Questions	Answers
1. What is CRISP-DM?	Cross-industry standard process for data mining, known as CRISP-DM, is an open standard process model that describes common approaches used by data mining experts. It is the most widely-used analytics model. CRISP-DM breaks the process of data mining into six major phases: Business Understanding,Data Understanding,Data Preparation,Modeling,Evaluation,Deployment.The sequence of the phases is not strict and moving back and forth between different phases as it is always required. The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram symbolizes the cyclic nature of data mining itself. A data mining process continues after a solution has been deployed. The lessons learned during the process can trigger new, often more focused business questions, and subsequent data mining processes will benefit from the experiences of previous ones.(Anon, 2021d)
2. How to understand “Decision Tree”?	A decision tree is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal, but are also a popular tool in machine learning.(Anon, 2021g)
3. What are the steps to establish a decision tree?	Drawn from left to right, a decision tree has only burst nodes (splitting paths) but no sink nodes (converging paths). Therefore, used manually, they can grow very big and are then often hard to draw fully by hand. Traditionally, decision trees have been created manually –although increasingly, specialized software is employed. Decision rules: The decision tree can be linearized into decision rules, where the outcome is the contents of the leaf node, and the conditions along the path form a conjunction in the if clause. In general, the rules have the form: if condition1 and condition2 and condition3 then outcome.(Anon, 2021g)

4. What is data mining?	Data mining is a process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.(Anon, 2021f)
5. What are the common data mining techniques?	Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.Association rule learning (dependency modeling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".Regression – attempts to find a function that models the data with the least error that is, for estimating the relationships among data or datasets.Summarization – providing a more compact representation of the data set, including visualization and report generation.(Anon, 2021f)
6. What are the common data mining tools?	RapidMiner : RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for data preparation, machine

	learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, model validation and optimization. RapidMiner is developed on an open core model.(Anon, 2021k) Weka: A suite of machine learning software applications written in the Java programming language. Advantages of Weka include:Free availability under the GNU General Public License.Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.A comprehensive collection of data preprocessing and modeling techniques.Ease of use due to its graphical user interfaces.(Anon, 2021m) R: R is a programming language and free software environment for statistical computing and graphics. It is supported by the R Core Team and the R Foundation for Statistical Computing. It is widely used among statisticians and data miners for developing statistical software and data analysis.(Anon, 2021j) NLTK: The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing (NLP) for English written in the Python programming language. NLTK includes graphical demonstrations and sample data. It is accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit, plus a cookbook.(Anon, 2021i) Google Cloud Platform (GCP), offered by Google, is a suite of cloud computing services that runs on the same infrastructure that Google uses internally for its end-user products, such as Google Search, Gmail, Google Drive, and YouTube. Alongside a set of management tools, it provides a series of modular cloud services including computing, data storage, data analytics and machine learning. Registration requires a credit card or bank account details.(Anon, 2021h)
7. What are the methods of data analysis?	Several analyses can be used during the initial data analysis phase: 1.Univariate statistics (single variable)2.Bivariate associations (correlations)3.Graphical techniques (scatter plots)It is important to take the measurement levels of the

	variables into account for the analyses, as special statistical techniques are available for each level: 1.Nominal and ordinal variables1.1Frequency counts (numbers and percentages)1.2Associations1.2.1circumambulations (crosstabulations)1.2.2hierarchical loglinear analysis (restricted to a maximum of 8 variables)1.2.3loglinear analysis (to identify relevant/important variables and possible confounders)1.3Exact tests or bootstrapping (in case subgroups are small)1.4Computation of new variables2. Continuous variables2.1Distribution2.1.1Statistics (M, SD, variance, skewness, kurtosis)2.1.2Stem-and-leaf displays2.1.3Box plots3.Nonlinear analysis.(Anon, 2021e)
8. What are the main functions of data mining?	Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.1.Association rule learning (dependency modeling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.2.Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.3.Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".4.Regression – attempts to find a function that models the data with the least error that is, for estimating the relationships among data or datasets.5.Summarization – providing a more compact representation of the data set, including visualization and report generation.(Anon, 2021f)
9. What is web mining?	Web mining is the application of data mining techniques to discover patterns from the World Wide Web. It uses automated methods to extract both structured and unstructured data from web pages, server logs and link structures. There are three main sub-categories of web mining. Web content mining

	extracts information from within a page. Web structure mining discovers the structure of the hyperlinks between documents, categorizing sets of web pages and measuring the similarity and relationship between different sites. Web usage mining finds patterns of usage of web pages.(Anon, 2021l)
10. What can web mining dig?	1. Get competitor and customer information. 2. Discover user access patterns.3. Anticompetitive intelligence activities.(Anon, 2021l)
11. What are the types of web mining?	Web mining can be divided into three different types – Web usage mining, Web content mining and Web structure mining.Web usage mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.Web structure mining uses graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided into two kinds:Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content.(Anon, 2021l)
12. What are the core technologies for big data?	Big data acquisition, big data preprocessing, big data storage, and big data analytics make up the core technologies in the big data lifecycle.(Anon, 2021b)
13. What the base Apache Hadoop framework includes?	Hadoop Common – contains libraries and utilities needed by other Hadoop modules;Hadoop Distributed File System (HDFS) – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster;Hadoop YARN – (introduced in 2012) a platform responsible for managing computing resources in clusters and using them for scheduling users' applications; Hadoop MapReduce – an implementation of the MapReduce

