Project

1. **Final report** An eight page report with additional page for references in the Neurips format (`https://nips.cc/Conferences/2020/PaperInformation/StyleFiles`). The final report should have the following structure.

   (a) Introduction (one paragraph on each item below)

   - State the problem you are trying to address.
   - Provide the motivation.
   - Provide a high-level overview of the three methods that you will use to analyze data.
   - Discuss the dataset you will analyze.

   Use the above as a checklist. Create a section for each of (a)-(d) above. This is a typical structure of a scientific paper. Your grade will reflect how closely you follow this structure.

   (b) Methods

   - Describe the three methods you applied to the data in detail, one in each subsection.

   (c) Results

   - Provide results as figures and tables. Each figure and table should have a caption.
   - In the text, explain your finding in figures. In each paragraph, discuss how each result was generated, make observations from the results in the figure, and draw mini conclusions.

   (d) Conclusions

   - Summarize your finding.
   - Discuss future work.

2. **Project proposal**: You can think of the project proposal as a short draft of the introduction of the final report. You will further revise this before submitting the final report at the end of the semester. The proposal should be in the Neurips format, with the title, team members as authors, and the introduction section. You can think of this as writing the first page in the Neurips format.

3. **Project presentation**: Each team will give a 10 minute presentation in class at the end of the semester to share your results with the rest of the class. You should prepare 10-15 slides, following the same structure as the final report listed above: 1-2 slides for introduction, 3-5 slides for methods, 4-6 slides for results, and 1 slide for conclusion.

4. **Suggested datasets**

- Genotype-Tissue Expression (GTEx) data (`https://www.gtexportal.org/home/`): Human gene expression data for many different tissue types. You can download the gene expression data from the web portal. Note that you cannot access the genetic data, as this requires an approval. You could perform unsupervised learning (clustering or network estimation) on the data from a single tissue type. You could compare your finding across multiple tissue types.
  - The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science, 2015.
- The Cancer Genome Atlas (TCGA) Data (`https://portal.gdc.cancer.gov/repository`): This is a consortium data with various types of genomic data collected from a large number of patients for many different cancer types. Gene expression, copy number variation, and epigenetic, clinical data are available for open access. You could perform unsupervised learning on a single type of data for different cancer types. You could perform supervised learning on two types of data. For example, you could predict gene expression levels based on copy number variation, or predict clinical outcome based on gene expression and/or copy number variation.
  - Hutter et al. The Cancer Genome Atlas: creating lasting value beyond its data. Cell, 2018.
- COVID cases (`https://github.com/CSSEGISandData/COVID-19`): This is a repository for covid case counts over time in different regions maintained by Johns Hopkins University. You could model the time-series of the covid case counts to predict the covid case counts in the future. You could perform spatio-temporal analysis by modeling both geographic and temporal aspects of the data.
- Mouse advanced intercross data (`http://palmerlab.org/protocols-data/`): The expression data appeared in Homework 2. Data are available for both brain gene expression and genotypes for the mouse population after 50 generations of inbreeding between two inbred founder mice. You could perform supervised learning to predict gene expression levels based on the genetic data.
  - Gonzeles et al. Genome-wide association analysis in a mouse advanced intercross line. Nature Communications, 2018.
- Human Cell Atlas (`https://www.humancellatlas.org/`): Human single-cell RNA-seq data for various organs are available in the data portal. This is similar to GTEx data above, but was collected at the single-cell level. A similar type of analysis described above for GTEx data could be done. Unlike the GTEx data, this has a large number of missing data points, so you will have to impute the missing values for analysis.
  - Rozenblatt-Rosen et al. Building a high-quality Human Cell Atlas. Nature Biotechnology, 2021.
- Allen Brain Map (`https://portal.brain-map.org/`): Various data types from human and mouse brains are available.
  - Shen et al. The Allen Human Brain Atlas: Comprehensive gene expression mapping of the human brain. Trends in Neuroscience, 2012.
  - Jones et al. The Allen Brain Atlas: 5 years and beyond, Nature Reviews, 2009.

5. Methods: As a team of two, we ask you to apply at least three different methods on the dataset of your choice. You should implement two of the three on your own. For the other

one, you are welcome to use an existing package. You are welcome to try out other packages in addition to the required minimum three methods.

6. Some suggestions

   - Feel free to find a different dataset that you find interesting.
   - In all data analysis project, one should take time to learn about the data. Go to the data portal and spend some time browsing the data. Go to `scholar.google.com` and search using the name of the dataset above as a keyword. You will find many papers published for the given dataset that you may find useful.

7. **Latex tutorial** The (not so) short introduction to Latex. `https://gking.harvard.edu/files/lshort2.pdf`

8. **What to submit** For the final report, please submit the report itself, the codes of the two methods you wrote with a brief instruction on how to run them, and the slides from the presentation.