Deep Knowledge Group

Data Sources, Quality Assurance and Accuracy Overview

Deep Pharma Intelligence Dashboard

May 2024

1. General Overview

1.1. **Deep Knowledge Group's** primary competitive edge lies in its analytical approach to sourcing data. In pursuit of this objective, we employ a diverse range of algorithms for collecting and processing data. Our team of experts is continuously innovating, creating new algorithms, and enhancing existing ones to further bolster our capabilities.

1.2. Our data collection process employs a range of sophisticated algorithms designed to efficiently gather information from diverse sources. These algorithms are tailored to specific data types and sources, enabling comprehensive data retrieval across various domains.

1.3. By utilising advanced parsing techniques and leveraging **APIs, NLP** and **web scraping** technologies, we ensure the accuracy and reliability of the collected data. These algorithms play a crucial role in enabling us to access, extract, and organise vast amounts of data, empowering us to derive valuable insights and make informed decisions.

2. Data Sources, Filling, Quality Assurance and Accuracy

Database of Clinical trials with all the requested parameters and categorisation

No.	Column	Percentage	Reliability Percentage	Negative DataPoints
1	Heading	99%		-
2	Sponsor name and logo	0%	0%	10030(name + logo)
3	Start date	99%	100%	5
4	Conditions	74%	100%	1285
5	Agency	99%	100%	2
6	Country	0%		2347(country/city)
7	City	0%		2347(country/city)
8	Study Type	100%	100%	0
9	Phase	13.7%	100%	4325
10	Status	92%	100%	395
11	AI-driven	0%	0%	5015

12	Therapeutic Focus	0%	0%	5015
13	Page with link	93%	100%	289

List of companies (names and links), categorisation and labelling

No.	Column	Percentage	Reliability Percentage	Negative DataPoints
1	Name	100%	100%	-
2	Link	100%	100%	-
3	Therapeutic Focus	37%	100%	183
4	Product type	79%	62%	389
5	Research focus	60%	86%	786
6	Funding Status	-	-	-
7	IPO Status	-	-	-

R&D Hubs Database + General Parameters

No.	Column	Percentage	Reliability Percentage	Negative DataPoints
1	Name	100%	100%	0
2	Website link	100%	100%	0
3	Description	58%		211
4	Email Address	82%	86%	92
5	Social media links	-	-	-
6	LinkedIn	30%	76%	344
7	Twitter	31%	65%	341
8	Facebook	58%	59%	209
9	Instagram	0.6%	43%	496
10	Logo	0%	-	-

13	Year of establishment	0%	-	-

General Information on the received list of companies

No.	Column	Percentage	Reliability Percentage	Negative DataPoints
1	Name	100%	100%	0
2	Website link	99%	100%	0
3	Description	53%	93%	822
4	Email address	73%	100%	579
5	Social media links	-	-	-
6	LinkedIn	72%	74%	543
7	Twitter	51%	58%	866
8	Facebook	59%	12%	691
9	Instagram	11%	88%	1123
10	Logo	86%	95%	294
11	Location	-	-	-
12	Country	38%	100%	514
13	City	38%	100%	717
14	Year of establishment	45%	97%	1142

AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Amount of R&D collaborations	Number of active collaborations focused on research and development; number, list of collaborations	1533	75%	566	80%

		with corresponding sources				
2	Amount of partners	Total number of business partners across all market sectors; number, list of partners with corresponding sources	1685	50%	414	75%
3	Amount of partnerships	Count of formalized partnerships with other companies across all market sectors; number, list of companies with corresponding sources	1533	88%	566	82%
4	Amount of countries where company is presented	Number of countries where the company operates; number and the list of countries	2022	100%	77	0%
5	Average revenue of collaborators	Average revenue of public companies with which a target company collaborates; the list of companies and their revenue	484	100%	1615	-
6	Collaboration with Big Pharma companies	Amount of collaborations with Big Pharma companies (J&J, Pfizer, Roche, etc.); the list of Big Pharma companies with corresponding sources	729	80%	1370	-
7	Collaboration with	Amount of collaborations with Big	385	100%	1714	-

