MATH 571/CSP 571
Data Preparation & Analysis
Class Hours: Wednesday 6:45PM - 9:35PM
Class Location: Hybrid (Synchronous): PS 111 + Internet

Jawahar J. Panchal
jpanchal@iit.edu / (312) 567-5871
Office Hours: TBD/Virtual
Office Location: SB 228A

This is a tentative syllabus - any subsequent changes will be communicated to students in class.

Course Description: This course will provide students with an introduction to the field of data science and the tools and techniques to analyze data and extract knowledge from it. This course surveys industrial and scientific applications of data analytics and will utilize case studies and programming exercises to explore opportunities and challenges involving data analysis and visualization including business opportunities, privacy concerns, and ethical issues. Students will work with a variety of real world data sets and learn how to prepare data sets for analysis by cleaning and reformatting.

Prerequisite(s): CS 425 or equivalent. Math 474 or equivalent. Graduate standing or permission of instructor.
Note(s): This is an *elective* course for MATH, CS, and CSP majors.
Credit Hours: 3 (3-0-3)

Required Text(s): *Introduction to Statistical Learning*, 2nd Edition (Online Edition) [Free]
Author(s): Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
ISBN-13: 978-1461471370

Required Text(s): *R for Data Science*, 1st Edition (Online Edition) [Free]
Author(s): Hadley Wickham, Garrett Grolemund
ISBN-13: 978-1491910399

Course Objectives:
This course provides students with the knowledge and skills to effectively analyze data to unearth knowledge including the ability to develop programs that support the analysis and presentation processes.

Course Outcomes:
Students successfully completing this course will be able to:

1. Discuss the concepts of data science
2. Gather requirements for data analysis projects
3. Prepare data for analysis
4. Programmatically analyze data
5. Develop meaningful data visualizations
6. Explore case studies

7. Examine ethical issues in data science
8. Report on findings of their analysis
9. Develop and present analysis effectively

Course Details:

The course will attempt to incorporate the following tools as part of the course material:

1. Use of the R programming language as a general-purpose data-analysis programming environment.
2. Use of the CRAN package ecosystem as a source of software libraries for extending/supplementing the R environment.
3. Use of various open data sources for projects and analytical work.
4. Use of various cloud provider systems for computational/modeling work.

Grading System:

Grade weighting is prescribed as below, with significant improvement between Midterm Exam to Final Exam scores taken into consideration in the final grade assignment.

Homework Assignments	20%
Reading Quizzes	20%
Semester Project	20%
Midterm Exam	20%
Final Exam	20%

Letter Grade Distribution:

Grade assignment is prescribed as below, with the student score rounded up if within 1%-2% of the next grade level, at the instructor's discretion.

>= 90	A
80 - 89	B
70 - 79	C
60 - 69	D
<= 60	E

Teaching Assistant:

The TA for this course is **Akshara Kudumula** - who will be available for questions regarding grading on assignments, exams, quizzes, and the project. The TA can be contacted at akudumula@hawk.iit.edu

Course Policies:

- **General**
  - Each course session will follow a lecture format with slides and examples - students are expected to take notes and participate in questions/answers during the session.
  - There will be a short 15 minute break, at the instructor's discretion, during each session - students are expected to return back to the classroom on time.
- **Reading Requirements & Homework Assignments**
  - Students are expected to complete reading assignments per the course schedule - homework assignments will be assigned by the instructor with details on submission deadlines.
  - **Late assignments submitted within one week of the due date will be subject to a full grade penalty - no late assignments will be accepted beyond a one week time period**.
- **Online Access & Electronic Submissions**
  - **Lectures will be recorded and made available online**, however attendance is strongly encouraged in order to understand the course material.
  - Students are expected to submit electronic documents for their homework assignments via the IIT Blackboard system.
- **Examinations**
  - Exams are closed book, closed notes - the Midterm Exam will be administered in class, the Final Exam will be administered according to university final exam schedules.
  - **No makeup exams will be given except in extreme circumstances/emergency situations, subject to department and university approval.**
- **Grades**
  - Student grades will be posted on Blackboard in a timely manner - students who wish to track their progress may do so online, or inquire with the instructor.
  - Any questions or discussions regarding grades should be directed to the instructor,
- **Attendance and Absences**
  - Attendance is expected for each session - students may contact the instructor regarding any missed classes due to sickness, emergencies, or other issues.
  - Students are responsible for all missed material, regardless of the reason for absence, and are expected to obtain notes/content independently.

University Policies:

- **General**

  Students should refer to the Illinois Tech Student Handbook as a reference to any and all policies listed below pertaining to this course.

- **Academic Honesty**

  Students are subject to the Code of Academic Honesty as part of being enrolled in this course. Issues related to academic honesty within this course will be handled according to university policies, regulations, and procedures.

