**9 Heilmeier questions (**[**source**](http://en.wikipedia.org/wiki/George_H._Heilmeier)**)**

1. What are you trying to do? Articulate your objectives using absolutely no jargon.
2. How is it done today; what are the limits of current practice?
3. What's new in your approach? Why will it be successful?
4. Who cares?
5. If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, ground truth data, etc.)?
6. What are the risks and payoffs?
7. How much will it cost?
8. How long will it take?
9. What are the midterm and final "exams" to check for success? How will progress be measured?

**Some Rough Guidelines (from a Piazza TA)**

Hi class,

I wanted to add more information to help with writing the Proposal Document.

The page limit is 2 pages excluding the references and must be in PDF format. Each person has to review at-least 3 papers. You may consider organizing your proposal based on the Heilmeier questions (e.g., each section addresses one question). You may also consider combining multiple references to fit within the page limit.

For the literature survey, please make sure you focus on the important areas so that you don't go over the page limit with extraneous details.

Few things to keep in mind:

* We want you to mention some conclusions / findings from the literature and how those impact the approach for your project.
* This should be a synthesis, so get the most important facts / assumptions from each paper first and then continue to add more details relevant to your project and structure to your paragraphs accordingly.
* We want to see that you have done research on your topic and using that existing research to back up and bring your idea to life. Hence, the requirement on the number of papers.
* Your survey can be combined with the Heilmeier questions. If that structure works better for your proposal and the page limit, feel free to use it.
* Please do follow the page limit requirements.
* You can read more about how to find papers for the literature survey in this thread<https://piazza.com/class/kd50xrr0s5l65a?cid=93>

And as always, if there are any more questions, please ask in this post!

Project Ideas:

* Education - Why does America suck at basically everything. Education success/outcomes in USA vs other countries (i.e. China), education success by various locations in US, educations success by teaching method in US
  + Countries - More than 1
    - Kshitij will look up some data for education.
    - Schools -
      * Economic
      * Academic Performance:

1. Student performance in virtual learning environment. Seems interesting and relevant. Example analysis: <https://analyse.kmi.open.ac.uk/resources/documents/mashupExample.pdf>. Dataset location - <https://analyse.kmi.open.ac.uk/open_dataset#description> - medium sized dataset
2. <https://data.nysed.gov/downloads.php> - New York State Education Department Data. Decent size if we consider data for multiple years
   1. Heatmap
   2. Geograpical map
   3. Time plot
      * + Demographics
      * Charter Schools?
      * Questions
        + Charter schools vs. public schools for graduation rates, college acceptance rates? Controlling for socioeconomic status
        + Geographical trends for helping students with lower socioeconomic conditions, disabilities have better outcomes?
        + Why do certain schools deviate from state baseline in graduation rate? (controlling to not build a racist model)
        + Expenditure per student, staff qualifications, average income, and outcomes
        + Urban, suburban, rural - geographic
      * Data Sets:
        + <https://datacatalog.worldbank.org/dataset/global-data-set-education-quality>

