EECS6414 Data Analytics and Visualization

Project Description

The most important learning component of the class consists of a substantial course project. The project will offer you an opportunity to develop a quantitative and qualitative intuition of data analysis and visualization methods and algorithms, and to obtain practical experience working with software and tools for large-scale data analysis and visualization. This can prepare you for applying state-of-the-art approaches in real-world settings and data. If you are interested in research, it will also equip you with necessary skills and knowledge to perform data science research. There can be *three types* of projects:

- **Type I**: An in-depth exploratory analysis of an interesting dataset that provides insights of the data that were otherwise hidden or non-obvious.
- **Type II**: A theoretical project that considers a model, an algorithm or a network property (measure) and derives a rigorous analysis and visualization of an interesting dataset. Experimental evaluation of algorithms and models should be in place.
- **Type III**: An efficient implementation of a data analytics and visualization solution that can scale to massive datasets.

Ideally, projects will be a combination of the three types of projects outlined above. All project should contain some experimentation on real data, and some amount of mathematical analysis. Projects will be evaluated based on:

- *Significance/Novelty*. Is the analysis "real" and "interesting", or just a "toy" analysis? How original, important and well defined are the questions posed? How novel is the approach? Is the analysis likely to be useful and/or have impact? Are there any novel/interesting applications of analysis?
- *Technical Quality*. Is the data analysis approach and methods appropriate and well described? Are sufficient details provided? Is the technical material correct? Are the data analytics methods and algorithms reproducible? Are the data visualization methods and tools creative and interesting? Is the interpretation (discussion and conclusion) well balanced and supported by the data?
- *Organization*: Is the report well organized? Is the write-up clear and the language adequate? Are results presented in the most appropriate manner? Are figures and tables used appropriately?

Project Deliverables

The table below presents the breakdown of the project deliverables, weights and due dates.

Project Deliverables	Weight	Due Date
Proposal (~2 pages)	10%	Wed, Jan 31
Midterm progress report (~5 pages)	20%	Wed, Feb 28
Midterm in-class presentation	10%	Wed, Mar 6
Final report (~7 pages)	40%	Wed, Mar 27
Final presentation	20%	Wed, Apr 3

Datasets

Check the course website, under resources.

Software Tools and Libraries

Check the course website, under resources.

Deliverable 1: Project Proposal (10%)

The project proposal should build on one or more large-scale open dataset(s) and/or APIs that are good candidates for analysis and visualization. The idea is to survey the data sets freely available online and identify what are their strengths and weaknesses for the needs of the analysis. The proposal should then focus on what are some promising intuitions or questions that can be answered through rigorous analysis of the data. You should try to provide a concrete proposal for a data analytics solution that can potentially be used for analysis of similar datasets over time. Emphasis should be given on how the solution scales or what methods can be used to alleviate the challenges of big data analytics.

The header of the project should include the course title and an indication that this is a project proposal, the title of the project, your name(s) and contact information. The content of the project should have the following parts:

- **Motivation and Domain Description**: What is the data domain? What is the goal of your project? What is the motivation for rigorous data analytics in this domain? What are the questions you want to answer? Why the analysis is important? What are a few potential applications?
- **Methodology**: What are the data analytics methods you will employ? What are the steps you need to perform? Are there any technical problems you need to solve? How will you address the problems? What type of data analysis you will perform? How this type of analysis is adequate for the data, problem and questions posed? Try to be as specific as possible.
- **Evaluation**: How will you evaluate your data analytics system? What experiments you plan to do? What other datasets can be used? What are the steps you need to take to scale it?
- **References**: The proposal should include the full reference of the papers that you want to base your project ideas on (full citation).