	Big Tech companies	Tech companies (Google, Microsoft, Nvidia, etc.); the list of Big Tech companies with corresponding sources				
8	Peer-reviewed papers published with Big Pharma	Amount of peer-reviewed papers published with Big Pharma companies (J&J, Pfizer, Roche, etc.); the list of publications, name of the journal, date of publication, authors, corresponding source	513	80%	1585	-
9	Papers published with Big Tech companies	Amount of papers published with Big Tech companies (Google, Microsoft, Nvidia, etc.); number of publications, the list of publications, name of the journal, date of publication, authors, corresponding source	552	80%	1546	-
10	Milestones with Big Pharma / Drugs developed	The number of drugs developed in collaboration with Big Pharma companies (the rights on the drug are shared between Big pharma and the company); number of drugs, the list of drugs with corresponding source	48	80%	2050	-

11	Collaboration with other AI in Drug Discovery	Number of collaborations with other AI in DD companies a target companies has; number of collaborations, the list of collaborations with corresponding source	748	100%	1351	-
12	Collaboration with women's health organisations	Number of collaborations with healthcare institutions and organisations specialising in women's health to enhance their product development and service offerings; number and a list with corresponding sources	211	100%	1888	-
13	Amount of collaborations with clinics	How many collaborations with clinics and medical organisations does a company have; number and a list of clinics/medical organisations	397	100%	1701	-
14	Amount of collaborations with consumer electronics	Number and a list of consumer electronics companies	406	100%	1692	-

Product Development Data Points for AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	*Total amount of products*	*Total amount of products a company provides to the market; number, the list of products*	*1030*	*90%*	*1068*	*90%*
2	*Percent of MVPs*	*MVP; percentage*	*506*	-	*1593*	-
3	*Amount of Drug Design stages are covered*	*How many stages (Preclinical Research, IND Application, Clinical Trials, NDA Submission, Regulatory Review, Approval and Post-Marketing Surveillance, Phase 4 Trials) end-to-end, a company covers; number and the list of stages*	*837*	*80%*	*1262*	*80%*
4	*Amount of active pipelines*	*How many active Drug Discovery pipelines a company has; number, the list of pipelines (drug names) with links*	*1301*	*80%*	*797*	*100%*
5	*Number of ongoing clinical studies*	*Number of ongoing clinical studies for the company's products; number of studies, the list with links*	*358*	*50%*	*1740*	*90%*
6	*Number of completed clinical studies during last 5 years*	*Number of completed clinical studies during last 5 years; number of studies, the list with links*	*258*	*50%*	*1840*	-

7	*Number of clinical studies that will be completed in the next 5 years*	*Number of clinical studies that will be completed in the next 5 years; number of studies, the list with links*	*358*	*50%*	*1740*	-
8	*Number of terminated clinical trials*	*Number of terminated clinical trials; number of studies, the list with links*	*117*	*50%*	*1981*	-
9	*Average phase of completed clinical trials*	*Early Phase 1 to Phase 4, Not Applicable; text*	*521*	*30%*	*1577*	-
10	*Amount of intervention models*	*The general design of the strategy for assigning interventions in clinical studies; text*	*394*	*65%*	*1704*	*90%*
11	*Average number of participants in clinical trials*	*Average number of participants in the clinical studies; number*	*403*	*0%*	*1695*	-%
12	*Amount of conditions covered by clinical trials ever conducted*	*The disease, disorder, syndrome, illness, or injury that is being studied; number with the list of conditions*	*433*	*90%*	*1665*	*100%*
13	*Amount of interventions in ongoing clinical trials*	*A process or action that is the focus of a clinical study; number with the list of interventions*	*238*	*90%*	*1860*	-

14	*Amount of approved drugs and therapies*	*How many drugs and therapies were approved by FDA/EMA; number with a list of names*	*59*	*100%*	*2040*	-	
15	*Amount of AI drugs that entered clinical trials*	*The number of drugs with a drug molecule or a drug target predicted using in silico AI tools that entered clinical trials; number with a list*	*2*	-	*2096*	-	
16	*Number of ongoing clinical trials for early diagnostic tools*	*Quantity of clinical trials conducted to test new diagnostic tools for early detection of diseases or conditions; number*	*36*	*100%*	*2062*	*0%*	
17	*Number of diseases/therapeutic areas of early diagnostics*	*Number, list of diseases/therapeutic areas subjected to early diagnostics by a company*	*964*	-	*1135*	-	
18	*Integration with electronic health record (EHR) systems*	*Company's ability to connect its healthcare-related products or services with EHRs; number of EHR software licences purchased*	*52*	-	*2047*	-	
19	*Number of FDA-approved products in the women's health tech market*	*Quantity of medical devices or products for women's health approved by the FDA; number and a list of names*	*2*	*100%*	*2104*	-	

