The A to Z of Understanding AI and Big Data

Algorithms, Generative AI, Retrieval Augmented Generation—understanding the language involved in Artificial Intelligence (AI) and big data initiatives can be daunting when you aren't a data scientist. But with AI predicted to transform the future of business, employment, and society as a whole, ignoring the jargon is no longer an option. Understanding the value of AI powered by credible data can help companies to grow, make better decisions, and manage risks.

A good first step an organization can take on that journey is to read this glossary. It aims to demystify AI by defining and explaining some of the key terms used by data scientists. In the process, we demonstrate how LexisNexis® supports companies to take advantage of the opportunities offered by AI and big data in an ethical and trustworthy way.

A

Aggregation – Identifying and collecting data, often combining different datasets.

LexisNexis brings together diverse and credible content in a wide range of areas, which companies can use to power their products and services. Our content covers news; company and financial information; sanctions data; Politically-Exposed Persons (PEP); legal data; and more.

Artificial Intelligence (AI) – Intelligent machines and software that can perceive their environment and act on it, often learning from those actions. AI can be applied in a wide range of fields, including risk and fraud detection, purchase and investment predictions, logistics and supplier management, news and entertainment creation, and online customer support interactions using chatbots.

Analytics – The discovery of insights based on data. There are three types:

- Descriptive analytics summarizes data to create an overall narrative.
- Predictive analytics analyses historical and current data to predict future behavior based on probabilities. For example, it can use trends in consumer preferences or in the stock market to inform buy-sell decisions.
- Prescriptive analytics builds on predictive analytics by analyzing its outcomes to decide the best action to take. It is the next evolution in deep learning to support decision-making without human interaction.

Algorithm – A mathematical formula that performs an analysis on a set of data, often embedded in technology.

Application – A software program that performs a specific function for the person (or application) using it.

Application Programming Interface (API) – An API provides a way to deploy the features of a specific application or service, which lets two applications interact with each other. For example, an API may specify how to retrieve data from an application.

API-first – When a company starts a digital transformation project by thinking about the API through which the data required will be delivered. Users across the organization can then pull data via the API to power their tools, models, internal systems, and other projects.

Archive – Usually a store of historical data that is no longer actively used. Data archives should be indexed for easy location and retrieval of files.

Artificial General Intelligence – Artificial General Intelligence (AGI) is the future many experts predict for AI and Gen AI tools, in which the intelligence matches or surpasses human capabilities.

B

Bias – Algorithmic bias occurs when the output from an algorithm is seen to be unfair, reflecting errors or even prejudices of the people who created the algorithm.

Big Data – Very large data sets that can be analyzed by computing technologies to reveal patterns and trends. Big data is the fuel for a wide range of AI applications.

Black Box – Where a technology or system produces results without humans being able to see how it arrived at them. Machine learning and generative AI are typical examples. This brings risks for users – see Bias, Hallucinations and Ethics.

C

Classification – Categorizing a data point based on traits it has in common with other data points. This allows the user to extract important and relevant information from a big dataset more quickly and easily.

Closed Loop – A type of AI model designed with privacy in mind, which does not let data leave the model. This data includes user prompts/ queries, uploaded documents, and output responses.

Correlation Analysis – Analyzing data to determine a positive or negative relationship between different variables.

Our comprehensive news data allows users to identify correlations between, for example, a company's actions and its reputation.

Credibility – It is widely understood that the results from AI will only be as effective as the data powering the technology. 9/10 of professionals interviewed in the LexisNexis Future of Work Report 2024 said their main consideration when using generative AI is the quality and accuracy of its output. Organizations need to acquire credible data sources from a trusted data provider.

As a trusted data provider of over 50 years, LexisNexis has extensive, long-standing – and in some cases, exclusive – content licensing agreements with publishers around the world.

Customer Relationship Management (CRM) – A system or strategy used by a company to manage its sales and business processes, which can be informed by big data. Integrate negative news, company, or legal data into a CRM system to provide additional context that empowers the Sales teams.

