**NPD Group** is a market research company (one of the world’s largest) that gets data from point-of-sale systems, retailers, and manufacturers.  Think of all the big ones – Target, Nike, Walmart, et al.  They turn that around and do research, product development, and merchandise industry analysis.  It’s privately owned by Tod Johnson, the CMU trustee.

The project they want is related to receipts they gather.  They have 50k retailers, but because of the overall size of the incoming data load, unsupervised methods need to be deployed to flag potential data corruptions, missing data, and structural changes in the data that would otherwise be missed and would affect trend projections later on in processing. They want help with tasking and cataloging possible anomaly types and developing approaches that will surface ongoing issues for analyst review.   These could be parsing errors, domain changes, missing data, reconciliation issues.  They’ll provide 2 year of data loads from an ecommerce source and a log of known issues that should be flagged.  They’ll provide the subject matter expertise, and want an anomaly taxonomy so they can triage issues.  They’re expecting something shallow for the project – 5-10 issue types, 1-3 actions suggested for each issue type.  The assigned team will attempt to automate issue detection.  They are a python shop.

**LinkedIn** is struggling with the Educational Partnership Agreement – the document that outlines IP ownership of capstone work, and sets the Non-Disclosure and Data Privacy agreement.  They are working with CMU legal on that document.  If that document gets signed, they want to do a very specific project that doesn’t require sensitive of confidential data.

As part of its portfolio, LinkedIn supports customers in identifying approaches to maximize their connectivity with potential employees and products through analysis of traffic, posts, etc.  However, the large variation in the types of companies, products, customers, and tracked data metrics means that there is no one-size-fits-all statistical approach for tackling these challenges.  For the project they want, the student team would gather inputs on common customer requests, statistical approaches, and data problems and build a taxonomy of customer profile segments, related data profiles (eg - periodicity, type, category, volume, and breadth of data set available), known data set types (marketing, sales, product).  The taxonomy build would be used to help suggest viable, appropriate statistical methods per problem and data profile with validation checklists and document use cases for each method with examples based on provided anonymized data.  This information could be used in an interactive application, like a ShinyApp, where based on user-selected inputs, end-users could input their relevant profile variables for a set of recommended methods, checklists, and high-level procedural documentation.

**IFDA** is the International Food Distributors Association.  It’s a trade association.  They signed up for a project previously, but didn’t do it at the time, I don’t know why.  They’re meeting this week to figure out if they have a project they want us to do this semester.  It’s higher probability because they’ve already signed an EPA and just need to allocate a problem and contact, so it could be a fast turnaround if they have a challenge that fits.

IFDA wants to move forward (Mike Harding from Advancement met with them this morning), I’m meeting them Friday to go through a more defined scope.  They want more of an advisory project – students interviewing stakeholders.  From Advancement notes from their meeting:  *U.S. every day. In the US, the foodservice distribution industry is suffering from high turnover rates in driver and warehouse selector positions in the US. The IFDA membership would like to gain more insight into the specific causes and total costs of this industry challenge. This goal of this project is to explore modeling turnover rate using information about work hours, type of work, quality of life, recruitment/retention incentives, the presences of competing industries, training programs, supervisor tenure, among other factors.*

**A list of the possible data elements from members:**

* ∙            Length of service
* ∙            Wages
* ∙            Benefits (values)
* ∙            Location
* ∙            Average weekly hours worked
* ∙            Time Off
* ∙            Bonuses and Incentives
* ∙            Drivers-Time away from home
* ∙            Reason for leaving
* ∙            Reasons people stay
* ∙            Typical work shifts
* ∙            Overlaying relevant supervisor data (TBD) to explore possible correlations
* ∙            Investigate the competitive factors in the various market regions (either from distributors or from other data sources).

It would entail stakeholder engagement and interviews upfront and define a meaningful analysis based on what’s available, as well as what else they should gather going forward.  Then a basic proof of concept or model if possible from what’s available.  I will confirm all of this Friday.

**DorothyAI** is a start-up that automates the search for patent documents.  I know little else about it, but put it at low probability because they haven’t responded to us in the last two weeks, and it’s a very small operation with under 10 people.

I will keep working to get a minimum of 7, ideally in the next week.  I would like to get the students assigned prior to the start of the semester.  Everyone would still hear all the project overviews in the first week to two weeks of the semester, but would like to get them pre-reading their own projects and get started as early as possible.  In any case, I’ll give the students a status update on Friday.  I should know about IFDA (more likely) and will call DorothyAI (less likely) by then.