	programming model for large-scale data processing.Hadoop Ozone – (introduced in 2020) An object store for Hadoop.(Anon, 2021a)
14. What is Hadoop?	Apache Hadoop is a collection of open-source software utilities that facilitates using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage and processing of big data using the MapReduce programming model. Hadoop was originally designed for computer clusters built from commodity hardware, which is still the common use. It has since also found use on clusters of higher-end hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.(Anon, 2021a)
15. What is Hadoop distributed file system?	The Hadoop distributed file system (HDFS) is a distributed, scalable, and portable file system written in Java for the Hadoop framework. Some consider it to instead be a data store due to its lack of POSIX compliance, but it does provide shell commands and Java application programming interface (API) methods that are similar to other file systems. A Hadoop instance is divided into HDFS and MapReduce. HDFS is used for storing the data and MapReduce is used for processing data.(Anon, 2021a)
16. What are the services provided by HDFS?	Name Node: Namer node can track files, manage the file system and has the metadata of all of the stored data within it.Secondary Name Node: his is only to take care of the checkpoints of the file system metadata which is in the Name Node. Job tracker: Job Tracker receives the requests for Map Reduce execution from the client. Job tracker talks to the Name Node to know about the location of the data that will be used in processing. The Name Node responds with the metadata of the required processing data.Data Node: A Data Node stores data in it as blocks. This is also known as the slave node and it stores the actual data into HDFS which is responsible for the client to read and write. Task Tracker: It is the Slave Node for the Job Tracker and it will take the task from the Job Tracker. It

	also receives code from the Job Tracker. Task Tracker will take the code and apply on the file. The process of applying that code on the file is known as Mapper.(Anon, 2021a)
17. What are the features of Cloud Computing?	Cloud computing shares characteristics with:Client–server model—Client–server computing refers broadly to any distributed application that distinguishes between service providers (servers) and service requestors (clients). Computer bureau—A service bureau providing computer services, particularly from the 1960s to 1980s.Grid computing—A form of distributed and parallel computing, whereby a 'super and virtual computer' is composed of a cluster of networked, loosely coupled computers acting in concert to perform very large tasks.Fog computing—Distributed computing paradigm that provides data, compute, storage and application services closer to the client or near-user edge devices, such as network routers. Furthermore, fog computing handles data at the network level, on smart devices and on the end-user client-side (e.g. mobile devices), instead of sending data to a remote location for processing.Mainframe computer—Powerful computers used mainly by large organizations for critical applications, typically bulk data processing such as census; industry and consumer statistics; police and secret intelligence services; enterprise resource planning; and financial transaction processing.Utility computing—The "packaging of computing resources, such as computation and storage, as a metered service similar to a traditional public utility, such as electricity." Peer-to-peer—A distributed architecture without the need for central coordination. Participants are both suppliers and consumers of resources (in contrast to the traditional client-server model).Green computing—Study and practice of environmentally sustainable computing or IT.Cloud sandbox—A live, isolated computer environment in which a program, code or file can run without affecting the application in which it runs.(Anon, 2021c)
18. What courses are offered for Master Degree?	Knowledge Representation and Reasoning,Machine Learning, Artificial Intelligence, Big Data Systems, Data Science, Programming for Data Science, Data Mining and Text Analytics, Cloud Computing, Advanced Software Engineering, Scientific

	Computation, Foundations of Modelling and Rendering, Geometric Processing, High-Performance Graphics, Animation and Simulation.(Anon, n.d.)
19. Can you provide the tutors' information to me?	Of course. You can learn about the tutor's information through the following website: https://eps.leeds.ac.uk/computing/stafflist. (Anon, n.d.)
20. How do I view basic information about my majors?	We offer six postgraduate majors:Advanced Computer Science MScAdvanced Computer Science (Cloud Computing) MSc Advanced Computer Science (Data Analytics) MSc Advanced Computer Science (Artificial Intelligence) MSc Data Science and Analytics for Health Mres High-Performance Graphics and Games Engineering MScYou can find basic information on this website：https://eps.leeds.ac.uk/computing-masters.(Anon, n.d.)

For the answers in the above, the references are listed below(Since the above answer comes from many encyclopaedia entries, there is no author name, and it is Anon):

Anon 2021a. Apache Hadoop. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Apache_Hadoop&oldid=1053186958.

Anon 2021b. Big data. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Big_data&oldid=1059261866.

Anon 2021c. Cloud computing. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Cloud_computing&oldid=1059266571.

Anon 2021d. Cross-industry standard process for data mining. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Cross-industry_standard_process_for_data_mining&oldid=1055377187.

Anon 2021e. Data analysis. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Data_analysis&oldid=1055861488.

Anon 2021f. Data mining. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Data_mining&oldid=1059125769.

Anon 2021g. Decision tree. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Decision_tree&oldid=1058119806.

Anon 2021h. Google Cloud Platform. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Google_Cloud_Platform&oldid=1059080772.

Anon 2021i. Natural Language Toolkit. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Natural_Language_Toolkit&oldid=1026979174.

Anon n.d. People | School of Computing | University of Leeds. [Accessed 8 December 2021a]. Available from: https://eps.leeds.ac.uk/computing/stafflist.

Anon 2021j. R (programming language). *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=R_(programming_language)&oldid=1058424566.

Anon 2021k. RapidMiner. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=RapidMiner&oldid=1048649117.

Anon n.d. Research and innovation | School of Computing | University of Leeds. [Accessed 8 December 2021b]. Available from: https://eps.leeds.ac.uk/computing-research-innovation.

Anon 2021l. Web mining. *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Web_mining&oldid=1045158979.

Anon 2021m. Weka (machine learning). *Wikipedia*. [Online]. [Accessed 8 December 2021]. Available from: https://en.wikipedia.org/w/index.php?title=Weka_(machine_learning)&oldid=1051745380.