Data Incubator Challenge Prep

1. **For one question you will be asked to propose a project for the fellowship**
   1. **A well formulated outline of the project goals and deliverables.**
   2. **Evidence of some exploratory data analysis.**
   3. **Two interesting graphics.**
   4. **General** 
      1. While your project does not have to be completely original, you should Google around to see if your analysis has been done to death. (Look back at the brainstorming Lawrence and I did)
      2. You can use design thinking techniques to consider how different kinds of end users will interact with your project and how it will add value they couldn’t get somewhere else (Focus on something with product/user implications)
      3. projects that require pulling data an API or scraping a webpage. (Civitas data or something new where I need to do that alternatively?)
      4. All things being equal, analysis of larger datasets is more impressive than analysis of smaller ones. Real world problems often involve working on large, multi-gigabyte (or terabyte) datasets (size matters)
   5. **Brainstorming** 
      1. **Notes**
         1. Avoid panels if possible for the time being -> Can do in the end but will prob give you a headache right now for this application
      2. Urban
         1. Inclusionary zoning
            1. Upzoning and increasing density **Building permits.** The Census Bureau’s [Building Permits Survey](https://www.census.gov/construction/bps/definitions/) collects data from [thousands of municipalities](https://www.census.gov/construction/bps/how_the_data_are_collected/) every month. For each municipality, metro area, and state, the datasets provide the number of permits issued for new residential housing, number housing units authorized, and total estimated value of the new construction. **Previously:** [The Census Bureau’s Annual Characteristics of New Housing survey](https://tinyletter.com/data-is-plural/letters/data-is-plural-2016-06-22-edition)(DIP 2016.06.22). [h/t [Susie Cambria](https://twitter.com/susiecambria) + [Issi Romem](https://www.buildzoom.com/blog/paying-for-dirt-where-have-home-values-detached-from-construction-costs)]
            2. <https://inclusionaryhousing.org/map/> (Could try to predict market elasticity or affordable housing based on all these policies) (Could merge with building permits data perhaps)
            3. To look at just DC: <https://opendata.dc.gov/datasets/affordable-housing/data>

But only 416 projects

* + - 1. Transit (If we’re serious about Civitas MVP)
         1. **Public transit, curated.** As a way to “lower the barrier“ for analyzing public transportation data, researchers at Finland’s Aalto University [have published](https://www.nature.com/articles/sdata201889) “a curated collection of [now more than] 25 cities' public transport networks in multiple easy-to-use formats including network edge lists, temporal network event lists, SQLite databases, GeoJSON files, and the GTFS data format.” On [the project’s website](http://transportnetworks.cs.aalto.fi/), you can browse, visualize, and download each city’s data. (The cities are mostly in Europe and Australia, but also include Detroit, Winnipeg, and Antofagasta, Chile.) **Previously:** [TransitLand](https://transit.land/) and [TransitFeeds](https://transitfeeds.com/) ([DIP 2016.07.27](https://tinyletter.com/data-is-plural/letters/data-is-plural-2016-07-27-edition)). [h/t [NYU Data Science Community Newsletter](https://cds.nyu.edu/newsletter/)]
      2. Predicting pricing
         1. Dataset: Financial Tweets ([link](https://kaggle.us1.list-manage.com/track/click?u=e4c8fb8b43860678deab268e5&id=51a6a4065f&e=5f49d31401)) （Merge with something else）
         2. **S&P 500 Simple Forecasting with Prophet (**[link](https://kaggle.us1.list-manage.com/track/click?u=e4c8fb8b43860678deab268e5&id=0c46c799e4&e=5f49d31401)**) (Can merge the above but the financial tweets dataset might already have it)**
      3. Break down valuation to its components and see if anything is in there
         1. **Financial statements.** The SEC’s [Office of Structured Disclosure](https://www.sec.gov/structureddata) publishes [data extracted from corporations’ public financial state ments](https://www.sec.gov/dera/data/financial-statement-data-sets.html). That dataset contains the numbers listed in each company’s primary financial statements — balance sheets, cash flows, et cetera. An [even more detailed version of the dataset](https://www.sec.gov/dera/data/financial-statement-and-notes-data-set.html) includes plain-text notes from the filings, plus numbers from a broader array of forms. Both datasets are updated quarterly and go back to 2009. [THIS IS AWESOME. JUST WOULD TAKE A LOT OF REFINEMENT ON MY PART TO TRY TO FIGURE OUT WHAT TO PREDICT]
         2. **Publically Traded Market Data:**[Quandl](http://www.quandl.com/) is an amazing source of finance data. [Google Finance](http://localhost:9000/project/www.google.com/finance) and [Yahoo Finance](https://finance.yahoo.com/) are additional good sources of data.  Corporate filings with the SEC are available on [Edgar](http://www.sec.gov/edgar.shtml).
         3. Something with predicting capital allocation structures for public companies