- **Code of Conduct**

  Students are subject to the Code of Conduct as part of being enrolled in this course. Issues related to conduct within this course will be handled according to university policies, regulations, and procedures.

- **Special Accomodations**

  Students requiring special accomodations, such as in the case of documented disabilities, should contact the Center for Disability Resources. Accomodations will be arranged via the Reasonable Accomodations process.

Tentative Course Outline:
Material coverage, lecture order, and content timing may change - the student is expected to maintain independent progress regarding reading requirements and homework assignments.

Week	Content
Week 1 08/25/2021	Statistical Learning - Descriptive/Inferential Statistics • Exploratory Data Analysis - Empirical/Parametric Dist., Visualization • Statistical Modeling - Model Accuracy, Error, Bias/Variance • **Suggested Reading**: James Ch1,Ch2; Wickham Ch1,Ch2-8
Week 2 09/01/2021	Linear Regression • Linear Regression - Simple/Multiple, Estimation • Model Accuracy - Diagnostics/Validation, Additional Considerations • **Suggested Reading**: James Ch3; Wickham Ch2-8 • **Homework Assignment**: Homework 1 Assigned
Week 3 09/08/2021	Classification • Logistic Regression - Generlized Linear Model/Least Squares (GLM/GLS) • Linear Discriminant Analysis (LDA) - Bayes Theorem • **Suggested Reading**: James Ch4; Wickham Ch9-16
Week 4 09/15/2021	Resampling & Cross-Validation • Cross-Validation: Leave-One-Out, k-Fold • Bootstrap: Overview • **Suggested Reading**: James Ch5; Wickham Ch9-16 • **Homework Assignment**: Homework 1 Due, Homework 2 Assigned
Week 5 09/22/2021	Regularized Regression • Model Selection: Subset Selection, Optimal Models • Shrinkage Methods: Ridge, Lasso • **Suggested Reading**: James Ch6; Wickham Ch9-16 • **Project Deliverable**: Project Group & Topic Form Due
Week 6 09/29/2021	Non-Linear Models • Splines: Piecewise, Polynomial, Smoothing • GAM: Regression, Classification • **Suggested Reading**: James Ch7; Wickham Ch17-21 • **Homework Assignment**: Homework 2 Due, Homework 3 Assigned
Week 7 10/06/2021	Supervised Learning - Decision Trees • Classification/Regression: CART, Additional Considerations • Ensemble Methods: Random Forests, Boosting/Bagging • **Suggested Reading**: James Ch8; Wickham Ch17-21
Week 8 10/13/2021	Support Vector Machines • Maximal Margin Classification: Hyperplanes, Seperability • SVM: Classification, Decision Boundaries • **Suggested Reading**: James Ch9; Wickham Ch22-25 • **Homework Assignment**: Homework 3 Due • **Project Deliverable**: Project Proposal & Outline Due - *Presentation* • **Quiz Assessment**: Quiz 1
Week 9 10/20/2021	**Midterm Exam** • *In Class/Online Examination*

Week	Content
Week 10 10/27/2021	Deep Learning - Neural Networks • ANN: Single/Multiple Layer, Use Cases • CNN/RNN: Fitting, Backpropogation • **Suggested Reading**: James Ch10; Wickham Ch22-25 • **Homework Assignment**: Homework 4 Assigned
Week 11 11/03/2021	Survial Analysis - Censored Data • Survival: Curves, Log-Rank, Hazards • Shrinkage: Cox Model, AUC • **Suggested Reading**: James Ch11; Wickham Ch22-25 • **Project Deliverable**: Project Plan & Detail Due - *Presentation*
Week 12 11/10/2021	Unsupervised Learning - Dimensionality Reduction/Clustering • Dimensionality Reduction: Principal Component Analysis (PCA) • Clustering: K-Means, Mixture Models (GMM) • **Suggested Reading**: James Ch12; Wickham Ch22-25 • **Homework Assignment**: Homework 4 Due, Homework 5 Assigned
Week 13 11/17/2021	Multiple Testing • Confirmatory Data Analysis: Hypothesis Testing, ANOVA • Error Rates: Type I/Type II, Family-Wise, False Discovery • **Suggested Reading**: James Ch13; Wickham Ch22-25
Week 14 11/24/2021	**Thanksgiving Break** • *No Class* • **Homework Assignment**: Homework 5 Due, Homework 6 Assigned
Week 15 12/01/2021	Catch-Up & Review • **Project Deliverable**: Project Presentation & Report Due - *Presentation* • **Quiz Assessment**: Quiz 2
Week 16 12/08/2021	**Final Exam** • *In Class/Online Examination* • **Homework Assignment**: Homework 6 Due