| **Patent Equity | MIT Collaboration Data Analysis** | |
| --- | --- |
| Contact | Jordi Goodman  Visiting Clinical Assistant Professor in the BU/MIT Tech Law Clinic  908-616-0497  [jordi@bu.edu](mailto:jordi@bu.edu) |
| Organization | BU School of Law |
| Organization Description | Founded in 1872, Boston University School of Law is a top-tier law school with a faculty recognized nationally for exceptional teaching and preeminent scholarship. You can explore virtually any area of the law in 200+ classes and seminars, 20+ clinics and externships, and 20 study abroad opportunities. BU Law offers a full-time JD degree, five LLM & master’s degrees, and 17 dual degrees. With the support of a global network of nearly 24,000 alumni and a robust Career Development & Public Service Office, our graduates achieve remarkable career success. BU Law is located in the heart of Boston and housed in the ultra-modern Sumner M. Redstone Building and Law Tower. |
| Project Type | Data Science |
| Project Description | Through a conference, I learned about the website<http://collaboration.mit.edu/>. It has a great database with all the professors in a department, the number of articles they’ve published, the number of conference proceedings, and the number of patents associated with each professor.  **I want to be able to show that there is (or isn’t) a correlation between the number of conference proceedings and/or articles and the number of patents by year.**  I’m also wondering if the race, gender, or tenure status of the professor makes a difference. I would love to download the information by professor and analyze how many articles, conference proceedings, and the number of patents are associated with each professor, and then determine by years at the faculty level, if the representation at conferences/on articles/on patents has improved  I already have labeled the race and gender of each professor on the website. |
| Data Sets & Sources | MIT Faculty Collaboration Tool   * <http://collaboration.mit.edu/>   Google Patents   * <https://patents.google.com/>   Value of a Patent Sources   * <https://patents.google.com/> |
| Suggested Steps | **One Approach**   1. Scraping data 2. Matching ratio to gender with X amount of certainty 3. Analyze   **Second Approach**   1. Use current data set (provided by Client) 2. Correlate each patent to a value |
| Questions to be answered in Analysis | 1. Is there a statistically significant disparity by race and/or gender in patent applications, patents, papers, and/or conference proceedings relative to the representation in each department or overall at MIT? 2. Has this disparity changed over time and, if so, in what way? 3. Are newer professors (professors who have been at MIT less than ten years) experiencing similar gaps to older professors? |
| Ideal Output + Final Deliverable | Ultimately the client would appreciate the following deliverables:   1. Analysis of data collected focused in on answering the questions listed above 2. A clean and searchable database that includes patents and patent dates and potentially also the longevity of the career of the MIT professor as well as race, gender, articles, and conference proceedings associated with each professor sorted by department. |
| Additional Information |  |
| Zoom Recording | 10/19 Client Meeting  [https://bostonu.zoom.us/rec/share/EvXHTWxYCP8UJB25t-jnuiWqR4rk5eaQZwE-6QwLkU\_AWnIaQr\_-d1XbsTlr8F1o.mGWlkf9LUIDCUuy](https://bostonu.zoom.us/rec/share/EvXHTWxYCP8UJB25t-jnuiWqR4rk5eaQZwE-6QwLkU_AWnIaQr_-d1XbsTlr8F1o.mGWlkf9LUIDCUuyN)N  (Access Password: zD\*d6Z0i)  11/2 Client Meeting  <https://bostonu.zoom.us/rec/share/8qCUiwqwi-fflyLvnNioD20n1dWzGtWpcx2_9xiA1WeBKRWbfcpcvZ-7_ZcjX0e5.yg_Yfztq2xwPltqv>  (Access Password: kZU\*9tGH)  11/9 Client Meeting  <https://bostonu.zoom.us/rec/share/jB92ispMLJjwCfKV5Eh6Gi3IQbgdl1BpxVn1QBWLs4sTvJeKY9zks4KGn2XltRO5.D1HaJii-I8uY_R0o>  (Access Password: @A4JO47.)  11/16 Client Meeting  <https://bostonu.zoom.us/rec/share/89CyD8WWhrPyQPblM2mxP3CCvV6j-4_lRJuLo-FTB8V-cfBrpF_KMK34o3QAM9hI.QTut2VFUfOmuZNaY>  (Access Password: n!sa\*5ty) |

