## Principali informazioni sull'insegnamento

| | |
|---|---|
| Denominazione dell'insegnamento | **Big Data** |
| Corso di studio | Computer Science (Masters' degree in Computer Science) |
| Anno Accademico | 2023/24 |
| Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS) | 6 CFU |
| Settore Scientifico Disciplinare | INF/01 |
| Lingua di erogazione | inglese |
| Anno di corso | Secondo |
| Periodo di erogazione | 1^ semestre, le date esatte sono riportate nel manifesto/regolamento |
| Obbligo di frequenza | La frequenza è fortemente raccomandata |
| Sito web del corso di studio | https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/computer-science/computer-science |

## Docente/i



| | |
|---|---|
| Nome e cognome | Michelangelo Ceci |
| Indirizzo mail | michelangelo.ceci@uniba.it |
| Telefono | +39 080 5442285 |
| Sede | Dipartimento di Informatica, Via Orabona 4, 70125, Bari. Stanza n.510, 5^ piano. |
| Sede virtuale | Piattaforma ADA - https://elearning.uniba.it/ |
| Sito web del docente | http://www.di.uniba.it/~ceci/ |
| Ricevimento (giorni, orari e modalità, es. su appuntamento) | In presenza presso lo studio del docente: Giovedi, 9:00-11:00, preferibilmente su appuntamento |



## Syllabus

| | |
|---|---|
| **Obiettivi formativi** | Conoscenza e comprensione<br>L'analisi dei dati ha sostituito l'acquisizione dei dati come collo di bottiglia per il processo decisionale basato sull'evidenza. Ciò implica che l'estrazione di conoscenza da set di dati grandi, eterogenei e rumorosi richiede non solo potenti risorse di calcolo, ma anche basi metodologiche e astrazioni di programmazione appropriate. Le astrazioni emerse nell'ultimo decennio fondono idee da database paralleli, sistemi distribuiti e linguaggi di programmazione per creare una nuova classe di piattaforme di analisi dei dati scalabili che costituiscono la base per la scienza dei dati su scala reale. Nel corso verrà fornita una panoramica dei principi e tecnologie rilevanti, sia in termini di organizzazione e gestione dei dati che in termini di analisi degli stessi.<br><br>Conoscenza e comprensione applicate. Progettazione di un processo di analisi di grandi volumi di dati, considerando la loro gestione e memorizzazione, la loro elaborazione e la loro analisi. |
| **Prerequisiti** | È necessario conoscere i principi base delle basi di dati, le basi di dati distribuite, i Data Warehouse, i concetti basilari di Apprendimento Automatico. |
| **Contenuti di insegnamento (Programma)** | **1. Big Data: introduction (5h).** Big Data: difinitions, Why and where, Characteristics of Big Data dimensions and scalability, Getting value from Big Data, A fourth paradigm for doing science.<br><br>**2. Big data modeling and storing (5h).** The aggregate model. The key-value, document and column family models. L'esempio di BigTable. Column-family: Google BigTable; Dynamo DB; Graph databases: concepts, queries and data organization<br><br>**3. Knowledge Discovery in Databases (10h).** Context and definitions. The KDD process. Crisp_DM: Business understanding, Data Understanding.Sampling. Feature selection, Feature transformation and construction. Factor analysis. Data representation, Data representation, model representation, Model building and evaluation; PMML.<br><br>**4. Correlation and variable associations (4h)**. Correlation Between two/many quantitative variables. Correlation between qualitative variables: Chi-square, lambda, Sparman Correlation Coefficient. Association rule mining: Apriori and FPGrowth.<br><br>**5. The Bayesian framework and Ensemble learning (5h).** MAP learning, Bayes optimal classifier. Ensemble learning: Boosting, Bagging and Stacking<br><br>**6. Programming framework for Big Data Analysis (5h).** Spark: The MapReduce framework, RDDs, DataFrames and DataBases. Spark SQL, Spark MLlib, Park streaming.<br><br>**7. Learning from data streams (4h).** Motivations, Data Streams definitions, Approximate Answers, Count-Min Sketch. Basic Methods: Estimating statistics over windows, Sampling. Very Fast Decision Trees, Conceptd Drift<br><br>**Laboratorio (30h)**<br>Cassandra DB: Introduction, the architecture of Cassandra, The logical Model. Hand-on sessions<br>Spark, SparkML and Spark streaming., Hand-on sessions. |
| **Testi di riferimento** | Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray, 2013 |