Chris told me **Brian Macdonald** would be the faculty advisor on the NHL and NBA projects.  I have time set up with him today to go through those.

Unrelated – I had a call with **Ryan Harty** last week and a follow-up this morning.  He asked if I talked to you to let you know he’s trying to get in touch and set up time to talk to you.

Thanks,

Jamie McGovern

Special Faculty, Director, Master of Statistical Practice Program

Department of Statistics & Data Science

Carnegie Mellon University

**From:** [Brian Junker](mailto:bj20@andrew.cmu.edu)  
**Sent:** Monday, January 11, 2021 5:37 PM  
**To:** [Jamie McGovern](mailto:jspm@cmu.edu)  
**Subject:** Re: MSP projects?

Thanks Jamie,  
  
Here's what I infer so far (please update as appropriate; some questions also in parentheses here, and also at the end):  
  
Definite:  
  
1. NPD (what's that?)  
2. NHL (Sam)  
3.  NBA with Kostas Pelechrinis (Pitt)  
4.  COVID with Seema Lakdawala (Pitt)  
  
Maybe:   
  
5. LinkedIn project through Mike Harding (what's "EPA" and what's the hold up?)  
6. IFDA (what's that and what's its status?)  
7. DorothyAI  (what is that and what's the current status?)  
  
Being developed by BJ:  
  
8. PPS - I'm talking to them tomorrow about a specific project  
9. PIER/HCII - conversation started  
10. biostat project (through a contact of Joel's)  
  
With sufficient faculty support, it's better to have more projects than fewer (teams of 2-3 students easier to manage and grade than teams of 5-6).  
  
According to the hub, the instructors for 726 are:  
  
BJ  
Valerie  
Zach  
  
...but I guess Brian Macdonald is also on this course (that would be great, of course)?  Is Brian also contributing a project (that would also be great!)  
  
lmk all of the above.  
  
thanks  
  
-BJ

On 1/5/2021 3:48 PM, Jamie McGovern wrote:

Hi Brian –

Happy New Year.  So far, we have NPD, a project on the NHL with Sam, and a project on the NBA with Kostas Pelechrinis (Pitt), a project on COVID with Seema Lakdawala (Pitt).

We’re trying to close a LinkedIn project through Mike Harding, we have it at 75% odds – the only outstanding issues are around the EPA.  We are still working getting info back from IFDA – we’re trying to push that to the end in the next two weeks.  I’m not sure, could go either way.  DorothyAI is in the same boat – hope to get more information by end of this week.  Almost everyone was out last week, so just trying to get some responses now that most are back in the office.

So that’s 4 locked.  3 iffy.  Maybe all three close, maybe 0 of 3.  I’d think just based on everything to date, not all 3 of those close – so will still have some gaps.  If all 3 closed and we ended up with more than 7 because we were mitigating risk, that wouldn’t be necessarily bad.

Given that - if there are any in your conversations that would be good fits that would be great and relieve some of the pressure; we can do them – I’m going to keep working as if I have to fill out the slate to get at minimum 7, and will keep you updated in the next week as we have movement.

Thanks,

Jamie McGovern

Special Faculty, Director, Master of Statistical Practice Program

Department of Statistics & Data Science

Carnegie Mellon University

**From:** [Brian Junker](mailto:bj20@andrew.cmu.edu)  
**Sent:** Monday, January 4, 2021 10:01 AM  
**To:** [Jamie McGovern](mailto:jspm@cmu.edu)  
**Subject:** MSP projects?

Hi Jamie,  
  
Hope you had a nice holiday break!  
  
Where do we stand so far with projects?  
  
I made some initial inquiries with PPS and PIER before xmas, and I'd be   
happy to follow up with those if needed.  
  
Also Joel may have an interesting biostat project.  
  