Formatting and Style

The suggested length of the project proposal is **2 pages** and it must be in PDF format. All reports should be formatted according to the ACM SIG conference proceedings template in LaTex and prepared using Overleaf, a free collaborative authoring tool. The template can be accessed here:

https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sigproceedingstemplate/bmvfhcdnxfty

How to Submit?

```
% submit 6414 project proposal.pdf team.txt
```

Deliverable 2: Project Midterm Progress Report (20%)

The project midterm progress report should represent a first (incomplete) draft of your final report. The expectation is that almost 50% of the work has been completed. At this stage, you should be able to provide a complete outline of the project, even if certain key parts have not yet been implemented/solved and any major results are not available. The header of the report should include the course title and an indication that this is a project midterm progress report, the title of the project, your name and contact information. The outline of the report should be structured as follows:

- **Introduction/Motivation**: What is the project about? What is the data domain? What is the goal of your project? What is the motivation for rigorous data analytics in this domain? What are the questions you want to answer? Why the analysis is important? Are you testing any specific hypothesis? What are a few potential applications?
- **Problem Definition:** Is there any algorithmic component in your project (e.g., a prediction model, classification model, clustering model, network model)? If yes, introduce notation, provide formal definitions as needed, define any constraints or restrictions, define what you try to optimize (e.g., maximize or minimize an optimization function, or an accuracy/error function). Describe the problem in a formal way. Describe the hardness of the problem in a formal way.
- **Related Work:** Position the problem among the body of existing research or technical studies. How does your project relate to previous work? How is your project replicating/ different/ complimentary to previous work? References to papers or other studies you cite should be explicit.
- **Methodology**: What are the data analytics methods you will employ? What are the steps you need to perform? What type of data analysis you will perform? How this type of analysis is adequate for the data, problem and questions posed? Describe your overall data analytics architecture. Describe the data collection/ingestion process, data storage process, data analysis, data serving, and data visualization. Are there any algorithmic components in your project? What is your approach to solve the problem? Provide any mathematical background necessary for the methods. Describe any algorithms or variations of the methods. Describe limitations or difficulties with your approach. Formally describe any important algorithms used from literature. Try to be as specific as possible.
- **Evaluation/Results**: How will you evaluate your work? What experiments you plan to do? What datasets have been used? Provide summary statistics of your dataset. Show representative results of your analysis, discuss important findings, discuss implications of your analysis to applications.
- **Conclusions**: What are the conclusions of your work? Are there any highlights? Is there need to discuss or further interpret the results? What are some ideas for future work?
- **References**: The proposal should include the full reference of the papers that you want to base your project ideas on, your approach to solve the problems, and the tools and datasets that you have employed. Full citation is required. References should be specific and found inside the text, as appropriate.

Parts of the above outline will (probably) be filled at a later stage, especially the ones that relate to the details of the methods, the evaluation and the conclusions. At this phase, you should try to fill in as many parts of the outline as possible so that it is clear what you plan to do for the final version.

Formatting and Style

The suggested length of the project proposal is **5 pages** and it must be in PDF format. All reports should be formatted according to the ACM SIG conference proceedings template in LaTex and prepared using Overleaf, a free collaborative authoring tool. The template can be accessed here:

https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sigproceedingstemplate/bmvfhcdnxfty

How to Submit?

```
% submit 6414 project midterm-report.pdf team.txt
```

Deliverable 3: Midterm In-class Presentation (10%)

The midterm in-class presentation should be seen as your opportunity to present your work in class, challenge your ideas, get feedback from peers and discuss methods and algorithms. The expectation is that almost 50% of the work has been completed and some interesting analysis results are available to share with an audience. At this stage you should be able to tell a story about your project, even if parts of it have not yet been completely implemented or solved. The title of the presentation should include the course title and an indication that this is the project's midterm in-class presentation, the title of the project, your name and contact information. The following are some guideline, tips and advice for preparing your presentation.