| | |
|---|---|
| | T. Mitchell Machine Learning, Morgan Kaufmann, 1997<br>Richard J. Roiger, Michael W. Geatz. Introduction to Data Mining McGraw-Hill, 2003<br>A. Azzalini, B. Scarpa Analisi dei dati e data mining, Springer, 2004<br><br>Articoli scientifici selezionati e messi a disposizione su ADA.<br>Slide presentate a lezione e messe a disposizione su ADA.<br><br>Gli studenti che lo desiderano possono ottenere i testi in prestito dalla Biblioteca. Può convenire verificarne la disponibilità mediante il Sistema Bibliotecario di Ateneo https://opac.uniba.it/easyweb/w8018/index.php? e contattare la biblioteca per concordare il prestito. |
| **Note ai testi di riferimento** | Sul sito ADA sono disponibili:<br>Le slide presentate a lezione, slide presentate durante i seminari e articoli scientifici di interesse. |

## Organizzazione della didattica

| **Ore** | | | |
|---|---|---|---|
| Totali | Didattica frontale | Pratica (laboratorio, progetto, esercitazione, altro) | Studio individuale |
| 150 ore | 32 ore | 30 ore | 88 ore |
| **CFU/ETCS** | | | |
| 6 CFU | 4 CFU | 2 CFU | |

## Metodi didattici

| | |
|---|---|
| | Lezioni frontali sugli argomenti teorici del programma e su esempi pratici da sviluppare in aula e a casa.<br>Attività in laboratorio per approfondire gli aspetti pratici e tecnologici. |

## Risultati di apprendimento previsti

| | |
|---|---|
| **Conoscenza e capacità di comprensione** | Lo studente acquisirà le principali conoscenze riguardanti il panorama dei sistemi rilevanti di Big Data, i principi su cui si basano i metodi di memorizzazione e il processo di analisi basato su metodologia CRISP-DM. Verranno trattati i database NoSQL, Spark e l'ecosistema che esso ha generato. |
| **Conoscenza e capacità di comprensione applicate** | Lo studente acquisirà le conoscenze pratiche che gli consentiranno di applicare le conoscenze sui sistemi di data storage, processing e analisi su grandi moli di dati. |

| | |
|---|---|
| **Competenze trasversali** | **Autonomia di giudizio**<br><br>Effettuare valutazioni e scelte informate è esattamente lo scopo del corso, gli studenti impareranno a progettare e implementare in modo autonomo uno strumento di Big Data Analytics ed eseguire analisi.<br><br>**Abilità comunicative**<br><br>Per rendere fruibile e credibile la conoscenza estratta, il risultato dell'analisi deve essere presentato adeguatamente. Inoltre, il processo deve essere chiaro basato su standard e trasparente. Questo è un passaggio fondamentale del processo KDD e, di conseguenza, questo è un passaggio fondamentale della Big Data Analytics.<br><br>**Capacità di apprendere in modo autonomo**<br>Lo studente apprenderà i concetti di base che lo renderanno in grado di utilizzare, comprendere e implementare qualsiasi metodo di data mining che estrae conoscenza da un grande volume di dati. |

