Project Guidelines

Our Project Focus

- Big data driven
- Big data sources
  - too large for your computer — at least GB level (1GB ~ 8 blocks!)
  - more than one data sets with spatio-temporal components
- Data Analysis and Mining at scale
  - classification, regression, clustering, dimensionality reduction
- Communication, Presentation and Reproducibility

Big data driven

- You must perform data preparation (Phase 2) using big data platforms
  - using any frameworks within the Hadoop ecosystem
- Modeling (Phase 4) can be done without “big data” depending on your problem
  - e.g. creating plots from the data prepared in Phase 2
- Phase 2 BIG and Phase 4 SMALL is okay but not the other way around
  - Discuss with me if your group really needs to pursue this

Big Data Sources

- Must use more than one data sources, combined
- Total size at GB level
- Should have both space and time component, separately or by joining them together
  - important to have good spatio-temporal coverage (e.g. not just a day of trip data)
- Encouraged to use data posted on the CUSP Data Catalog
  - but external data are also welcome
- Open data are preferred (or a must if using NYU ITS cluster)
  - Proprietary data sets have to be ingested through the CUSP Data Facility

Data Analysis and Mining

- Tasks are derived from the main objectives of the project (hypothesis!)
  - e.g. for the objective of study taxi driver behavior, the main task would be performing classification of pick-up and drop-off patterns across time and space
- Many techniques in mining, recommendation, classification and clustering can be implemented using big data platforms
  - e.g. available techniques on Spark
- Descriptive analysis are okay but must be comprehensive (fact book)
  - TLC Taxi Fact Book, Rudin Center's Citibike report

Presentation and Reproducibility

- Must drawing some insights from the analysis results (to support or deny your objective hypothesis)
- Clearly document the methodologies and data sources
- Commit all the code into a GIT Repository
- Deliverables are presentations and reports

Logistics

- Can work in teams, of at most 4 members
  - There must be enough work to split among the team members!
- All members will receive the same grade by default
- Can share projects with other classes (as long as your project meets the big data requirements)
- Can work on either NYU CUSP or Amazon EMR cluster
- Project materials must be committed into a GIT Repo

Project Proposal

- Motivation and project objectives
- Data sources
- Methodologies: how are you going to tackle the challenges?
- What are the big data challenges?
- What will be the project deliverables?
  - Good to have some mock-up or proof-of-concept visualizations
- A written document and a 10-min presentation

Final Project Submission

- Updates from the proposal
- What are the outcome of your project? What insights did you learn?
- How did you tackle your big data challenge?
  - Present your solution in details
  - What are the bottlenecks?
- A detailed report and a presentation

Accounts at NYU-CUSP

- Everyone should receive an email about your account
- First, please reset your password

  https://serv.cusp.nyu.edu/ipa/ui/reset_password.html
- Second, sign the Term of Use agreement

  https://datahub.cusp.nyu.edu/forms/terms-of-use.html
- Please do not request for access on the hub (you already got an account)
- You will need your account for the lab