D

Data as a Service – Providing data to users over a network on-demand. This allows users to acquire and use external datasets, often in combination with their own data.

Data Analyst – An employee with the data and statistical skills to interpret and analyze data for insights. This job role is in high and growing demand from companies.

Data Cleansing – Reviewing data to see if it is still valid, as well as correcting errors, eliminating duplicates, and standardizing data formats for greater consistency.

Data Engineering – The behind-the-scenes work to build systems that allow data scientists to do their analysis more quickly and efficiently.

Data Feed – A stream of data, for example an RSS feed or a social media feed.

Data Governance Framework – The set of rules and processes for how data is organized, aggregated and managed.

Data Journalism – Using data to tell stories and identify patterns and trends. Data journalists have gained prominence with analysis of topics ranging from the impact of political ads in the media to spread of global COVID-19 pandemic and effectiveness of various responses.

LexisNexis offers an extensive collection of reliable media content dating back more than 50 years. We also provide news data from thousands of licensed titles that has been approved and optimized for generative AI (see: G).

Data Lake – A way of storing a vast amount of raw data, whether structured, semi-structured or unstructured. This data can be stored within an organization's data center or using cloud services.

Data Visualization – Communicating data visually, often using infographics, color- coded graphs, or data dashboards.

Data Wrangling – Taking raw data and formatting and restructuring it to make it useful. Data scientists often spend more than half of their time on data wrangling. As such, data that is already enriched and normalized can free up employee time to focus on higher-value work.

Deep Learning – Using very large neural networks to solve complex problems, such as facial recognition.

Developer Tools – A way for a firm to test and vet a third party's API against their organization's needs and use cases, before committing to using it to bring in data to power their products and services. The testing should be carried out by experienced developers, data scientists or related experts.

Nexis® Data+ offers users the opportunity to extensively test and vet our flexible API.

E

Enrichment – The process of making a raw dataset more useful and insightful by normalizing the format and applying tags that make it easier to search and use.

Nexis Data+ complements its comprehensive content coverage with enrichments which allow users to find relevant results and insights more quickly and easily. These can be integrated in AI initiatives via our API.

Ethics – The way data is used by AI tools can prompt serious concerns. There have been examples of breaches of data privacy regulations and unethical harvesting of individuals' data. Using a trusted, transparent and compliant data provider is essential.

From data acquisition to customer onboarding, we pride ourselves on offering data which is up-to-date, compliant with licensing agreements and applicable laws, and safeguarded by robust data security and privacy measures. By partnering with LexisNexis, you can be assured that your AI initiatives are built on a foundation of trust, transparency, and ethical principles.

F

Fuzzy Logic – An approach to logic that is widely used in AI. Rather than judging whether a statement is true or not, it judges how close to the truth it is.

G

Generative AI – A tool which generates content in response to a prompt or query from a user. These prompts might ask the algorithm to produce a written answer, imagery, videos, audio, and more. OpenAI's ChatGPT is perhaps the best-known Gen AI tool to date. Its potential applications are extremely broad, and LexisNexis' Future of Work Report 2024 found that generative AI is already used by 87% of professionals and is "shaping the future of work".

Our extensive news coverage, enriched with robust metadata, is readily available for integration into your generative AI projects. Over the past year, we have worked diligently and transparently with our publishers to secure the rights to use their data with generative AI tools. Our portfolio covers over 20,000 licensed titles, with thousands of sources available for use with generative AI technology.

H

Hallucinations – When an AI model generates a result or output which is not grounded in the training data or the prompt it was given. This is a particular risk of generative AI, and guardrails can be put in place to mitigate its risk – see RAG and Human Oversight.

Human Oversight – AI has enormous potential for companies, but it is important that staff members with expertise oversee the technology to help detect hallucinations, biases and other errors. 97% of professionals told the LexisNexis Future of Work survey that human validation of AI outputs is important.

I

Internet of Things – Interrelated computing devices, machines and physical objects that exchange or transfer data with each other over the internet. The term is commonly used to describe 'smart homes' in which thermostats, lighting and security cameras can be controlled by connected devices like smartphones.

