Course Wise Syllabus

Mathematical Foundations for Data Science Vector and matrix algebra, systems of linear algebraic equations and their solutions; Eigenvalues, eigenvectors and diagonalization of matrices; multivariate calculus, vector calculus, Jacobian and Hessian, multivariate Taylor series, gradient descent, unconstrained optimization, constrained optimization, nonlinear optimization, stochastic gradient descent, dimensionality reduction and PCA, optimization for support vector machines. Introduction to Statistical Methods Basic probability concepts, Conditional probability, Bayes Theorem, Probability distributions, Continuous and discrete distributions, Transformation of random variables, estimating mean, variance, covariance, Hypothesis Testing, Maximum likelihood, ANOVA single factor, dual factor, time series analysis: AR, MA, ARIMA, SARIMA, sampling based on distribution, statistical significance, Gaussian Mixture Model, Expectation Maximization. Data Warehousing Introduction, evolution of data warehousing; decision support systems; goals, benefit, and challenges of data warehousing; architecture; data warehouse information flows; software and hardware requirements; approaches to data warehouse design; creating and maintaining a data warehouse; Online Analytical Processing (OLAP) and multi-dimensional data, multi-dimensional modeling; view materialization; data marts; data warehouse metadata; data mining. Computer Organization & Software Systems Programmer model of CPU; Basic concept of buses and interrupts; Memory subsystem organization; I/O organization; Concept of assembler, linker & loader; Types of operating systems; Concept of process; OS functions: Process scheduling, Memory management, I/O management and related issues. Graphs - Algorithms and Mining Basic concepts of graphs and digraphs connectivity, reachability and vulnerability; Trees, tournaments and matroids; Planarity; Routing and matching problems; Representations; Various algorithms; applications, introduction to graph mining, Graph Pattern Mining, Graph Classification, Graph Compression, graph model, graph dynamics, social network analysis, visualization, summarization, graph clustering, link analysis, applications of graph patterns.

Big Data Systems What is big data - are existing systems sufficient?; Data Warehouse v/s Data Lakes; Hadoop – Components; Storage - Relational DBs/ NoSQL dbs / HDFS / HBase / Object Data stores - S3; Serialization; Interfaces - Hive/ Pig; Stream Processing; Spark; Mahout. Deep Learning Common Architectural Principles of Deep Networks; Building Blocks of Deep Networks; Convolutional Neural Networks (CNNs); Recurrent Neural Networks; Recursive Neural Networks; Building Deep Networks with ND4J; Applications to Sequence Data, Anomaly Detection; Tuning Deep Networks; Vectorization. Probabilistic Graphical Models HMM, Markov Random Field, Bayesian networks, Representation, Learning, Inference; Dynamic Bayesian Networks and Temporal Bayesian networks, applications. Ethics for Data Science Nature of data - data as a by-product of computing, operations data (e.g., sales/marketing), surveillance data (business or government), data collected for research; Ethics - What are ethics, need for ethics, Ethical concerns in computing and analytics; Why data science needs ethics?; Issues -political/social, liberty and justice, fairness and equality, business competitiveness, privacy, anonymity, and security; Data Ownership, Informed Consent, Security Risks (Privacy, Anonymity, Integrity, and Provenance); Ethical methods for sourcing/collecting data, and for storage/ distribution of data. Data validation. Algorithmic Fairness and Case Studies; Solutions to address ethical issues for government, corporations/organizations, research, public use of data, social norms, legal compliance, and case studies. Data ethics in specific domains - e.g. health care, finance, and social studies/research. Optimization Techniques for Analytics Role of optimization in different types of analytics, Introduction to Linear Programming, LP Model and graphical solution, Primal Simplex method, Dual Simplex and Post Optimality Analysis, Revised Simplex method with examples, Application of linear programming in transportation, assignment problems, Integer linear programming, mixed integer programming, complexity analysis, branch and bound techniques, goal

programming, Network models - critical path method and PERT, Dynamic programming, game theory, additional meta heuristic techniques, 2-3 case studies from relevant industry domains.

Data Management for Machine Learning Data Models and Query Languages: Relational, Object-Relational, NoSQL data models; Declarative (SQL) and Imperative (Map Reduce) Querying; Data Encoding: Evolution, Formats, Models of dataflow; Machine learning workflow; Data management challenges in ML workflow; Data Pipelines and patterns; Data Pipeline Stages: Data extraction, ingestion, cleaning, wrangling, versioning, transformation, exploration, feature management; Modern Data Infrastructure: Diverse data sources, Cloud data warehouses and lakes, Data Ingestion tools, Data transformation and modelling tools, Workflow orchestration platforms; ML model metadata and Registry, ML Observability, Data privacy and anonymity. Natural Language Processing Natural Language Understanding and Generation, N-gram and Neural Language Models, Word to Vectors / Word Embedding (Skip gram/CBOW, Glove, BERT/ XLM, MURIL), Part of Speech Tagging, Hidden Markov Models, Parsing - Syntactic, Statistical, Dependency, Word Sense Disambiguation, Semantic Web Ontology. Design of Experiments for Data Science Introduction and importance of Experimental Design, Testing of Hypothesis, Designs with One Source of Variation, Multiple Comparison Testing, Interaction Effect, Factorial Experiment, Fractional Factorial Designs & Confounding, Latin Squares and Graeco-Latin Squares, Fractional-Factorial Designs, Taguchi Design, Designs with Random Effects, Optimal Designs and Model Uncertainty, Design for Nonlinear Model, Sequential Designs. Introduction to Data Science Data Analytics, Data and Data Models, Data wrangling, Feature Engineering, Classification and Prediction, Association Analysis, Clustering, Anomaly Detection, exploratory / explanatory data analysis with visual storytelling, Ethics for Data Science. Information Retrieval Organization, representation, and access to information; categorization, indexing, and content analysis; data structures for unstructured data; design and maintenance of such data structures, indexing and indexes, retrieval and classification schemes; use of codes, formats, and standards; analysis, construction and evaluation of search and navigation techniques; search engines and how they relate to the above. Multimedia data and their representation and search.