20	*Regulatory approvals for medical technologies*	*Number of FDA approved medical devices; number and a list of names*	*64*	*100%*	*2042*	*90%*
21	*Number of genetic and molecular tests developed*	*Number of genetic and molecular diagnostic tests available in Personalised Medicine; number and a list*	*104*	*100%*	*1995*	-
22	*Number of FDA or regulatory approvals for diagnostic tests using biomarkers*	*Number of FDA-approved biomarker tests; number and list*	*12*	*100%*	*2094*	-
23	*Number of disease/conditions for which biomarkers are found*	*Number, list of disease/conditions*	*77*	*80%*	*2022*	-
24	*Integration of AI*	*Does the company use AI for product development and in what capacity; count the number of distinct AI applications or algorithms in drug discovery processes. Number (low threshold)*	*510*	*100%*	*1589*	*0%*
25	*Amount of therapeutic areas*	*How many therapeutic areas a companies products targets; number for each product and a list*	*1231*	*100%*	*868*	*100%*

26	*Amount of users*	*How many people use a service, number of users*	*529*	*100%*	*1570*	*0%*
27	*Amount of partnerships with clinics*	*How many partnerships with clinics a company has; number and a list of partners names*	*1071*	*100%*	*1028*	*90%*
28	*Number of conditions*	*Number of conditions for which company's products were approved to use; number and a list of conditions*	*49*	*100%*	*2050*	*90%*
29	*Accessibility*	*a percentage of company products that can be easily purchased (no prescription needed, off the shelf) (low threshold)*	*1868*	-	*231*	-

Financial Position Data Points of AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Number of subsidiary companies	The sum of all the subsidiary companies; number, the list of the subsidiary companies with corresponding sources	146	47%	1953	78%
2	Number of investors	The sum of all the investors that invested in the company; number of investors, the list of investors with corresponding sources	1198	85%	901	85%

3	Annual revenue	The amount of money that a company actually receives during the last year; number, currency	86	73%	2013	30%
4	Funding status	Indicates what type of funding a company received last time; text	1378	70%	722	37%
5	Funding velocity	% change of volume of funding across successive funding rounds; percentage	549	-	1550	-
6	Last funding amount	Last funding amount raised; number, currency symbol	943	95%	1156	95%
7	Total funding amount	Total funding amount raised across all funding rounds; number; currency symbol	1035	90%	1064	90%
8	Series Funding Rounds	The number of Series Funding Rounds the company got after the first seed funding; for each funding round: funding type, funding date, money raised (number, currency symbol), investor(s) name(s).	581	-	1518	*41%*
9	Amount of investments	The sum of all investments a target company made; number, currency symbol)	113	-	1986	-
10	Grant		160	-	1940	-

Investors Database of AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Website link	URL address; link	4337	88%	253	79%
2	Name	Name of the investor; text	4530	100%	0	100%
3	Investor Description	3 to 5 sentences about the investor (at least 100 words), briefly presenting the main goals of the investor, its scope of services, and industry; text	4538	100%	0	100%
4	Logo/Photo	Logo of the investor in high quality if it's a company and photo if it's a person, preferably in png/jpeg format; png/jpeg picture	2820	95%	2124	31%
5	Location	Name of country and city of the headquarters, also country/city for any other locations if present; text	5639	81%	0	100%
6	Company Invested	List of AI in Biopharma Companies they invested in; text	2868	-	2771	-
7	Amount	Amount of money in USD they invested for each investment round; number	948	-	4591	-
8	Year of	Year of establishment	5276	100%	363	91%

	establishment	when investor was founded; date				
9	Social Media links	Links to the investor's accounts on LinkedIn, Instagram, Facebook, Twitter; links	Twitter: 2772 Facebook: 1915 LinkedIn: 3810	88%	Twitter: 2867 Facebook: 3724 LinkedIn: 873	63%
10	Email address	Investor email address; text	3497	100%	2142	68%
11	Invested in GCC	Indicates if they invested in a Gulf Region company from a database; boolean	5639	-	0	-
12	Located in GCC	Indicates if they are located in the Gulf Region; boolean	5639	-	0	-