* Natural Sciences:
  + Bee science - Nick and Ben
    - Problem / issue to analyze
      * [Science article from Feb2020 on climate change and bee declines](https://science.sciencemag.org/content/367/6478/685) (can access through GT library)
    - Available data
      * [Data repository for science article](https://figshare.com/articles/Climate_change_contributes_to_widespread_declines_among_bumble_bees_across_continents_-_DATA_REPOSITORY/9956471) (seems like it may be large enough, includes one .csv that is 95,000 observations and 8 dimensions)
      * Many observations, few features. Change in species by temperature and location.
      * Combine with agricultural data over time.
  + Frogs / climate - Nick
    - Problem / issue to analyze
      * [Possible increase in a spread of a frog disease](https://www.bbc.com/news/science-environment-48219217#:~:text=Climate%20change%20is%20having%20an,vanish%2C%20according%20to%20a%20study.)
      * [Nature paper on threads to global amphibian diversity](https://www.nature.com/articles/nature10650)
      * [USDA article on amphibians and climate change](https://www.fs.usda.gov/ccrc/topics/amphibians-and-climate-change)
    - Available data
      * [Frog survey at Yosemite NP, ~7600 observations and 12 dimensions (probably too small)](https://datadryad.org/stash/dataset/doi:10.5061/dryad.rm382)
      * Datasets overall are pretty small for this topic -- option to use a climate change dataset w/o frog data and relate it to frogs
  + Air Quality Data - Coal Plants + Asthma?
  + Avalanche Prediction - driven by available data - Dave will find some Data
    - <https://www.cambridge.org/core/journals/annals-of-glaciology/article/applying-machine-learning-methods-to-avalanche-forecasting/FBCEFDC8F6A8A81B40369AAB8BAA8B93/core-reader>
    - Not a lot of big enough data for this problem
    - Some snow science data available (snowtell?)
    - Major GIS problem, may not be feasible
  + MicroBiome - Effect of gut health/biome on health outcomes
    - Problem Description
      * Use microbiome strains as features to predict holistic health outcomes.
      * Existing literature has studied: autism, mental health, IBD and geographical distribution. Do Colorado biomes make people more fit than Houston biomes :)?
      * Machine learning and data analysis is a proven useful tool in general in this area:
        + <https://medium.com/tribalscale/early-diagnosis-of-autism-with-microbiome-and-machine-learning-2e90abb2437e?source=rss------artificial_intelligence-5>
        + <https://medium.com/tribalscale/early-diagnosis-of-autism-with-microbiome-and-machine-learning-2e90abb2437e>
        + <https://towardsdatascience.com/visualizing-high-dimensional-microbiome-data-eacf02526c3a>
        + <https://www.thecardiologyadvisor.com/general-cardiology/leveraging-machine-learning-for-gut-microbiota-based-cvd-screening/>
        + <https://anesthesiology.duke.edu/?p=846744>
    - Data
      * Readily available: <https://github.com/cduvallet/microbiomeHD>
      * Other
        + <https://www.arb-silva.de/>
        + <https://qiita.ucsd.edu/static/doc/html/index.html>
* Transportation - Nick & Dave Will look up on this.
  + CDOT meeting with data science guy from MIT -- Eric Subina, 30 min, 01Oct2020
    - Option 1-- GIS Data Format
      * GIS datasets are in different format/structure depending on source
      * Hard to tell what differences in structure in data are between sources
      * Can we use deep learning to detect dataset format in any way? To understand which datasets are in similar formats?
      * Build a graph of similar data formats?
    - Option 2 -- Visualize Data Sources
      * Interactive checklist/viz about difference sources of data used by CDOT
    - Option 3 -- Detour Mapping
      * PhD MIT Scott Hamming
      * Re-routing traffic during road closures (flood, rocks, avalanche)
      * Quantify impact of detours and capacity
      * Interactive map with blocked areas (manually add a flood onto the map)
      * Redundancy rating of roads (based on other roads available for detour)
      * Consider as a network directed graph with nodes/edges with different degrees. Edges can be defined as distance, time
  + Colorado DOT meeting with Lizzie Kemp, 1 hr, 01Oct2020
    - Flooding and rockfall are two biggest risks to highway system
      * Flooding is user side, rockfall is user + infrastructure
      * 700MM in asset damage in 2013 floods
      * Brain varrella, hydraulic engineer working on flood recovery, involved in 2013 efforts
    - Rockfall analysis is an underdeveloped problem, not a lot of available data
    - Dave’s spitballing on project idea
      * Map of denver with all roads in interactive web file
        + Focus on state highways
        + Rank redundancy index of roads, if they get closed how screwed are we?
      * Circle is a flood -- re-reoute traffic around circle
      * Use flood plane data to predict where floods may happen
        + Per Lizzie not available everywhere in Colorado
        + Where it is available, easy to get
      * Calculate optimal routes (option 1, 2, 3)
      * Eric sabina and Scott Ramming working on this problem
* Finance - Vuong will come up with
  + a. Topic and
  + b. Data Ideas

Primary Proposal Question:

What are the most important school, economic, or academic performance features that inform a notable deviation from NYS wide graduation rates in each county.

County Aggregate Income Statistics:

<https://labor.ny.gov/stats/employ/county.shtm>

**Meeting 04Oct2020 on refining project idea**

* What responses do we need to include in the proposal
* Ben to add current state of machine learning / practices into the “current practices” section of the paper
* Success criteria and evaluations and citations and bibliographies
  + Ben and Kshitij to help out