- You (or your group) will have 10 minutes to present your work in the classroom. Another 5 minutes will be allocated for questions and discussion.
- You should prepare a set of ~10 slides, given that a slide should take a minute to talk about on average.
- Presentations should be organized into thematic units. A typical outline includes:
    - Motivation of the project, its importance and potential applications.
    - Definition of problem (if any), including input, constraints, desirable output and hardness
    - Main project approach ideas, data, methods and the type of analysis proposed
    - Highlights of the results (visual, experimental, theoretical)
    - Interesting variations and limitations of the analysis
    - Concluding remarks
- The talk should be self-sufficient, meaning that you should not make any assumption about prior knowledge of the audience or previous well-known results. All concepts should be introduced and appropriate notation should be used consistently throughout the presentation.
- Focus on the essential parts of the project and avoid too-many technical details. The goal is to give a summary of the project and convey the contribution of your work to other people. At the same time, you should make sure that important content is adequately covered.
- Prepare the slides carefully. Text should be easily readable and slides should not be overloaded with content. Avoid full text sentences and use of math symbols, unless necessary.
- Practice the talk several times, and time yourself to make sure you are within the time bounds.

More Advice & Tips

Some interesting advice on how to give a bad talk by David A. Patterson (UC Berkeley):
https://people.eecs.berkeley.edu/~pattrsn/talks/BadTalk.pdf

Some design tips for beautiful presentations:
https://visage.co/11-design-tips-beautiful-presentations/

PowerPoint Templates

Please find below links to PowerPoint Templates. These are designed with the York brand's primary and secondary colours and fonts embedded for ease of use and can be customized. Template files are available in a widescreen (16:9) format. Branded layouts featuring the York colours and campus photos are also included for title and section slides. This PowerPoint template is AODA compliant (last updated January 2021)

Widescreen with built-in instruction (.potx)

Widescreen without built-in instruction (.potx)

How to Submit?

```
% submit 6414 project midterm-presentation.pptx midterm-presentation.pdf team.txt
```

Deliverable 4: Project Final Report (40%)

The project final report should represent all the completed work. The expectation is that most of the work has been completed and any major results are available. At this stage, you should be able to provide a complete description of the project, even if a few minor parts have not yet been implemented or solved. The header of the report should include the course title and an indication that this is the project's final report, the title of the project, your name and contact information. The report should be structured as follows:

- **Abstract**: The abstract (limited to **150-200** words) should be a comprehensive but concise description of your project that aims to attract potential readers. It should briefly discuss the motivation, problem of interest, approach to solve it and main results of your work.
- **Introduction/Motivation**: What is the project about? What is the data domain? What is the goal of your project? What is the motivation for rigorous data analytics in this domain? What are the questions you want to answer? Why the analysis is important? Are you testing any specific hypothesis? What are a few potential applications?
- **Problem Definition:** Is there any algorithmic component in your project (e.g., a prediction model, classification model, clustering model, network model)? If yes, introduce notation, provide formal definitions as needed, define any constraints or restrictions, define what you try to optimize (e.g., maximize or minimize an optimization function, or an accuracy/error function). Describe the problem in a formal way. Describe the hardness of the problem in a formal way.
- **Related Work:** Position the problem among the body of existing research or technical studies. How does your project relate to previous work? How is your project replicating/ different/ complimentary to previous work? References to papers or other studies you cite should be explicit.
- **Methodology**: What are the data analytics methods you have employed? What are the steps you need to perform? What type of data analysis you have performed? How this type of analysis is adequate for the data, problem and questions posed? Describe your overall data analytics architecture. Describe the data collection/ingestion process, data storage process, data analysis, data serving, and data visualization. Are there any algorithmic components in your project? What is your approach to solve the problem? Provide any mathematical background necessary for the methods. Describe any algorithms or variations of the methods. Describe limitations or difficulties with your approach. Formally describe any important algorithms used from the literature. Try to be as specific as possible.
- **Evaluation/Results**: How did you evaluate your work? What experiments did you perform? What datasets have been used? Provide summary statistics of your dataset. How your evaluation provides support (or not) of your methods. Show results of your analysis, discuss important findings, discuss implications of your analysis to applications.
- **Conclusions**: What are the conclusions of your work? Are there any highlights? Is there need to discuss or further interpret the results? What are some ideas for future work?
- **References**: The final report should include the full reference of the papers that you have based your project ideas on, your approach to solve the problems, and the tools and datasets that you have employed. Full citation is required. References should be specific and found inside the text, as appropriate.