Go directly through Quandl R api if we want anything to couple with the valuation stuff

* + 1. Other
       1. What if we actually did the music application we talked to Jack about? (Do this later but not now)
          1. <https://labrosa.ee.columbia.edu/millionsong/tasteprofile?fbclid=IwAR3Kc7qsteW61KcAYa9AI7gFjqndxtc2SmZaLAKc_CPBeyapc2hr6tEyTaU>

<https://labrosa.ee.columbia.edu/millionsong/challenge>

* + - * 1. <https://pudding.cool/2018/05/similarity/?fbclid=IwAR3C5-M4kGPNZUJUIP4xg0HpKi1iANaH3nktjfqAC1vpuSLUeCY0heT3De4>
        2. <http://people.ischool.berkeley.edu/~nikhitakoul/capstone/index.html>
        3. <https://qz.com/797313/how-to-invest-according-to-rappers/>

With 2019 will come the IPO of ride-sharing company Uber, which many are now calling a “decacorn” due to its $10 billion valuation. The much anticipated IPO will represent the third largest IPO in the history of NYSE behind Facebook and Alibaba, and other hot tech companies like Airbnb and Pinterest will not be far behind Uber in their own respective paths to IPOs. While these companies have acquired massive userbases through competitive strategies in monopolistic markets, what is most interesting about these companies is that of the dozen unicorns (valuation of at least $1 billion) that went public last year, they posted a combined $14 billion in losses. Ten years ago, only about 33% of companies pursuing IPOs had no recorded profits. Today that number is a staggering 84% of companies pursuing IPOs.

Naturally, many who suffered through the dot-com bubble see the current trend in revenue-neutral or revenue-negative IPOs as eerily familiar. However, the major difference is that while there was only $6 billion in VC funds circling between startups in the US in 1994, that number has now reached over $130 billon. Despite the negative profit and loss statements, the allure of massive user bases and the greater scale of VC funding means that companies are able to carry on in the red for far longer than they used to be before funders start to demand a return. As we saw with the disastrous IPO of ephemeral messaging platform Snapchat in 2017, however, the great wall protecting revenue-negative firms like Uber is starting to crack, causing many to focus more on whether these companies might be massively overvalued.

The objective of my proposal therefore is to use financial statement data on hundreds of public companies to predict whether firms or over or undervalued and identify which features on balance sheets, income statements, and cash flow statements have the greatest predictive power in determining overvaluation and undervaluation. Armed with this knowledge, investors can better protect themselves from looming bubbles arising from a critical mass of overvaluations. At the same time, by accessing financial metrics beyond the questionable daily active users (DAUs) indicators, investors can better identify value investment opportunities that can guarantee significant returns in the long-run. This of course all rests on the well accepted argument that the stock market does not in fact behave in accordance with the efficient market hypothesis and it is therefore possible to produce predictions the market has not taken into account.

To accomplish the above objective, I will use the eXtensible Business Reporting Language (XBRL) dataset that scrapes line-item information from the financial statements publically traded companies are required to publish annually under SEC regulations. The SEC established the XBRL data collection program to address the issues that cause stock mispricing and it is therefore perfect for accomplishing my objective. The full dataset includes information on 122,815 financial statement items collected across 5,711 companies dating back to 1985 and up through 2018. There are a number of potential target variables we can use, but we will mainly focus on the common stock price and the earnings per share price initially.

Within the realm of supervised learning, there are a number of different data science methodologies that can be applied to accomplishing my core objective, and I will initially organize those approaches based on the two ways of organizing the data for specific insights.

In the first approach, I will organize the data into a panel consisting of a cross-section of publically traded companies and a time series of years dating back 20 years capturing 1998 to 2018. Because there is missingness across companies and years, I will attempt two approaches to address missingness. In the first approach, I can conduct multiple imputation for missing values using the Amelia Package in R. This process uses bootstrapping and an Expectation-Maximization algorithm to impute the missing values in a data set. In the second approach, I will evaluate which financial statement variables are best represented in the data by calculating a proportion through dividing the frequency of an indicator and the data and dividing it by how much the indicator would appear if it was included for every company and every year. After eliminating the indicators that fall below 50% coverage, I will then subset the dataset to eliminate the companies that do not have 50% of the subsetted indicators across the time periods in the data.