**11/30 Last Client Meeting**

Note: D2 for Team 2

Who’s missing: Team1 - Liqiu Li [lm10@bu.edu](mailto:lm10@bu.edu); Team 2 - Taowen Dong [taowend@gmail.com](mailto:taowend@gmail.com)

* Please upload final presentation slide here:

<https://drive.google.com/drive/folders/1cE3Vjz9hcKWXgVXpHCZBLa5PN7C3COPl?usp=sharing>

**11/23 Meeting**

Last Meeting: 11/30 with Prof. Goodman

* Questions from Prof. Goodman

**Question 1:** **The question of the gender gap in patents vs. the gender gap in articles for MIT professors**

Team1: yes, gender gap in patents (male:5019+female:3086 ) > gender gap in articles (male:49386+female:42478)

Team2: yes, gender gap in patents (5572) > gender gap in articles(135,050)

Team3: yes, gender gap in patents (3873) > gender gap in articles (131307)

**Question 2: The academic age goes (from youngest to oldest): assistant, associate, professor, professor emeritus.**

**Deliverable 4 (12/1)**

This is a draft of your final report that has been reviewed by your client. It includes all visualizations, results, data, and code up to this point, along with proper documentation on how to reproduce your results, compile and use your codebase, and navigate your dataset. Your team will submit this as a PR.

**Suggestion:**

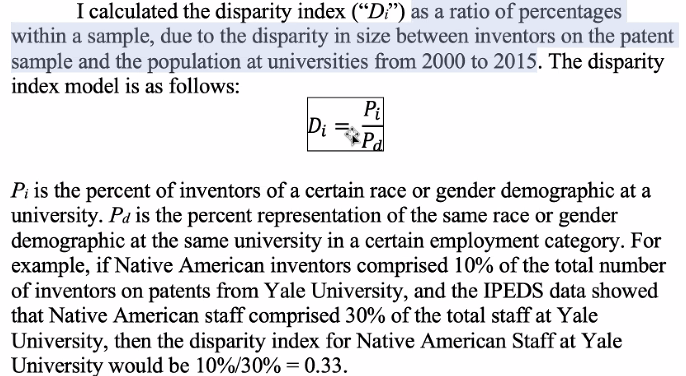
1. Give few hints on scraping the data; providing more guidance on scraping the data
2. Provide more help on coding
3. Challenging knowing what’s

**11/16 Client meeting**

Todos:

* Deliverable 3 deadline extension to 11/17
* Upload the early insight presentation slide to the google folder: <https://drive.google.com/drive/folders/1K8CG-yWTDHIksVsJ2s7lsv0JadXO4K-S?usp=sharing>
  + Name format: Team x Early Insight Presentation

* The overarching question in deliverable 3 is pointing to:
  + Informed the sample population at MIT, per capita patterns, take percentage of patents of m/w compare to the percentage of m/w;

****(example)

* Clean and searchable database: evey code and excel.

**11/9 Client meeting**

Anqi: check the cookie regulation

* Deliverable 2:

1. Race Classification Method
2. Preliminary analysis
3. Answer another key question
4. Include a list of departments you included

* Early Insight Presentation (11/15)

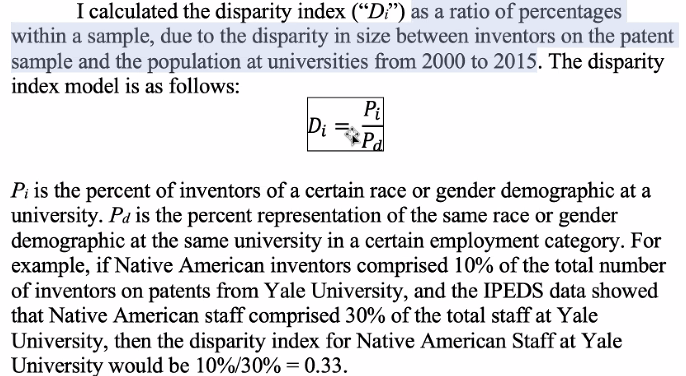
<https://piazza.com/class/ks4yrhftd3j237?cid=330>

* Deliverable 3 (11/17):

1. All data is collected
2. Refine the preliminary analysis of the data performed in PD1&2
3. Answer another key question
4. Attempt to answer overarching project question
5. Create a draft of your final report
6. Refine project scope and list of limitations with data and potential risks of achieving project goal
7. Submit a PR with the above report and modifications to original proposal

