Project Instruction
CSE 40647/60647 Data Science

Project teams are welcome to discuss with the instructor on ideas, progress, and any other thing. Send an email (mjiang2@nd.edu) to make an appointment (for either 15 mins or 30 mins). Monday afternoon, Tuesday after class, and Friday morning are preferred time.

Project goal:

For the course project, students will be expected to collect one or multiple datasets (online or otherwise), formulate a question of interest, and perform aspects of data science to address that question by using whatever tools they find appropriate. The project will involve a **proposal**, **milestone**, and **final paper** with **oral presentations** of the project.

Project introduction:

- The students should work in **team of 3 – 4.**
- The class project may involve some or all stages of the **knowledge discovery** process, depending on the chosen project. All project topics should be preapproved by the professor.
- The class project will require a **proposal and milestone assessment** during the semester with respect to the data science process.
- The students will be required to write **project papers** (**proposal, milestone, and final)** and make a **class presentation** on their project.
    - The **project paper** must be in PDF format and formatted according to the new Standard ACM Conference Proceedings **Template**.
    - The **project paper** should include **sections** about Introduction, Related Work, Problem Definition, Methodology, Experiments, Discussion, Conclusion and Future Work.
        - There is no page limit.
        - For LaTeX users: unzip acmart.zip, make, and use sample-sigconf.tex as a template; Additional information about formatting and style files is available at: https://www.acm.org/publications/proceedings-template
        - For Word users: export into PDF format.

Grading policy: (20% of the final score)

Grading distribution: (100+5 points)

- **Proposal paper (10 points)**
- **Milestone paper (30 points)**
- **Final paper & presentation (40 points)**
- **Code package and data (20 points)**
- **Discussion bonus (+5 points), if the team discussed with the instructor 3 times in the semester AND at least once before the milestone.**

Students are required to submit their **data and code package + "readme" (.ZIP) and term paper (.PDF)**.

Students are encouraged to **implement** algorithms such as Apriori, FP-Growth, Decision Trees, Naïve Bayes, SVM, and K-Means Clustering by themselves instead of calling Python packages. Students are also encouraged to **use Python packages** (e.g., Numpy and Scipy) when they use **advanced techniques** (e.g., Neural Networks, word2vec) to address challenging problems.

Project Policies:

Grading distribution

Proposal (10%), Milestone (30%)

Final paper & Presentation (40%), Code/Data (20%)

Teaming

- Each team should have three to four students. A mixture of undergraduates and graduates is allowed. The teams will be evaluated uniformly no matter they have undergraduates or graduates.

Required items and dues

- Proposal
  - Paper (PDF): title, problem definition, potential solutions, data sources, proposed evaluation methods, project plan/timeline **due 9/9**
- Milestone
  - Paper (PDF): introduction, related work, problem definition, one or more method that has been used, some other potential solutions, data and experiment settings, evaluation methods, preliminary results and experimental analysis, timeline **due 10/18**
  - Presentation (PPTX) for selective projects: **due 10/15, 10/17** in class
- Final
  - Paper (PDF): introduction, related work, problem definition, solutions and methods that have been used, data and experiment settings, evaluation methods, experimental results and analysis (tables and figures), discussion and future work, conclusions **due 12/13**
  - Presentation (PPTX): **due 12/3, 12/5, 12/10** in class.
- Data and code package (ZIP): **due 12/13**

The papers (PDF) should be formatted according to the new Standard ACM Conference Proceedings Template: https://www.acm.org/publications/proceedings-template

There will be -33% penalty if the paper is not formatted correctly.

Deadlines

Every project required item is due at midnight (11:55pm) with some grace period. There will be -33% penalty for each 24 hours past the deadline.

Both **milestone paper** and **final paper** will be graded using the **project paper** rubric.

The **project proposal** will be graded as follows:

Title of Project:	5%	What's the title of the project?
Project Plan:	30%	What do you plan to do? Please clearly define the data science functionality/task. Define the input and expected output. For example, if the task is classification, define basic concepts in data science: data objects, features, and labels.
Data Sources:	20%	What data do you plan to use? From where will this data come? Please try your best to describe the datasets to readers. Use the methods/concepts you learned from Data Description, Visualization, Cleaning and Integration if possible.
Proposed Evaluation:	30%	How do you plan to evaluate your proposed method? How will you determine whether the method is successful?
Writing Quality:	15%	Clarity of expression (5%), organization (5%), and grammar (5%).