Experts Database in the field of AI in Biopharma

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Name	Name of the expert; text	755	90%	0	100%
2	Description	The main information about the expert; text (3-5 sentences)	755	100%	0	100%
3	H-index	Hirsch index (if applicable); numeric value	58	100%	687	100%
4	Leader Category Label	Entrepreneur/Research & Academia/Investor; text	755	100%	0	100%

Collaborations Database

No.	Column	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	**URL**	1201	100%	0	0%
2	**Partner1**	1570	100%	0	0%
3	**Partner2**	1569	100%	1	0%
4	**Date**	1115	71%	455	0%
5	**Capital**	286	100%	1284	0%
6	**Description**	1567	100%	0	0%

Intellectual property Data Points of AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Number of other IP assets	Overall number of other assets (Trademarks, licences, copyrights, etc., any IP apart from patents)	1652	100% (is not checked)	447	100% (is not checked)
2	Trademark	Registered trademark of the company or not; boolean value	566	100%	1533	50%
3	Business Licences	List of jurisdictions where the company could perform business; text	2022	100%	77	0%
4	Average expiration time	When, on average, do the patents of a company expire?; date	71	100% (is not checked)	2028	20%

Marketing Data Points of AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Amount of organised online events	How many online events a company organised; number of events, the list of events with corresponding sources	187	100%	1911	91%
2	Amount of organised offline events	How many offline events a company organised; number of events, the list of events with corresponding sources	486	100%	1612	92%
	Amount of subscribers at Linkedin	How many subscribers at Linkedin a company has as of date of search; number, date of search	1699	100%	447	23%
4	Amount of subscribers at X	How many subscribers at X a company has; number, date of search	1059	100%	1071	32%
5	Amount of mentions in news articles	How many times a company was mentioned in the news by the date of search; number, date of search	2000	100%	98	100%

Team Composition Data Points of AI in Biopharma Companies

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Amount of	How many people	1604	68%	494	73%

	co-founders	co-founded a company; number				
2	Average experience in the field of co-founders	What is average years of experience of co-founders in the field; the list of names with corresponding number of years, average number	1604	*100%*	494	*100%*
3	Average years co-founders worked at BigTech/BigPharma companies	How long on average co-founders worked at Big Tech/Big Pharma companies; the list of names with corresponding number of years, average number	1541	*100%*	557	*100%*

Companies data:

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Companies Descriptions	3 to 5 sentences about the company (at least 100 words), the description must present briefly the main goals of the company, it's scope of services and industry; text *53% of descriptions are already collected, the remaining number of required data points is indicated in the table.	1914	80%	184	0%

2	Companies Locations	name of country and city of the headquarter, also country/city for any other locations if present; text *38% of locations are already collected, the remaining number of required data points is indicated in the table.	1639	70%	459	10%
3	R&D Hubs Descriptions	R&DHubs Descriptions: 3 to 5 sentences about the company (at least 100 words), the description must present briefly the main goals of the company, it's scope of services and industry; text *58% of descriptions are already collected, the remaining number of required data points is indicated in the table.	495	100%	5	100%
4	R&D Hubs Classification	assign labels from the list: Start-up Incubator Collaborative Research Consortia Government Research Facilities Innovation Platform Academic Research Center Hubs can have more than one label from each category.	494	90%	6	100%
5	R&D Hubs		498	100%	2	100%

	Locations					

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Companies' product development metrics	Amount of articles about company products: How many articles mentioned products of a company; number; the list of articles with links. *26% of data is already collected, the remaining number of required data points is indicated in the table. Data needs to be proved because of some unnatural values.	1938	87%	160	85%
2	Companies' product development metrics	Amount of App downloads: How many downloads a company products has in total; total number, the list of app names with links.	117	83%	1982	80%
3	Companies' Financial position metrics (ticker to be provided)	IPO status: Private/Public, money raised on IPO (number, ticker symbol). *37% of data is already collected, the remaining number of required data points is indicated in the table.	163	100%	1936	100%

