Graduate Material Covered

DATA601 – Probability with Statistics

- Basic notions of probability
- Independence
- Random variables/vectors
- Expectation, discrete and continuous variables
- Common types of random variables (uniform, binomial, Bernoulli, Poisson, etc)
- Common distributions (normal, Chi-square, etc) and their properties
- Marginal distributions/conditional expectation
- Law of Large Numbers, Monte Carlo method, Central Limit Theorem
- Parameter Estimators (Biased and Unbiased)

DATA602 – Principles of Data Science

- Basic tools of data science (Python, Jupyter notebooks, Git, cloud computing, SQL, containers such as Docker)
- Basic statistics and probability theory
- Data collection, loading, and modeling
- Data wrangling and cleaning
- Handling missing data (e.g. imputing values)
- Data visualization
- Natural language processing (NLP)

DATA603 – Principles of Machine Learning

- Nomenclature
- Bayes' Decision Theorem, Cost Functions
- Maximum Likelihood Estimators
- Anomaly Detection
- Classification
- Regression
- Gradient Descent
- Learning Curves, Precision vs Recall
- K-Means Clustering
- Neural Networks/Deep Learning
- Principle Component Analysis (PCA)/Dimensionality Reduction
- Decision Trees/Forests
- Ranking Problems

DATA604 – Data Representation and Modeling

- Statistical machine learning
- Topological machine learning
- Approximation theory
- Data transformations
- Big data
- Applications of Dimensionality Reduction
- Parameter Tuning
- Computational Complexity Optimization

DATA605 – Big Data Systems

- Git and Docker
- SQL and NoSQL
- Relational Data Models
- Sorrentum
- Airflow
- MapReduce/Hadoop
- Spark/Dask
- Parallel Databases
- Cloud Computing
- AWS
- Graph Databases
- Streaming Analytics

DATA606 – Algorithms for Data Science

- Random sampling
- Jackknife and Bootstrapping
- Conjugate Priors
- Information Theory (entropy, cross-entropy, KL-divergence, mutual information)
- Expectation Maximization
- Topic Modeling
- Dynamic Programming
- Graphs/Data Structures
- Bayes Nets/Causal Graphs
- Optimization

DATA607 – Communication in Data Science

- R
- RMarkdown
- Effective Data Visualization
- Communicating Uncertainty
- Data Science Ethics

DATA612 – Deep Learning

- Mathematical Fundamentals of Deep Learning
- Classic ML Models
- Neural Networks
- Convolutional Neural Networks (CNN's)
- Recurrent Neural Networks
- Advanced Convolution
- Autoregressive Models
- Diffusion Models
- Word Embedding

DATA641 – Natural Language Processing

- N-grams
- Word and sentence sequential structure
- Syntactic Structure
- Machine Learning in NLP
- Vector Semantics/Embeddings
- Deep Learning NLP Models
- Pre-trained language models
- Model Alignment
- Prompting
- In-Context Learning
- AI/NLP Ethics

DATA650 – Cloud Computing

- AWS (in-depth)
- Computing/virtualization
- Storage/databases
- Networking
- Commonly used cloud services such as EC2, S3, RDS, VPC
- Security through IAM policy, authentication methods, encryption, and firewalls