| **Valutazione** | |
|---|---|
| **Modalità di verifica dell'apprendimento** | L'esame si compone di una parte orale e di una parte pratica. La parte orale ha lo scopo di verificare i concetti appresi e si riferisce a tutti gli argomenti discussi durante le lezioni. La parte pratica è una discussione orale di un piccolo elaborato pratico e mira a verificare le capacità di creare un processo CRISP-DM per l'analisi dei big data.<br><br>Le due parti non sono propedeutiche l'una all'altra (cioè sono indipendenti l'una dall'altra). I voti ottenuti per entrambe le parti saranno mantenuti fino al termine dell'Anno Accademico.<br><br>La valutazione è in trentesimi per entrambe le prove e il voto finale è derivato applicando la media delle valutazioni delle prove. |
| Criteri di valutazione | La valutazione mira a verificare il raggiungimento di una buona capacità di problem solving nell'ambito dell'analisi di grandi volumi di dati. In particolare, rispetto ai risultati di apprendimento attesi, si considerano i seguenti criteri:<br><br>Conoscenza e capacità di comprensione:<br>Capacità di descrivere con accuratezza e chiarezza le tecniche e i fondamenti teorici del processo di analisi dei Big Data;<br><br>Conoscenza e capacità di comprensione applicate:<br>Capacità di progettare e definire una architettura per l'analisi dei dati.<br><br>Autonomia di giudizio:<br>Capacità di giustificare la soluzione adottata tra più opzioni;<br><br>Abilità comunicative:<br>Chiarezza di risposte/soluzioni a domande/problemi;<br><br>Capacità di apprendere: |

| | |
|---|---|
| | Astrazione, ragionamento per analogia e dimostrazione di adattività nella risoluzione dei problemi, anche considerando il dominio di riferimento e i dati a disposizione del processo di analisi. |
| Criteri di misurazione dell'apprendimento e di attribuzione del voto finale | Lo studente dovrà dimostrare di conoscere tutti i concetti discussi durante le lezioni, nonché dimostrare di essere in grado di progettare una analisi completa secondo il modello CRISP-DM, realizzato mediante parte delle tecnologie discusse. |
| **Altro** | Si suggerisce agli studenti di affidarsi esclusivamente alle informazioni/comunicazioni fornite sui siti ufficiali del Dipartimento di Informatica, ovvero sui gruppi social solo se costituiti e amministrati esclusivamente dai docenti dei relativi insegnamenti:<br><br>● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea<br>● https://www.uniba.it/it/ricerca/dipartimenti/informatica<br>● https://elearning.uniba.it/<br><br>I programmi degli insegnamenti sono disponibili qui:<br>● https://elearning.uniba.it/course/index.php?categoryid=287<br><br>Le informazioni che tutti gli studenti dovrebbero conoscere sono scritte nei Regolamenti didattici e manifesti degli studi disponibili nel sito:<br>● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea<br><br>Si suggerisce agli studenti di diffidare delle informazioni e dei materiali circolanti su siti o gruppi social non ufficiali, poiché spesso sono risultati non affidabili, non corretti o incompleti. Per ogni dubbio, chiedere un incontro al docente secondo le modalità previste per il ricevimento. |

## Main information on the course

| | |
|---|---|
| Course name | **Big Data** |
| Degree | Computer Science (Masters' degree in Computer Science) |
| Academic year | 2023/24 |
| European Credit Transfer and Accumulation System (ECTS), in Italian Crediti Formativi Universitari (CFU) | 6 CFU |
| Scientific Disciplinary Sector | ING-INF/05 |
| Course language | English |
| Academic year | Second |
| Delivery period | First semester, exact dates are specified in the program/regulations. |
| Attendance requirement | It is highly recommended to attend classes |
| Course of study's website | https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/computer-science/computer-science |