4	For experts database	Entrepreneurs (Companies' CEO's and CMO's): Name Currant workplace Current position Country Description (3-5 sentences) Image (high quality, the face is visible, no frames) - png/jpeg Links to websites and media: personal website, researchgate, google scholar, linkedin, instagram, facebook, twitter Email address *65% of data is already collected, the remaining number of required data points is indicated in the table	1004	100%	0	100%
5	CEO	Name of the current CEO of the company, provide social media links and email address where available; name, links to LinkedIn, X, Facebook, Instagram, email	1603	100%	495	78%
6	Founder/co-Founders	Name of company founder or co-founders of the company, provide social media links and email address where available; name, links to LinkedIn, X, Facebook, Instagram, email	1604	100%	494	78%
7	Number of employees	Number of employees in a target company; number	1644	100%	434	40%

No.	Column	Description	Positive Share	Positive Reliability	Negative Share	Negative Reliability
1	Companies' Intellectual property: Number of patents	Number of patents and Amount of patents categories: Number of active patents; number and a list with corresponding sources; number of different patent categories.	650	100	1449	100
2	Companies' Intellectual property: Amount of patents categories	Number of patents and Amount of patents categories: Number of active patents; number and a list with corresponding sources; number of different patent categories	650	100	1449	100

3. Overall Data Filling, Quality Assurance and Accuracy

Positive Objects - non-zero and non-N/A
Negative Objects - zero objects and N/A
Negative Reliability = share of zero-valued Negative Objects, TN/(FN+TN) - correct zero and N/A for randomly selected subset
Positive Reliability = share of non-zero-valued and non-NA, TP/(FP+TP) - correct non-zero and non-N/A for randomly selected subset
(1 - TN/(FN+TN))
For entity of list = T(entity list) / (F(entity list) + T(entity list))
For n entities = (n * (T(entity list) / (F(entity list) + T(entity list)))) / n * 100
E - overall entitis = P + N
Qualified Data Points = P * Reliability (P) + N * Reliability (N)

TRUE DATA (1% dataset) = P * Reliability (P)
TRUE FALSE (1% dataset) = N * Reliability (N)
SUMM = TRUE DATA (1% dataset) + TRUE FALSE (1% dataset)

4. Data Sources

4.1. Bing Answer + NER (Named Entity Recognition) is a structured process for gathering specific data entities from various online sources mainly from google and bing. This methodology combines search engine capabilities and advanced entity recognition techniques to extract and verify relevant information. Bing Answer + NER provides a robust framework for collecting and verifying a wide range of data entities. By integrating search engine capabilities with advanced NER methods, this approach ensures comprehensive and reliable data extraction across various domains.

4.2. The process of collecting logos takes place by making a request to google, bing search engines and obtaining the URL of the image according to the entered information, downloading the image of the logo from the appropriate source or resource using the URL of the image, and checking whether the image corresponds to the required formats (eg JPEG, PNG, etc.). Validation of process images using NNtool validation classification.

4.3. Also was used such open sources as ClinicalTrials.gov, PubMed, Semantic Scholar, Companies websites, Bing, Yahoo Finances, LinkedIn, Google Patents, USPTO, Twitter, SEMrush

5. Data Classification

5.1. At DKG, we specialise in advanced text preprocessing solutions that optimise data analysis processes for businesses across industries. Our expertise lies in refining raw text data, extracting key objects, and ensuring data integrity through thorough cleaning and structuring. Text pre-processing is an important step in the data analysis pipeline, as it lays the foundation for accurate and insightful conclusions. By cleansing the data and extracting relevant objects, we reduce noise and improve the quality of information, thereby contributing to more effective decision-making and strategic planning for our clients.

5.2. Our text pre-processing services offer several advantages:

5.2.1. Improved data quality: We eliminate inconsistencies, errors, and redundancies in text, resulting in cleaner, more reliable datasets.

5.2.2. Improved object recognition: With advanced algorithms and natural language processing techniques, we excel at identifying and extracting key objects such as dates, names, locations, organisations, and more, enriching the dataset with valuable contextual information.

5.3. **Optimized analysis**: By organising and structuring text according to predefined criteria, we simplify the data analysis process, saving time and resources while maximising the depth of insights gained from the data.

5.4. **Increased accuracy**: Our rigorous pre-processing methods ensure that data is properly formatted and standardised, minimising the risk of misinterpretation and errors during analysis.

5.5. **Entity recognition and extraction**: To extract named entities such as dates, names, locations, organisations, etc., we use models trained specifically for named entity recognition (NER). These models are typically based on deep learning architectures such as bidirectional LSTMs (long short-term memory networks) or transformer-based models such as BERTs (bidirectional encoder representations of transformers).