At this phase, you should try to fill in as many parts of the report as possible so that it is clear what your term work has been.

Formatting and Style

The suggested length of the final report is **7-8 pages** and it must be in PDF format. All reports should be formatted according to the ACM SIG conference proceedings template in LaTex and prepared using Overleaf, a free collaborative authoring tool. The template can be accessed here:

https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sigproceedingstemplate/bmvfhcdnxfty

Final Report Evaluation

The *final report* will be evaluated based on the following mark breakdown.

Final Report Component	Weight	Due Date
Introduction/Problem Definition	20%	Clear motivation that encourages the reader to read on; clear notation and problem/analysis definition (if any) including input and desirable output
Related Work	10%	Important references are not missing; explicit citations
Model/Methodology/Algorithms	30%	Clear and well written so that we can fully understand what you did
Evaluation/Results	30%	Comprehensive evaluation plan; clear and conclusive set of experiments
Style and Language	10%	Overall organization, language, and style

How to Submit?

```
% submit 6414 project final-report.pdf code.zip team.txt
```

Deliverable 5: Final In-class Presentation (20%)

The final in-class presentation should be seen as your opportunity to present your hard work in class, your ideas, your methods and algorithms, your results, and discuss further implications for future work or research. The expectation is that most of the work has been completed and any major results are available to share with an audience. At this stage you should be able to tell a story about your project, even if parts of it have not been completely implemented or solved. The title of the presentation should include the course title and an indication that this is the project's final in-class presentation, the title of the project, your name and contact information. The following are tips and advice for preparing your presentation.

- You (or your group) will have 15 minutes to present your work in the classroom. Another minute will be allocated for question answering.
- You should prepare a set of ~15 slides, given that a slide should take a minute to talk about on average.
- Presentations should be organized into thematic units. A typical outline includes:
  - Motivation of the project, its importance and potential applications.
  - Definition of the problem (if any), including input, constraints, output and hardness
  - Main project approach ideas, data, methods and the type of analysis proposed
  - Highlights of the results (visual, experimental, theoretical)
  - Interesting variations and limitations of the analysis
  - Concluding remarks
- The talk should be self-sufficient, meaning that you should not make any assumption about prior knowledge of the audience or previous well-known results. All concepts should be introduced and appropriate notation should be used consistently throughout the presentation.
- Focus on the essential parts of the project and avoid too-many technical details. The goal is to give a summary of the project and convey the contribution of your work to other people. At the same time, you should make sure that important content is adequately covered.
- Prepare the slides carefully. Text should be easily readable and slides should not be overloaded with content. Avoid full text sentences and use of math symbols, unless necessary.
- **Practice** the talk several times, and time yourself to make sure you are within the time bounds.

More Advice & Tips

Some interesting advice on how to give a bad talk by David A. Patterson (UC Berkeley):
https://people.eecs.berkeley.edu/~pattrsn/talks/BadTalk.pdf

Some design tips for beautiful presentations:
https://visage.co/11-design-tips-beautiful-presentations/

PowerPoint Templates

Please find below links to PowerPoint Templates. These are designed with the York brand's primary and secondary colours and fonts embedded for ease of use and can be customized. Template files are available in a widescreen (16:9) format. Branded layouts featuring the York colours and campus photos are also included for title and section slides. This PowerPoint template is AODA compliant (last updated January 2021)

Widescreen with built-in instruction (.potx)

Widescreen without built-in instruction (.potx)

How to Submit?

```
% submit 6414 project final-presentation.pptx final-presentation.pdf team.txt
```