L

Large Language Model (LLM) – A model that applies machine learning and deep learning to large datasets to understand and generate text. This is similar to generative AI, but more narrowly about text and language rather than images and other content forms.

M

Machine Learning – An application of AI in which computer systems are able to learn, adapt and improve through experience and without following express instructions. These systems use algorithms and statistical models to analyze patterns of data and draw insights.

Nexis Data+ empowers companies to leverage relevant datasets for machine learning, predictive analytics, and other big data applications.

Metadata – Data that describes and gives information about other data - known as "data about data". By summarizing basic information, it makes it easier to find and use the data.

N

Natural Language Processing – A type of AI concerned with the interactions between computers and the human language, particularly how to program computers to process and analyze large volumes of natural (ordinary) language data. The technology can 'understand' text documents, including nuances in the language, and accurately extract information and insights from them. NLP is an example of machine learning.

Neural Networks – A system of connected nodes like neural connections in the brain that are used as a method of machine learning. Connections between the layers lead to outputs and a prediction.

Normalization – The process of reorganizing data in different databases to make comparisons between the data easier and more meaningful.

P

Pattern Recognition – Identifying patterns in data, usually via algorithms, which allows predictions to be made when similar data is encountered.

Q

Quantitative Analysis – Using algorithms to find insights from large amounts of quantitative data. This is particularly useful in the financial sector, where trading decisions are often made by quantitative analysis of high volumes of numerical, financial data.

R

Retrieval Augmented Generation (RAG) – A technique to improve the output from a large language model or generative AI tool by bringing in an external, up-to-date and authoritative dataset to help shape the response to a user's query. This mitigates the risk of hallucinations.

As well as offering licensed data for generative AI solutions via our API, we employ a RAG approach in our generative AI tool, Nexis®+AI, to provide cited sources in our responses to companies' research requests.

Robotic Process Automation (RPA) – Software that is programmed to do repetitive and often mundane tasks. RPA deploys robots to improve efficiency and free human resources for more high value tasks. It can have a dramatic impact on productivity, efficiency, and accuracy within business processes, such as fraud detection and risk mitigation.

S

Semi-Structured Data – Data that does not have a structured format and cannot be contained in a database of rows and columns, but a hierarchy has been established using tags or other markers. Nexis Data+ allows for standard and flexible integration of a semi-structured XML data feed into any database or application.

SmartIndexing – A classification technology that helps researchers to find relevant information from large volumes of data by tagging documents. This is particularly useful for research.

T

Text Analytics – Deriving insight or meaning from text-based sources. This can be done by applying linguistics, machine learning and statistical techniques.

Training and Testing – This is a key part of the process of machine learning. A predictive model uses a set of traced data to build understanding, then it uses what it has learned to predict outcomes based on similar data.

U

Unstructured Data – Data that has not been organized in a pre-defined manner. It is often full of text, dates, numbers, and facts, and requires additional effort to make it useful.

X

XML – A way of tagging data to describe it.

Z

Zettabytes – A measurement for an enormous amount of data – bigger than an Exabyte and a Terabyte, but smaller than a Yottabyte. It is estimated that 181 Zettabytes of data will be generated in 2025. Data is being created at an exponential rate.

Conclusion

If this glossary has demystified some of the complex language behind AI and Big Data, why not talk to us about how LexisNexis can support your business to find new insights and manage risk?

In today's fast-paced business environment, maintaining ethical standards in AI is paramount to its successful implementation. Our powerful combination of credible, licensed content and sophisticated technology can transform your AI and generative AI initiatives and set you up for success.

Contact us today to learn more about how our trusted and ethical data and AI solutions can drive your business forward.

LexisNexis.com/**Data** or call **1-888-466-3947**

LexisNexis, Nexis and the Knowledge Burst logo are registered trademarks and Nexis Diligence is a trademark of RELX Inc. Other products or services may be trademarks or registered trademarks of their respective companies.
© 2024 LexisNexis. All rights reserved. US-DATA-Glossary-AI-and-Big-Data 1024

LexisNexis®