5.6. **GTP prompt classification**: Regarding the conditions of classification, it is assumed that the framework of the relation of classes, sub-classes, etc. is formed, the classification is one-way or multi-way. For each specific task of classification, the optimal GTP prompt is formed, and the corresponding filters are formed to form a structured answer.

6. Entity Recognition

6.1. Deep Knowledge Group (DKG) has spearheaded the development of advanced Entity Recognition algorithms, revolutionising our ability to extract validated data from vast quantities of unstructured sources. These algorithms represent a breakthrough in data processing, enabling us to efficiently identify and validate entities within text and images, thereby enhancing the quality and reliability of our dataset.

6.2. Our Entity Recognition algorithms are designed to tackle the inherent challenges posed by unstructured data, such as news articles, reports, social media posts, and other textual sources. Leveraging state-of-the-art natural language processing (NLP) techniques, machine learning models, and semantic analysis, these algorithms can accurately identify and extract entities, including entities such as organisations, people, locations, events, and more. By discerning context, relationships, and relevance, our algorithms ensure that extracted entities are not only identified but also validated for accuracy and reliability.

6.3. In addition to text, our Entity Recognition algorithms extend to image data, enabling us to validate and extract entities from visual content with precision and efficiency. Through advanced image recognition and pattern recognition techniques, we can identify objects, logos, landmarks, and other visual elements, enriching our dataset with valuable insights derived from multimedia sources.

6.4. Furthermore, our data enrichment and classification validation processes are augmented by querying search engines, leveraging their vast repositories of indexed information to corroborate and validate extracted entities. By cross-referencing data obtained from multiple

sources, we enhance the accuracy and completeness of our dataset, ensuring that only validated information is incorporated into our database.

6.5. A key aspect of our Entity Recognition framework is the integration of a data validator, which provides continuous evaluation and feedback to the system. This feedback loop enables our algorithms to learn and adapt over time, incorporating new patterns, insights, and corrections to improve their accuracy and performance. By leveraging machine learning techniques, the system can dynamically adjust its algorithms based on real-world feedback, ensuring that the database remains clean, updated, and reflective of the most current information available.

7. Advantages of Utilising APIs, NLP and Web Scraping Technologies

7.1. **Efficiency**: By employing APIs, NLP, and web scraping, we optimise the data collection process, reducing manual efforts and enhancing efficiency. These technologies automate data retrieval tasks, allowing our team to focus on analysis and decision-making rather than data acquisition.

7.2. **Accuracy**: The use of advanced parsing techniques and structured data access through APIs ensures the accuracy and reliability of the collected data. NLP algorithms enhance data understanding and interpretation, enabling us to extract meaningful insights with precision.

7.3. **Comprehensive Data Coverage**: By combining APIs, NLP, and web scraping technologies, we achieve comprehensive data coverage across various sources and domains. This approach enables us to gather data from diverse sources, including structured databases and unstructured web content, facilitating holistic analysis and informed decision-making.

7.4. **Real-Time Access**: APIs provide real-time access to data, ensuring that our analyses are based on the latest information available. This real-time data access enhances the timeliness and relevance of our insights, enabling agile responses to changing market conditions and emerging trends.

7.5. **Flexibility**: The flexibility of web scraping allows us to extract data from any website, regardless of whether an API is available. This flexibility enables us to adapt to evolving data sources and requirements, ensuring continuity and adaptability in our data collection efforts.

7.6. **Innovation**: Deep Knowledge Group's commitment to innovation drives continuous improvement in our algorithms and technologies. By staying at the forefront of advancements in APIs, NLP, and web scraping, we continually enhance our capabilities and maintain a competitive edge in data analytics and decision support.

8. Data Quality Control, Accuracy and Lawfulness

8.1. Data quality control and accuracy are paramount in Deep Knowledge Group's (DKG) analytical endeavours, particularly given the diverse array of data sources we rely on. Our commitment to excellence begins with the careful selection of source material, drawn from a multitude of reputable outlets, including news articles, press releases, organisation websites, scientific publications, patent databases, government websites, and various open sources encompassing a wide spectrum of information.

8.2. Recognising the inherent variability and potential biases across different data sources, we prioritise independent data from external sources outside our organisation. This approach not only enriches our dataset with diverse perspectives but also serves as a foundational principle for ensuring the reliability and credibility of our analyses. By cross-validating each data point across multiple sources whenever possible, we mitigate the risk of inaccuracies and inconsistencies, thereby bolstering the overall quality and trustworthiness of our database.