Once I’ve addressed missingness, I will split the data into train and test sets where the test set comprises the last year of the data. At this point, I will employ Lasso regularization to evaluate which financial statement indicators to include in the middle to predict the common stock price or earnings per share. After dropping the variables that are kicked out in the lasso process, I will then develop a Panel Linear Regression model to predict the stock prices. To fine-tune the model, I will lag the dependent variable to consider the possibility that the previous year’s stock price could be the best predictor of this year’s stock price. To asses how well the model performs over all, I will then calculate the Mean Averege Percent Error (MAPE) to see accurately the model uses training values to predict rental prices in the test set.

In the second approach, I will restrict the data to one year of data. After applying the same lasso process laid out above for feature selection, we will employ a variety of methodologies to identify the best model to train. We will use decision trees to subset the population into smaller more homogenous subpopulations using the input features. We will grow a forest of decision trees where observations are bootstrapped and variables are sampled using the random forest approach. We will evaluate a number of other approaches as well, including gradient boosting and bagging approaches.

Follow Up research

1. I’m kind of taking a moneyball approach: Using statistical analysis, the A’s competed by buying assets that were undervalued by other teams and selling ones that were overvalued by other teams.
   1. So whereas they focused on number of runs (gained and lost) as the dependent variable, I’d be focusing on stock price
   2. They used strong batting, I’d use profit drivers
   3. They used strong fielding and pitching, I’d use cost saving measures
   4. Goal: To find out (predict)if the OaklandAs will make it to the 2002 playoffs, using past data
      1. First we establish what it takes a baseball team to make it to the playoffs. We do that by showing thata team has to win at least 95 games. What do the pink, orange and red lines tell you?
      2. What does it take to win games? It requires a team to score more runs and give away fewerruns in absolute as well as relative (to otherteams) terms.In essence, the run difference (runs scores – runs allowed) is established as a useful metric for predicting the number of wins.
      3. Having established that run difference (for a team)is a great predictor of the number of wins (in a season),we now turn to developing models to predict runs scored and runs allowed (against otherteams)
      4. We start by choosing all data up until 2001. It is generally accepted that batting average (BA)is a good predictorfor runs scored (for a player) and, consequently,for a team.Oakland also considered “on base percentage”(OBP) and “slugging percentage” (SLG)to be useful predictors of runs scored.Themodelthapredictsruns scored (RS)usingBA,OBPandheSLGis:
      5. What stands outhere is thatthe coefficient of BA (-368.97)implies an inverse relationship between runs scored and batting average. This can not be right. It so happens that BAis highly correlated with bothOBP andSLG (see correlation matrix below):
      6. The modelfor runs allowed(RA) is built on a similar logic. The predictors include the “other teams on base percentage”(OOBP) and the “other teams slugging percentage”(OSLG). That model comes outto be:
2. <https://www.sec.gov/dera/data/financial-statement-data-sets.html>
   1. The Financial Statement Data Sets below provide numeric information from the face financials of all financial statements.
      1. This data is extracted from exhibits to corporate financial reports filed with the Commission using eXtensible Business Reporting Language (XBRL).
      2. As compared to the more extensive Financial Statement and Notes Data Sets, which provide the numeric and narrative disclosures from all financial statements and their notes, the Financial Statement Data Sets are more compact.
3. <https://www.sec.gov/dera/data/financial-statement-and-notes-data-set.html>
   1. The Financial Statement and Notes Data Sets below provide the text and detailed numeric information from all financial statements and their notes.
      1. This data is extracted from exhibits to corporate financial reports filed with the Commission using eXtensible Business Reporting Language (XBRL)
      2. As compared to the more compact [Financial Statement Data Sets](https://www.sec.gov/dera/data/financial-statement-data-sets.html) which provide only the numeric information from face financials, the Financial Statement and Notes Data Sets provide significantly more disclosure data
4. <https://www.sec.gov/files/aqfs.pdf>
   1. These data sets currently include quarterly and annual numeric data appearing in the primary financial statements submitted by filers. (Obv need to eliminate all quarterly reported data and focus just on annual)
   2. The data extracted from the XBRL submissions is organized into four data sets containing information about submissions, numbers, taxonomy tags, and presentation.
      1. • TAG – Tag data set; includes defining information about each tag. Information includes tag descriptions (documentation labels), taxonomy version information and other tag attributes.
      2. • PRE – Presentation data set; this provides information about how the tags and numbers were presented in the primary financial statements.
      3. So the big question is whether tag and/or pre could be used to synchronize financial statement line items that have slightly different names but are the same for all intents and purposes
   3. The scope of the data in the financial statement data sets consists of: • Numeric data on the primary financial statements (Balance Sheet, Income Statement, C ash Flows, Changes in Equity, and Comprehensive Income) and page footnotes on those statements;
      1. So I want to somehow eliminate the changes in equity and comprehensive income items -> This should be in tag or pre
      2. Might be best to just organize directly into the interactive data attachments themselves -> 10-K, 10-K/A, 10-KT, 10-KT/A, 10-Q, etc.
   4. Useful from Sub
      1. Adsh- Accession Number. The 20- character string formed from the 18-digit number assigned by the SEC to each EDGAR submission
5. <https://www.sec.gov/info/edgar/edgartaxonomies.shtml>
   1. We only want GAAP submissions, and so we should also eliminate any non-Gaap submissions
      1. I am even tempted to restrict just to 10-k to start as that is the standard and would most likely be cleanest
         1. We’d still end up with 3890 out of 4351 -> Think about it
6. Merging in Corporate Bond Rating Data
   1. I’ve found historical corporate bond rating that could give me the bonds for companies for year end 2017 and 2018. However’ the problem is the unique identifier for merging
   2. While we can try to merge on company name text, that often doesn’t work due to inc., llc, acronyms, or any number of reasons
   3. So here is what we’ve got other than that
      1. Financial Statement Datasets: adsh (but that’s for the EDGAR submission so one company will have more than one), cik (assigned by SEC to each registrant submitting filings), ein (IRS assigns to business entities operating in the US)
      2. Corporate bond rating datasets: Legal entity identifier
7. So what we need is some king of ein or cik to lei crosswalk
8. <https://bipartisanpolicy.org/wp-content/uploads/2018/07/Employer-Data-Matching-Workgroup-White-Paper.pdf>
   1. There are four primary unique identifiers currently in use by Federal Agencies: the Employer Identification Number (EIN), Data Universal Numbering System (DUNS©) numbers, Commercial and Government Entity (CAGE) codes, and the Legal Entity Identifier (LEI).
      1. We could use Dun and Bradstreet like we used to at AED to upload in company names and get DUNS if that is better… Don’t remember if Dun and Bradstreet gave LEI – don’t think they did
9. Using the more extensive SEC dataset
   1. <https://www.sec.gov/files/aqfsn_1.pdf>
      * 1. Still use usual suspects