## Teacher(s)



| | |
|---|---|
| Name and Surname | Michelangelo Ceci |
| Email | michelangelo.ceci@uniba.it |
| Phone | +39 080 5442285 |
| Office | Dipartimento di Informatica, Via Orabona 4, 70125, Bari. Room n.510, 5th floor. |
| E-learning platform | Platform ADA - https://elearning.uniba.it/ |
| Teacher's homepage | http://www.di.uniba.it/~ceci/ |
| Office hours | In-person at the office: Thursday, 9:00-11:00 a.m. preferably by appointment |

| **Syllabus** | |
|---|---|
| **Course goals** | Knowledge and Understanding<br><br>The analysis of data has replaced data acquisition as the bottleneck for evidence-based decision-making processes. This implies that extracting knowledge from large, heterogeneous, and noisy datasets requires not only powerful computing resources but also appropriate methodological foundations and programming abstractions. The abstractions that have emerged in the last decade merge ideas from parallel databases, distributed systems, and programming languages to create a new class of scalable data analytics platforms that form the basis for real-world scale data science. The course will provide an overview of relevant principles and technologies, both in terms of data organization and management, and in terms of data analysis.<br><br>Applied Knowledge and Understanding<br><br>Designing a process for analyzing large volumes of data, considering their management and storage, processing, and analysis. |
| **Prerequisites/requirements** | It is necessary to be familiar with the basic principles of databases, distributed databases, Data Warehouses, and fundamental concepts of Machine Learning. |
| **Course program** | **Big Data: Introduction (5h):** Definitions of Big Data, reasons and contexts for its use, characteristics of Big Data, scalability dimensions, extracting value from Big Data, and the fourth paradigm for scientific research.<br><br>**Big Data Modeling and Storing (5h):** The aggregate model, key-value, document, and column family models. Examples of BigTable, column-family databases like Google BigTable, Dynamo DB, and concepts, queries, and data organization in graph databases.<br><br>**Knowledge Discovery in Databases (10h):** Context and definitions, the KDD process, Crisp_DM (Business understanding, Data Understanding, Sampling, Feature selection, Feature transformation, and construction). Factor analysis, data representation, model representation, model building and evaluation, PMML.<br><br>**Correlation and Variable Associations (4h):** Correlation between two/many quantitative variables, correlation between qualitative variables (Chi-square, lambda, Spearman Correlation Coefficient). Association rule mining using Apriori and FPGrowth.<br><br>**The Bayesian Framework and Ensemble Learning (5h):** MAP learning, Bayes optimal classifier, ensemble learning techniques including Boosting, Bagging, and Stacking.<br><br>**Programming Framework for Big Data Analysis (5h):** Introduction to Spark, the MapReduce framework, RDDs, DataFrames, and Databases. Overview of Spark SQL, Spark MLlib, and Spark streaming.<br><br>**Learning from Data Streams (4h):** Motivations, definitions of data streams, approximate answers, Count-Min Sketch. Basic methods for estimating statistics over windows, sampling, very fast decision trees, and concept drift.<br><br>**Laboratory (30h):**<br><br>**Cassandra DB:** Introduction, architecture of Cassandra, logical model. Hands-on sessions.<br>**Spark, SparkML, and Spark streaming:** Hands-on sessions. |

| | |
|---|---|
| **Books of reference** | Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray, 2013<br>T. Mitchell Machine Learning, Morgan Kaufmann, 1997<br>Richard J. Roiger, Michael W. Geatz. Introduction to Data Mining McGraw-Hill, 2003<br>A. Azzalini, B. Scarpa Analisi dei dati e data mining, Springer, 2004<br><br>Selected scientific articles made available on ADA.<br>Slides presented during the lecture and provided on ADA.<br><br>Students who wish to can borrow the texts from the Library. It may be advisable to check their availability through the University Library System at https://opac.uniba.it/easyweb/w8018/index.php?<br>and contact the library to arrange the loan. |
| **Notes to the books** | On the ADA website, you can find:<br>The slides presented in class, slides presented during seminars, and relevant scientific articles. |

## Organization of the didactic activities

| **Hours** | | | |
|---|---|---|---|
| Total | Lectures | Practice sessions | Individual study |
| 150 hours | 32 hours | 30 hours | 88 hours |
| **CFU/ETCS** | | | |
| 6 CFU | 4 CFU | 2 CFU | |

## Teaching methods

Frontal lectures on the theoretical topics of the program and practical examples to be developed in the classroom and at home.
Laboratory activities to delve into practical and technological aspects.

