NAAIC

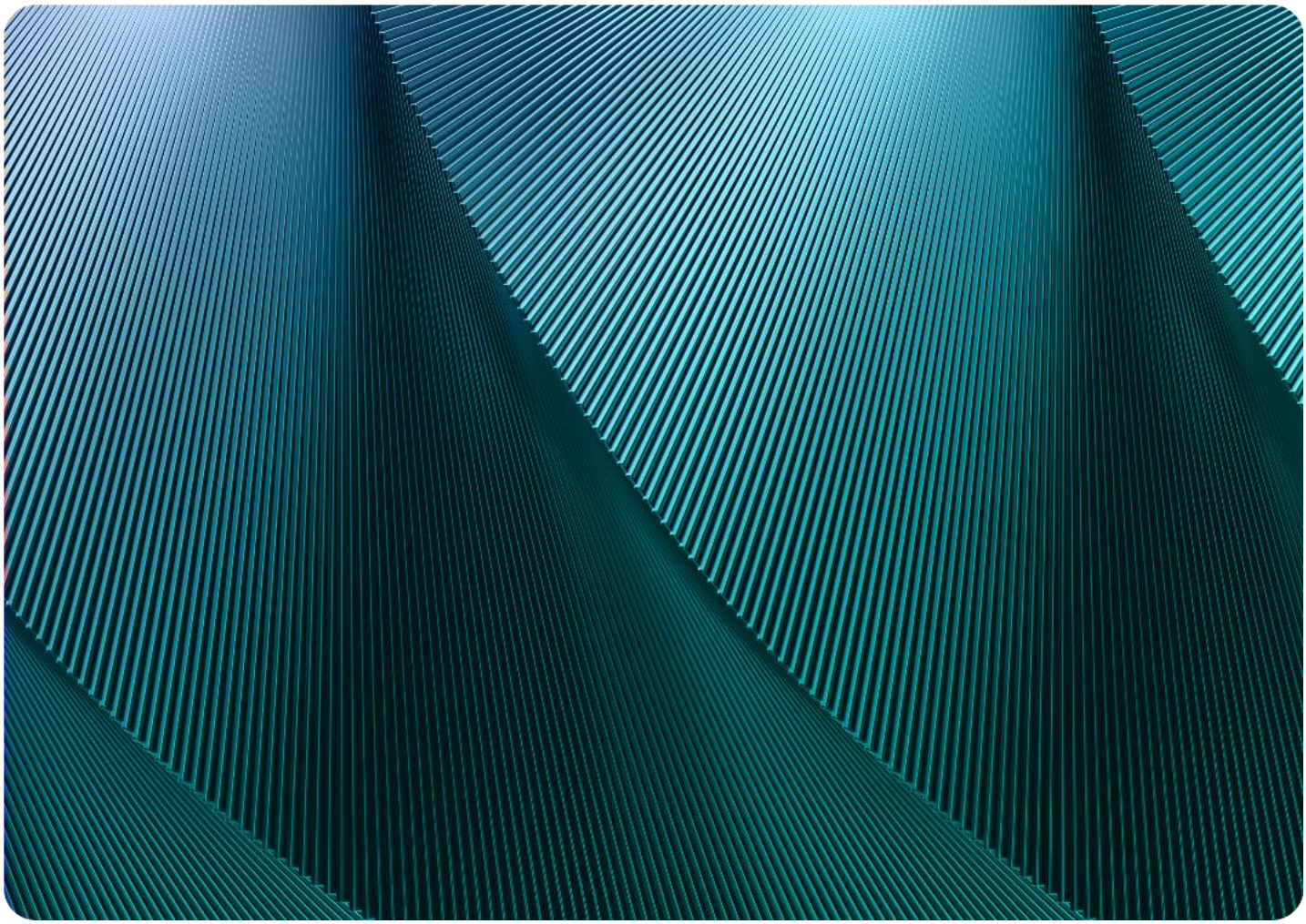
Resource provided by

HCC | HOUSTON COMMUNITY COLLEGE

Course Outline Details

Data Science for AI and Robotics

This is an intermediate course that explores essential data science principles for artificial intelligence and robotics applications.