**Final Interview Prep**

The Project

1. **Introduction for Project:** Why should your project be interesting to a broad audience? Being able to relate your project to a personal interest or experience makes the message more interesting.

Ride-sharing company Uber IPOed with a valuation of close to $70 billion last week, a number which actually fell short of investor’s expectations. The much anticipated IPO represented the third largest IPO in the history of NYSE behind Facebook and Alibaba, and other hot tech companies like Airbnb and Pinterest will not be far behind Uber in their own respective paths to IPOs.

While these companies have acquired massive userbases through competitive strategies in monopolistic markets, what is most interesting about these companies is that of the dozen unicorns (valuation of at least $1 billion) that went public last year, they posted a combined $14 billion in losses. Ten years ago, only about 33% of companies pursuing IPOs had no recorded profits. Today that number is a staggering 84% of companies pursuing IPOs.

Naturally, many who suffered through the dot-com bubble see the current trend in revenue-neutral or revenue-negative IPOs as eerily familiar. Indeed, we already saw one crack in the great wall protecting revenue-negative firms like Uber with the disastrous IPO of ephemeral messaging platform Snapchat in 2017.

The objective of my proposal therefore is to use financial statement data on hundreds of public companies to predict whether firms are over or undervalued and identify which features on balance sheets, income statements, and cash flow statements have the greatest predictive power in determining overvaluation and undervaluation.

Armed with this knowledge, institutional and hobbyist investors alike can better protect themselves from looming bubbles arising from a critical mass of overvaluations. At the same time, by accessing financial metrics beyond the questionable daily active users (DAUs) indicators, investors can better identify value investment opportunities that can guarantee significant returns in the long-run.

1. **Specifics of data set and size**

To accomplish the above objective, I am using the eXtensible Business Reporting Language (XBRL) dataset that scrapes line-item information from the financial statements publicly traded companies are required to publish annually under SEC regulations. The full dataset includes information on 122,815 financial statement items collected across 5,711 companies dating back to 1985 and up through 2018.