8.3. The process of data validation and quality control at DKG is meticulous and thorough. We employ a multifaceted approach that involves rigorous scrutiny of each data point, leveraging advanced algorithms, machine learning techniques, and human expertise to assess its validity, relevance, and reliability. Through automated checks and manual verification procedures, we identify and weed out low-quality, outdated, or questionable data, ensuring that only the most accurate and up-to-date information is retained in our database.

8.4. Furthermore, we recognise that maintaining data accuracy is an ongoing endeavour that requires constant vigilance and proactive measures. As such, we have implemented robust protocols for continuous monitoring, validation, and updates, enabling us to promptly address any discrepancies, errors, or changes in the underlying data sources. Whether it's through regular audits, real-time alerts, or feedback mechanisms from users and stakeholders, we remain vigilant in our efforts to uphold the integrity and accuracy of our data.

8.5. In addition to stringent quality control measures, we also prioritise transparency and accountability in our data management practices. We provide clear documentation regarding the sources, methodologies, and assumptions underlying our analyses, allowing users to assess the reliability and credibility of the insights derived from our data. Moreover, we actively engage with stakeholders to solicit feedback, address concerns, and refine our data collection and validation processes, thereby fostering a culture of continuous improvement and trust.

8.6. From a **legal** standpoint, we rely on certain legal exclusions provided by various legislation in the UK, EU and US concerning data collection and mining. In particular, based on the jurisdiction and aims of data collection, its further use, we rely on temporary reproduction in the UK and EU, the text and data mining exception in the EU, and fair use in the US.

8.7. At each stage, specific legal criteria receive and filter each individual data source:

8.7.1. Access to the content: Initial access to the content is subject to stringent legal scrutiny to ensure compliance with relevant regulations and standards;

8.7.2. Extraction and copying of content: The process of extracting and copying content is conducted within the bounds of applicable laws, with careful attention paid to intellectual property rights and data protection considerations;

8.7.3. Text and data mining: Text and data mining activities are conducted in accordance with the provisions and exceptions afforded by relevant legislation, balancing the imperative of innovation with the imperative of safeguarding privacy and intellectual property rights;

8.7.4. Data storage: The storage of mined data is executed in compliance with established data protection regulations, encompassing measures aimed at safeguarding the confidentiality, integrity, and availability of the data.

8.8. Furthermore, our adherence to privacy regulations is unwavering, with a particular focus on the EU General Data Protection Regulation, the UK General Data Protection Regulation, and the Data Protection Act 2018. With each occurrence of personal data collection, we establish legal grounds and implement a comprehensive array of supplementary measures. For instance, we provide users with expedited mechanisms such as quick contact forms for the deletion or rectification of personal data, ensuring that individuals retain agency over their personal information within the parameters of the law.

8.9. We diligently observe and abide by AI regulations promulgated by regulatory bodies such as the European Commission and the UK government. Specifically, we adhere to the EU Regulation on Artificial Intelligence (AI Act) and the UK's AI Regulatory Framework, which provide comprehensive frameworks for the ethical and lawful deployment of AI technologies, including those involved in data collection and mining endeavours.

8.10. Our AI regulatory compliance efforts encompass several key facets:

8.11. **Transparency and Accountability**: We prioritise transparency in AI-driven data collection processes, ensuring that individuals are informed about the nature, scope, and implications of data collection activities. Moreover, we uphold principles of accountability, assuming responsibility for the ethical and lawful utilisation of AI technologies in data mining endeavours.

8.12. **Fairness and Non-discrimination**: In accordance with regulatory mandates, we uphold principles of fairness and non-discrimination in AI-driven data collection practices. We employ measures to mitigate biases and ensure equitable treatment across diverse demographic groups, thereby fostering trust and inclusivity in our data collection initiatives.

8.13. **Data Protection and Privacy**: As indicated above we prioritise the safeguarding of individuals' privacy rights and personal data integrity throughout the data collection lifecycle.

8.14. **Ethical Considerations**: We remain attuned to ethical considerations inherent in AI-driven data collection and mining activities, striving to uphold principles of beneficence, autonomy, and justice in our practices. Our ethical framework guides decision-making processes, ensuring alignment with societal values and ethical norms.