NAAIC

Course Outline

Week starting on / Module	Module Topic
Week 1	Module 01: The Evolution of Data Science Across AI Paradigms
Week 2	Module 02: AI Data Taxonomy, Sources & Infrastructure
Week 3	Module 03: Data Preprocessing for Machine Learning
Week 4	Module 04: Data Preprocessing for Deep Learning
Week 5	Module 05: Data Preprocessing for Generative AI
Week 6	Module 06: Feature Engineering in ML
Week 7	Module 07: Word Embeddings
Week 8	Module 08: Evaluating and Optimizing ML Models
Week 9	Module 09: Advanced Neural Network Data Handling
Week 10	Module 10: Enterprise Data Pipelines for AI
Week 11	Module 11: Cloud Platforms for AI Implementation
Week 12	Module 12: Data Science for Agentic AI
Week 13	Module 13: Ethical and Responsible AI Data Practices
Week 14	Module 14: Ethical and Responsible AI Data Practices
Week 15	Module 15: Course Wrap up and Review
Week 16	Module 16: Final Exam / Project Due

NAAIC

Course Outline Details

Module 01: The Evolution of Data Science Across AI Paradigms

- Historical Context of AI Evolution: Development from traditional ML to deep learning to generative AI and agent systems
- Five Key Paradigms: Traditional Machine Learning, Deep Learning, Generative AI, Large Language Models, and Agent Thinking
- Comparative Analysis: Data requirements, processing methods, and training approaches across paradigms
- Future Trends and Implications: Emerging developments in AI data science and their potential impact
- **Lab: Comparative Data analysis across AI Paradigms**

Module 02: AI Data Taxonomy, Sources & Infrastructure

- Data Taxonomy in AI: Structured, semi-structured, and unstructured data types and their applications
- Data Preparation Approaches: Requirements for traditional ML vs. deep learning vs. generative AI
- Sources of AI Training Data: Public datasets, private collections, and synthetic data generation
- Cloud Platforms for AI Data Management: Solutions like Snowflake, AWS, Azure, and Google Cloud Storage
- **Lab: Exploring AI Data Sources and Taxonomy**

Module 03: Data Preprocessing for Machine Learning

- Data Quality Assessment: Identifying and addressing missing values, outliers, and format issues
- Data Transformation Fundamentals: Scaling, normalization, and encoding techniques
- Feature Selection Introduction: Methods for identifying the most relevant features
- Dimensionality Reduction: PCA, t-SNE, and UMAP techniques for handling high-dimensional data
- **Lab: Data Quality and Preprocessing Pipeline**

Module 04: Data Preprocessing for Deep Learning

- Deep Learning Data Pipeline: Specialized preprocessing requirements for neural networks
- Image Data Preprocessing: Resizing, normalization, augmentation, and feature extraction
- Text Data Preprocessing: Tokenization, stop word removal, and text normalization
- Additional Data Types: Preprocessing for audio, time series, graph, and multimodal data
- **Lab: Image Data Preprocessing for Neural Networks**

Module 05: Data Preprocessing for Generative AI

- Generative AI Data Requirements: Quality, diversity, and specificity considerations
- Model-Specific Preprocessing: Techniques for GANs, diffusion models, and transformers
- Data Quality and Ethics: Best practices, quality checks, and ethical guidelines
- Data Augmentation Methods: Innovative techniques for generating additional training examples
- **Lab: Text Data Preprocessing for Generative Models**

Module 06: Feature Engineering in ML

- Feature Engineering Evolution: Changes in approach from traditional ML to deep learning
- Feature Generation Techniques: Transformation, encoding, and feature combination methods
- Feature Analysis and Selection: Evaluating importance, distributions, and relationships
- Feature Stores: Centralized repositories for managing and serving features
- **Lab: Feature Engineering for Improved Model Performance**

Module 07: Word Embeddings

- Text Representation Fundamentals: Converting words into meaningful vectors
- Sparse Representations: One-hot encoding, Bag of Words, and TF-IDF
- Dense Representations: Word2Vec, GloVe, and FastText
- Contextualized Embeddings: ELMo, BERT, GPT, and specialized embeddings like CLIP
- **Lab: Building and Comparing Word Embeddings**