## Expected learning outcomes

| | |
|---|---|
| **Knowledge and understanding** | The student will acquire the main knowledge regarding the landscape of relevant Big Data systems, the principles underlying storage methods, and the analysis process based on the CRISP-DM methodology. Topics covered include NoSQL databases, Spark, and the ecosystem it has generated. |
| **Applying knowledge and understanding** | The student will acquire practical knowledge enabling them to apply their understanding of data storage, processing systems, and large-scale data analysis. |

| | |
|---|---|
| **Other skills** | **Judgment autonomy**<br><br>Making informed evaluations and choices is precisely the aim of the course; students will learn to independently design and implement a Big Data Analytics tool and conduct analyses.<br><br>**Communication skills**<br><br>To make the extracted knowledge usable and credible, the results of the analysis must be presented appropriately. Moreover, the process must be clear, based on standards, and transparent. This is a fundamental step in the KDD process and, consequently, a fundamental step in Big Data Analytics.<br><br>**Self-learning ability**<br><br>The student will learn the basic concepts that will enable them to use, understand, and implement any data mining method that extracts knowledge from a large volume of data. |

| **Assessment** | |
|---|---|
| **Assessment methods** | The exam consists of an oral part and a practical part. The oral part aims to assess the understanding of acquired concepts and covers all topics discussed during the lectures. The practical part involves an oral discussion of a small practical project and aims to evaluate the ability to create a CRISP-DM process for big data analysis.<br><br>The two parts are not interdependent (i.e., they are independent of each other). Grades obtained for both parts will be retained until the end of the academic year.<br><br>The evaluation is in thirtieths for both components, and the final grade is derived by averaging the scores from both parts. |
| **Evaluation criteria** | The evaluation aims to assess the achievement of a good problem-solving ability in the context of large-scale data analysis. In particular, considering the expected learning outcomes, the following criteria are taken into account:<br><br>Knowledge and understanding:<br><br>The ability to accurately and clearly describe the techniques and theoretical foundations of the Big Data analysis process.<br>Applied knowledge and understanding:<br><br>The ability to design and define an architecture for data analysis.<br>Judgment autonomy:<br><br>The ability to justify the chosen solution among multiple options.<br>Communication skills:<br><br>Clarity in providing answers/solutions to questions/problems.<br>Learning ability:<br><br>Abstraction, reasoning by analogy, and demonstration of adaptability in problem-solving, considering the reference domain and the data available for the analysis process. |

| | |
|---|---|
| **Measurements and final grade** | The student must demonstrate knowledge of all the concepts discussed during the lectures and show the ability to design a comprehensive analysis according to the CRISP-DM model, implemented using some of the discussed technologies. |
| **Further information** | It is recommended that students rely exclusively on information and communications provided on the official websites of the Department of Computer Science or on social groups only if they are established and managed solely by the instructors of the respective courses:<br><br>• https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea<br>• https://www.uniba.it/it/ricerca/dipartimenti/informatica<br>• https://elearning.uniba.it/<br><br>The course syllabi are available here:<br>• https://elearning.uniba.it/course/index.php?categoryid=288<br><br>The information that all students should know is detailed in the educational regulations and course handbooks available on the website:<br>• https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea<br><br>Students are advised to be cautious regarding information and materials circulating on unofficial websites or social groups, as they can often be unreliable, incorrect, or incomplete. For any doubts, students should arrange a meeting with the instructor following the specified office hours.<br><br>———————<br><br>Link to the course on the department's ADA e-learning platform: https://elearning.uniba.it/ |