The **project presentation** will be graded as follows:

Introduction:	15%	Provide context. What questions are being addressed?
Solution/Method:	30%	What did you do? Why did you choose this method? What tools and techniques did you use?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable?
Evaluation and Results:	30%	What evaluation did you do? Do your conclusions match your results?
Presentation Quality:	15%	Clarity of speaking (5%), organization (5%), and visuals (5%).

The **project paper** will be graded as follows:

Introduction:	15%	Provide context and motivation. What questions are being addressed? Why are these questions interesting or important?
Related Work:	10%	What other methods have addressed these or similar questions? How do these methods differ from your method?
Solution/Method:	25%	What did you do? What tools and techniques did you use? Was any innovation attempted?
Data and Experiments:	10%	What data did you use? Are your experimental methods reliable? What preprocessing was done the data?
Evaluation and Results:	25%	Did you properly evaluate your experiments? Did you test for statistical significance? Do your conclusions match your results?
Writing Quality:	15%	Clarity of writing (5%), organization (5%), and grammar (5%).

Data Portals:

- Kaggle: https://www.kaggle.com/
- DATA.GOV: https://www.data.gov/
- City of Chicago Data Portal: https://data.cityofchicago.org/
- City of South Bend Open Data: http://data-southbend.opendata.arcgis.com/
- Index of Complex Networks: https://icon.colorado.edu/
- The Koblenz Network Collection: http://konect.uni-koblenz.de/
- Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/

Other Resources

Data Sources

KDnuggets Data Repositories List — Data repository list maintained by KDnuggets, a popular data mining website
UCI Datasets — The UC Irvine Machine Learning Repository, a popular source of machine learning datasets
mldata.org — A public repository for machine learning data
Wikipedia Database — Webpage for access to complete Wikipedia database dumps
IMDb Datasets — Webpage for access to IMDb datasets
Last.fm Datasets — Webpage for access to Last.fm datasets
Census.gov — US government source of data about the nation's people and economy
Data.gov — Source of machine readable datasets generated by the US government
UK's Office for National Statistics — Source of datasets generated by the UK's Office for National Statistics
UK's Met Office Data — Climate station records from the UK's National Weather Service
CDC Data — Medical data from the Centers for Disease Control and Prevention
World Bank Catalog — World Bank data
RealClimate Data — Aggregator for selected sources of code and data related to climate science
Google Public Data Explorer — Google's public data portal to explore, visualize, and communicate large datasets
Dataverse Network — Repository for research datasets
Linked Data — Linkage site for distributed data
Datamob — Aggregator for public datasets
Quandl — Search engine for financial, economic, and social datasets
Data Market — Portal for shared business data
CKAN — Open-source data portal platform
Hilary Mason (bitly) Data Links — Hilary Mason's bookmarked research-quality datasets
Peter Skomoroch (LinkedIn) Data Links — Peter Skomoroch's bookmarked machine learning data resources
Jake Hofman Data Links — Jake Hofman's bookmarked computational social science data resources
Reddit Open Data — Forum on the social news site reddit for open APIs and datasets
Guardian DataBlog — Data journalism and data visualization from the Guardian
Free SVG Maps — Website for free geographic maps
StateMaster — Reference site for data on US states
Wolfram|Alpha — Computational knowledge engine or answer engine

Data Visualization Resources
Many Eyes — Web community that connects visualization experts, practitioners, academics, and enthusiasts
Visual Complexity — Resource space for anyone interested in the visualization of complex networks
Thumbs Up Viz — Collection of elegant, efficient, and (above all) effective data visualizations
WTF Visualizations — Visualizations that make no sense

Python
Python.org — The Official Python Website
The Python Tutorial — The Python.org Python tutorial
Learn Python in X Minutes — Whirlwind tour of Python programming
Learn Python the Hard Way — Teaches Python by slowly building and establishing skills through practice and application
Learn Python (interactive) — Engaging Python tutorials
Google's Python Class — Teaches Python via written materials, lecture videos, and lots of code exercises
pyvideo.org — Python-related video index
yhat Data Science in Python Tutorial — Uses IPython to teach data science
Anaconda Python Distribution — Free Python distribution for large-scale data processing and predictive analytics
The Python Package Index — Repository of Python software
pip — Tool for installing and managing Python packages
NumPy — Python package for scientific computing
SciPy Library — Python package for mathematics, science, and engineering
Matplotlib — Python package for 2D plotting
pandas — Python package for high-performance, easy-to-use data structures and data analysis tools
IPython — Architecture for interactive computing with Python
scikit-learn — Python package for machine learning