***Course Number:* 25879**

Course Name: **Foundations of Data Science**

Course Type: Theory
Prerequisite: Probability and Statistics Note: Taking at least an online basic course in Machine Learning required
Level: Undergraduate
Group: Communication Systems

Type & Max Unit: Constant 3
Co-requisite: Nothing.
First Presentation: Fall 2022

Objectives:

- Presenting basic ecosystem and workflow of applied data science
- Review of basic math/statistics concepts required for practical data analysis
- Providing hands-on experience through data science programming on real-world data sets
- Working on real world datasets and becoming familiar with their challenges
- Enhancing vision for understanding new technologies and applications in this field

Topics

1. **Introduction to Data Science Ecosystem: A Systems View**
2. **Basics of Data Models: Data Sources, Database basics (SQL), Data Wrangling**
3. **Data Visualization**
4. **Review of Statistical Analysis: Distributions, Hypothesis Testing, Sampling, Data Leakage**
5. **Regression analysis**
6. **Basics of Causality**
7. **Review of ML: Supervised (K-NN, Decision Tree and Random Forest), Logistic Regression, Classification Pipeline, Labeling, Feature Selection and Normalization**
8. **Review of ML: Unsupervised (Hierarchical Clustering, K-Means)**
9. **Overview of Data Science Workflow** : Model Selection and Evaluation
10. **Basics of Text Analysis** Introduction to NLP basics
11. Recent advances in NLP methods: Basics of DL, BERT
12. **Network and Graph Data Analytics**
    - 12.1. Basics of Graph Theory
    - 12.2. Databases for Graph Analysis
    - 12.3. Basics of Graph Algorithms
13. ML Deployment and Scaling Up
14. **Summary**

References

1. *Designing Machine Learning Systems, Chip Huyen, O'Reilly, May 2022*
2. *Introduction to Computation and Programming using Python: with application to computational modeling and understanding data*, Guttag, The MIT Press, 2021.

3. *Foundations of Data Science*, Avrim Blum, John Hopcroft, and Ravindran Kannan, Cambridge University Press, 2020
4. *Computational and Inferential Thinking, The Foundations of Data Science*, Ani Adhikari and John DeNero, UC Berkeley, 2021.
5. *Causal Inference what if*, Hernan and Robins, CRC Press, 2020.
6. *SQL for Data Science*, Badia, Springer, 2020.
7. *Graph Representation Learning*, Hamilton, McGill University Press, 2020.
8. *Machine Learning in Production*, Kelleher, Addison Wesley, 2019.
9. *Practical Statistics for Data Scientists*, O'Reily Press, Bruce et. Al., 2017.
10. *Networks, Crowds and Markets: Reasoning about a highly connected world*, Easley and Kleinberg, Cambridge University Prezs, 2010.
11. *Data Analysis using Regression*, Gelman and Hill, 2007
12. *Foundations of Statistical NLP*, Christopher D. Manning and Hinrich Schütze, 1999.