customer service call logs reveals that customers calling about a specific new phone model have a significantly higher rate of unresolved issues, leading to increased call volume.

Content data science methodologies are then applied to investigate why call volume is up. This process involves analyzing the content of online support resources and the phone's user manual. The conclusion is that the instructions for setting up a particular feature are unclear or missing, leading to customer confusion and frustration.

Product Development

A regular analysis of sales figures and online reviews for a new line of organic snacks reveals that sales are below projections and customer feedback frequently mentions a bland flavor profile.

Content data science practitioners, focusing on understanding consumer preferences and perceptions, analyze customer interviews and surveys to uncover the "why" behind the negative feedback. They discover that the target audience associates organic snacks with stronger, more distinctive flavors and that the product's current taste doesn't meet expectations. They also analyze the packaging and marketing materials to ensure that they effectively communicate the intended flavor profile and brand values.

Human Resources

Through analysis of employee engagement surveys and performance reviews, traditional data science practitioners identify a significant drop in employee satisfaction and productivity within a specific department following the implementation of a new project management system.

Content data science practitioners then explore the "why" behind this trend by analyzing recordings of interviews between employees and managers, training materials, and digital communication patterns within the department. They find that the training materials for the new system are inadequate, leading to a lack of understanding and adoption among employees. They go on to investigate whether communication channels are effectively facilitating collaboration and knowledge sharing within the team.

Sustainability Reporting
Analysis of structured data related to a manufacturing company's environmental impact reveals that water usage has increased significantly over the past year, exceeding targets set in the company's sustainability plan.

Analysis of unstructured data from documentation related to operational processes, maintenance records, and employee training materials reveals that a new manufacturing process — while intended to be more efficient — is using more water than the previous method due to a miscalculation in its design.

Key Takeaway

By embracing a comprehensive approach to content data science — which relies on AI to make sense of unstructured data in a wide variety of content — organizations can move beyond just analyzing "what" is happening and understand "why" it is happening. The resulting insights can lead to better decision-making and a stronger competitive edge across multiple business functions.

Taking this comprehensive approach to content data science involves sophisticated tools and techniques. In our next blog post, we will explore the most important of these tools and techniques and how Lingaro can help you make the most of them.

Meet our Experts:

Norbert Fijałek
AI Engineering Team Leader

Norbert is passionate about AI and Machine Learning Operations. He has spent more than 10 years exploring and implementing new technologies, designing and developing innovative products, and leading teams to deliver successful AI and product management solutions across multiple industries.

Norbert enjoys sharing his knowledge with others and has been a featured speaker at various conferences on AI and related topics. He also participates in successful accelerators and competitions.

Overall, he is a lifelong learner who loves being on the cutting edge of the latest technology trends and exploring new ways to apply them to real-world problems.

Krystian Jabłoński
Head of GenAI Practice

Krystian has spent over seven years honing his skills in the data and analytics field, with a focus on creating and developing generative AI services. He has demonstrated leadership in managing product and product teams, as well as AI R&D activities, and is appreciated for his ability to translate business challenges into practical AI solutions.