1. **My current approach and progress on the overvaluation prediction project**
2. Load in the four financial datasets
3. Subset to the appropriate predictor data
   1. Subset to just the standard 10-K SEC reporting form
   2. Subset to just US GAAP reporting (i.e. standard)
   3. Subset to just annual data (i.e. not including quarterly reporting)
   4. Subset to the three main financial statements -> income statement, cash flow statement, and balance sheet
   5. Subset to only include standard numeric indicators
4. Create the valuation variable that will be compared against the share price value to indicate whether a company is over, under, or appropriately values
   1. Instead of choosing only those vars with greatest coverage to serve as key predictors, now we subset to those IS and BS variables used in the DCF
   2. Method 1: Discounted Cash Flow
      1. Requires projecting out five years into the future and discounting future cash flows by a weighted average cost of capital (WACC)
      2. Once we add up the sum of the projected five years and the terminal value estimate (a company is not expected to just close in five years), we subtract debt and divide by the number of shares outstanding to get the valuation
      3. I was unable to do this because of challenges getting the data needed to calculate the WACC, namely the corporate bond yields (I got the ratings), the Betas, and the Market Risk Premium
         1. I have already identified the Wharton Research Data Service Bond Returns database as the best source for the bond yield data and I’ve reached out to my MBA friend to see if she can help get me access
         2. Beta can be manually calculated and would require I get the ratio of the moving average of the company’s stock and the S&P 500 index
   3. Method 2: Multiples
      1. We calculate EV and divide by the EBIAT or EBITDA and then multiply that by the EPS to get the valuation
5. Where from there (Wing it)
   1. Amelia: In the first approach, I will organize the data into a panel consisting of a cross-section of publically traded companies and a time series of years dating back 20 years capturing 1998 to 2018. Because there is missingness across companies and years, I will attempt two approaches to address missingness. In the first approach, I can conduct multiple imputation for missing values using the Amelia Package in R. This process uses bootstrapping and an Expectation-Maximization algorithm to impute the missing values in a data set.
   2. Feature selection and regularization (Going back to all the possible indicators)
   3. Once I’ve addressed missingness, I will split the data into train and test sets where the test set comprises the last year of the data. At this point, I will employ Lasso regularization to evaluate which financial statement indicators to include in the middle to predict the common stock price or earnings per share. After dropping the variables that are kicked out in the lasso process, I will then develop a Panel Linear Regression model to predict the stock prices. To fine-tune the model, I will lag the dependent variable to consider the possibility that the previous year’s stock price could be the best predictor of this year’s stock price. To asses how well the model performs over all, I will then calculate the Mean Averege Percent Error (MAPE) to see accurately the model uses training values to predict rental prices in the test set.
6. Identified challenges
   1. There is significant missingness that will require diving further into the xblr dataset and seeing why such discrepancies exist if every company is required to report
      1. Alternative would be to manually create web scraping crawl to get information from 10-Ks manually
7. **Takeaway:** There are two types of takeaways: technical (e.g. improved accuracy by 20%) and fun facts (e.g. charitable giving as a percentage of income is highest for the poorest and wealthiest Americans)

Questions

1. General background
   1. We expect Fellows to be in attendance for a standard 9 am - 5 pm work day, including occasional evening events
   2. We ask that you interview with our hiring companies during and immediately after the program
2. How does the recruitment and hiring process take place?
   1. Important to me as the pipeline is my top reason for taking the course
   2. My second reason is three primary skills:
      1. Python
      2. Big Data and Parallel Processing
      3. Data Engineering (This is where most of my challenges are)
3. We are required to work at hiring companies, correct?
   1. If we find an opportunity outside of those companies, what is that process like?
      1. Website says that we decline to do anything with companies not associated with the program
   2. Is there a list of hiring partners? 250+ evidently
      1. What is the geographic spread? How global? How about government vs private?
   3. Does the fellow only have access to those partners in the city where he or she does the program or globally?
4. How long has the placement process occurred on average?
   1. 1-2 months after the program on the website
   2. This is an unpaid opportunity, so the timeline is obviously very important to me
   3. I am applying for data science jobs now and feel I have the skillset to get in the door, so the big question is beyond the training, how will this make this process better for me?
      1. Resume prep, interview prep, and identifying employers is something career advisers do at Georgetown
      2. Do the employers that help fund the program guarantee spots for program graduates?
5. Is there any relationship with Amazon ahead of their entrance into the DC metro area market?