What other stuff do we have in the bag or in the works?  
  
all best,  
  
-BJ  
  
--   
Brian Junker                    (412) 268 - 2718  
Department of Statistics        [brian@stat.cmu.edu](mailto:brian@stat.cmu.edu)  
232 Baker Hall                  FAX: (412) CMU-STAT  
Carnegie Mellon University        or (412) 268-7828  
Pittsburgh PA 15213 USA  
  
WWW: <http://www.stat.cmu.edu/~brian/>

--

Brian Junker                    (412) 268 - 2718

Department of Statistics        [brian@stat.cmu.edu](mailto:brian@stat.cmu.edu)

232 Baker Hall                  FAX: (412) CMU-STAT

Carnegie Mellon University        or (412) 268-7828

Pittsburgh PA 15213 USA

WWW: <http://www.stat.cmu.edu/~brian/>

Thanks Jamie,  
  
Here's what I infer so far (please update as appropriate; some questions also in parentheses here, and also at the end):  
  
Definite:  
  
1. NPD (what's that?)  
2. NHL (Sam)  
3.  NBA with Kostas Pelechrinis (Pitt)  
4.  COVID with Seema Lakdawala (Pitt)  
  
Maybe:   
  
5. LinkedIn project through Mike Harding (what's "EPA" and what's the hold up?)  
6. IFDA (what's that and what's its status?)  
7. DorothyAI  (what is that and what's the current status?)  
  
Being developed by BJ:  
  
8. PPS - I'm talking to them tomorrow about a specific project  
9. PIER/HCII - conversation started  
10. biostat project (through a contact of Joel's)  
  
With sufficient faculty support, it's better to have more projects than fewer (teams of 2-3 students easier to manage and grade than teams of 5-6).  
  
According to the hub, the instructors for 726 are:  
  
BJ  
Valerie  
Zach  
  
...but I guess Brian Macdonald is also on this course (that would be great, of course)?  Is Brian also contributing a project (that would also be great!)  
  
lmk all of the above.  
  
thanks  
  
-BJ

On 1/5/2021 3:48 PM, Jamie McGovern wrote:

Hi Brian –

Happy New Year.  So far, we have NPD, a project on the NHL with Sam, and a project on the NBA with Kostas Pelechrinis (Pitt), a project on COVID with Seema Lakdawala (Pitt).

We’re trying to close a LinkedIn project through Mike Harding, we have it at 75% odds – the only outstanding issues are around the EPA.  We are still working getting info back from IFDA – we’re trying to push that to the end in the next two weeks.  I’m not sure, could go either way.  DorothyAI is in the same boat – hope to get more information by end of this week.  Almost everyone was out last week, so just trying to get some responses now that most are back in the office.

So that’s 4 locked.  3 iffy.  Maybe all three close, maybe 0 of 3.  I’d think just based on everything to date, not all 3 of those close – so will still have some gaps.  If all 3 closed and we ended up with more than 7 because we were mitigating risk, that wouldn’t be necessarily bad.

Given that - if there are any in your conversations that would be good fits that would be great and relieve some of the pressure; we can do them – I’m going to keep working as if I have to fill out the slate to get at minimum 7, and will keep you updated in the next week as we have movement.

Thanks,

Jamie McGovern

Special Faculty, Director, Master of Statistical Practice Program

Department of Statistics & Data Science

Carnegie Mellon University

**From:** [Brian Junker](mailto:bj20@andrew.cmu.edu)  
**Sent:** Monday, January 4, 2021 10:01 AM  
**To:** [Jamie McGovern](mailto:jspm@cmu.edu)  
**Subject:** MSP projects?

Hi Jamie,  
  
Hope you had a nice holiday break!  
  
Where do we stand so far with projects?  
  
I made some initial inquiries with PPS and PIER before xmas, and I'd be   
happy to follow up with those if needed.  
  
Also Joel may have an interesting biostat project.  
  
What other stuff do we have in the bag or in the works?  
  
all best,  
  
-BJ  
  
--   
Brian Junker                    (412) 268 - 2718  
Department of Statistics        [brian@stat.cmu.edu](mailto:brian@stat.cmu.edu)  
232 Baker Hall                  FAX: (412) CMU-STAT  
Carnegie Mellon University        or (412) 268-7828  
Pittsburgh PA 15213 USA  
  
WWW: <http://www.stat.cmu.edu/~brian/>

--

Brian Junker (412) 268 - 2718

Department of Statistics [brian@stat.cmu.edu](mailto:brian@stat.cmu.edu)

232 Baker Hall FAX: (412) CMU-STAT

Carnegie Mellon University or (412) 268-7828

Pittsburgh PA 15213 USA

WWW: <http://www.stat.cmu.edu/~brian/>