Module 08: Evaluating and Optimizing ML Models

- Evaluation Metrics: Accuracy, precision, recall, F1-score, ROC curves, and AUC
- Optimization Techniques: Cross-validation, regularization, and hyperparameter tuning
- Bias-Variance Tradeoff: Understanding and addressing overfitting and underfitting
- Model Comparison and Selection: Techniques for choosing the best model for specific data
- **Lab: Model Evaluation and Hyperparameter Tuning**

Module 09: Advanced Neural Network Data Handling

- Architecture-Specific Optimization: Data techniques tailored to CNNs, RNNs, and Transformers
- Professional Annotation Systems: Platforms, quality assurance, and semi-supervised approaches
- Production-Ready Data Pipelines: Scaling from development to production environments
- Data Debugging and Diagnostics: Identifying and resolving data-related issues in neural networks
- **Lab: Optimizing Data Pipelines for CNNs**

Module 10: Enterprise Data Pipelines for AI

- Evolution of AI Data Requirements: From experimental to production-scale systems
- Components of Enterprise Data Pipelines: Ingestion, storage, processing, and serving
- Implementation Patterns: Feature engineering pipelines, training data preparation, and deployment
- Workflow Orchestration: Tools and techniques for coordinating complex data workflows
- **Lab: Designing an Enterprise Data Pipeline**

Module 11: Cloud Platforms for AI Implementation

- Introduction to Cloud for AI: Benefits of cloud computing for AI data management
- Cloud Service Models: IaaS, PaaS, and SaaS options for AI data workloads
- Major Cloud Platforms: Microsoft Fabric, AWS, Google Cloud Platform, and Snowflake
- Implementation Strategies: Best practices for leveraging cloud platforms for AI data
- **Lab: Deploying a Data Workflow on a Cloud Platform**

Module 12: Data Science for Agentic AI

- Data Requirements for Agentic Systems: Contrasting data needs between predictive models and agentic systems
- Environmental State Representation: Structuring and vectorizing state data for agent environments
- Agent Memory Systems: Data structures for episodic experiences and semantic knowledge
- Tool Use and API Data Patterns: Standardized formats for agent-tool interactions
- Agent Feedback Data: Reward signal design, human feedback collection, and performance metrics
- Data Pipeline Design for Agents: Real-time processing considerations and multi-modal input handling
- **Lab: Data Pipeline for an Agentic AI System**

Module 13: Synthetic Data and Data Augmentation

- How synthetic data improves AI training (e.g., GANs for image generation)
- Applications: Robotics (e.g., simulated environments), medical AI, computer vision
- LLMs for data generation: Synthetic text or sensor data
- Hands-on Lab: Generate synthetic data (e.g., images or time-series) for an AI model, such as a robotics simulation
- **Lab: Generating Synthetic Data for Robotics Simulation**

Module 14: Ethical and Responsible AI Data Practices

- Ethical Foundations in AI Data Science: Key principles and potential harms in data-driven systems
- Bias Detection and Mitigation: Methods for measuring and addressing bias in training data
- Privacy-Preserving Data Science: Techniques including differential privacy, federated learning, and data anonymization
- Data Documentation and Transparency: Creating comprehensive dataset documentation and tracking provenance
- Data Governance Frameworks: Designing systems for responsible data management throughout the AI lifecycle
- Case Studies in Responsible Data Practices: Real-world examples across healthcare, computer vision, and NLP domains

Module 15: Course Review and Wrap-up

- Integrated Data Science Workflow: Connecting preprocessing, feature engineering, and model evaluation
- Advanced Topics Discussion: Emerging trends and future directions in AI data science
- Future Learning Pathways: Resources and directions for continued growth in AI data science

Module 16: Final Submission

- Final date for Comprehensive project submission

Teaching Methods and Strategies

- Total Course Duration: 96 contact hours; 16 weeks
- Weekly Contact Time: 6 hours
- Weekly Structure:
    - Lecture: 2-3 hours
    - Lab: 3-4 hours