**Notes for deliverable 3:**

1. Informed the sample population at MIT, per capita patterns, take percentage of patents of m/w compare to the percentage of m/w;

****(example)

1. Where’s the greatest disparities in representation? (patents, article…) 2001-2020

* Check the paper in google folder for more information: <https://drive.google.com/drive/u/3/folders/1HRSs-78JZkJRZg3OEo8RGHKmBwzg8yHv>

1. Final report will exclude some departments that’s not relevant at the end.
2. Find the work year of the prof in MIT on IPED website (progress shown in 11/9 recording) <https://nces.ed.gov/ipeds/datacenter/CDS.aspx>
3. Excluding the doubles of professors if they worked for multiple departments
4. List the reason why you have excluded some departments

**11/9 Updates**

Team 2 (Deliverable 2 extended to 11/10 b/c data set)

**11/2 Client meeting**

Notes about the deliverables (other than the ones mentioned in syllabus):

\*Remember to include a contribution session in your PR

Deliverable 1:

1. Gender Classification Method
2. Data set
3. Preliminary analysis
4. Answer one key question

Deliverable 2:

1. Race Classification Method
2. Preliminary analysis
3. Answer another key question
4. Include a list of departments you included

Todo for Jordi and Anqi: Group the department

* Team 1 (11/3 - deliverable 1)
  + Almost done with web scraping, try to answer question
  + Gender Classification: train the dataset and generate the model; implement that to the MIT professors on the website. 80% accuracy now
* Team 2 (11/3 - deliverable 1)
  + Finished scraping the data
  + Found good classification method
  + Will implement the method and answer the key question
* Team 3 (have submitted deliverable 1)
  + Used model to predict gender according to their first names

**Reply from Prof. Goodman:**

* I am looking for a list with all the professors, their department, their patents, their articles, their conference proceedings, and the associated year for each of them. For example, Marshall bautz – physics. Article 1 (unveiling the merger…) is a 2020 article (continue for all articles). Conference proceeding 1 (reducing the athena…) is from 2020 (continue with all conference proceedings). Patent 1 is a patent granted in 2019 and has a priority date of 2017 https://patents.google.com/patent/US10206637 (continue with all patents). I also want it associated with the gender/race I associated with each professor in my last excel sheet.
* I then want to be able to “sort” by asking the inquiry “what is the ratio of articles (or conference proceedings) to patents for women vs men in 2017”
* I only care about the data on the website. Before then, there are so many variables contributing to why some people might get more or less patents. If everyone’s at the same institution, we’ve eliminated some variables.

**10/26 weekly meeting**

**\*We can use outside source to classify the gender as long as it’s a reliable source (with a paper or published article)**

Updates:

* Team 1: my team is looking at the data, and we want to separate the data into groups of older professors and newer professors, scale the data and take a look at them. We haven’t decided on which algorithms to use so far. We are concerned about the data regarding the lack of info since there might be data missing before the professors come to MIT.
* Team 2: No for now
* Team 3: collected names of 1312 professors, applied gender\_api for assigning gender to each professor using their names, gathered available patents (with USPTO numbers) from each professors
  + Aggregate the name of each gender of new born baby 1984- 2020

**10/19 meeting with client**

Prep:

* Question: limited data downloaded
* Maybe share the methodology and limit to a time frame of 5 years

Notes:

* Timeframe 2000 - current (past 20 years)
* Patent, articles, courses, grants, conference proceedings
* E.g. fewer grants from female professor
* The number in the [excel sheet](https://docs.google.com/spreadsheets/d/1bVff2IgrYXBfmuOSim-hhuOxVthvfuq3/edit?rtpof=true): population at school/ representative population of patent (don’t need to use this data)
* Patent (filing date)
* Gender classification

1. Use paper reference
2. Search online datas that includes gender information

**10/6 meeting with client**

1. Download the data

Different ways of categorizing the professor.

* Patent (amount)
* Article (amount)

1. Challenge

* The same time you apply the patent and article (5 year expansion)

1. Gender differences

* What, how much, why later

1. Match it to date, and match it to gender (file date is more relevant)

Current dataset: set of patterns downloaded

1. Algorithm classifying gender
2. Analysis on the ratio of article/patents