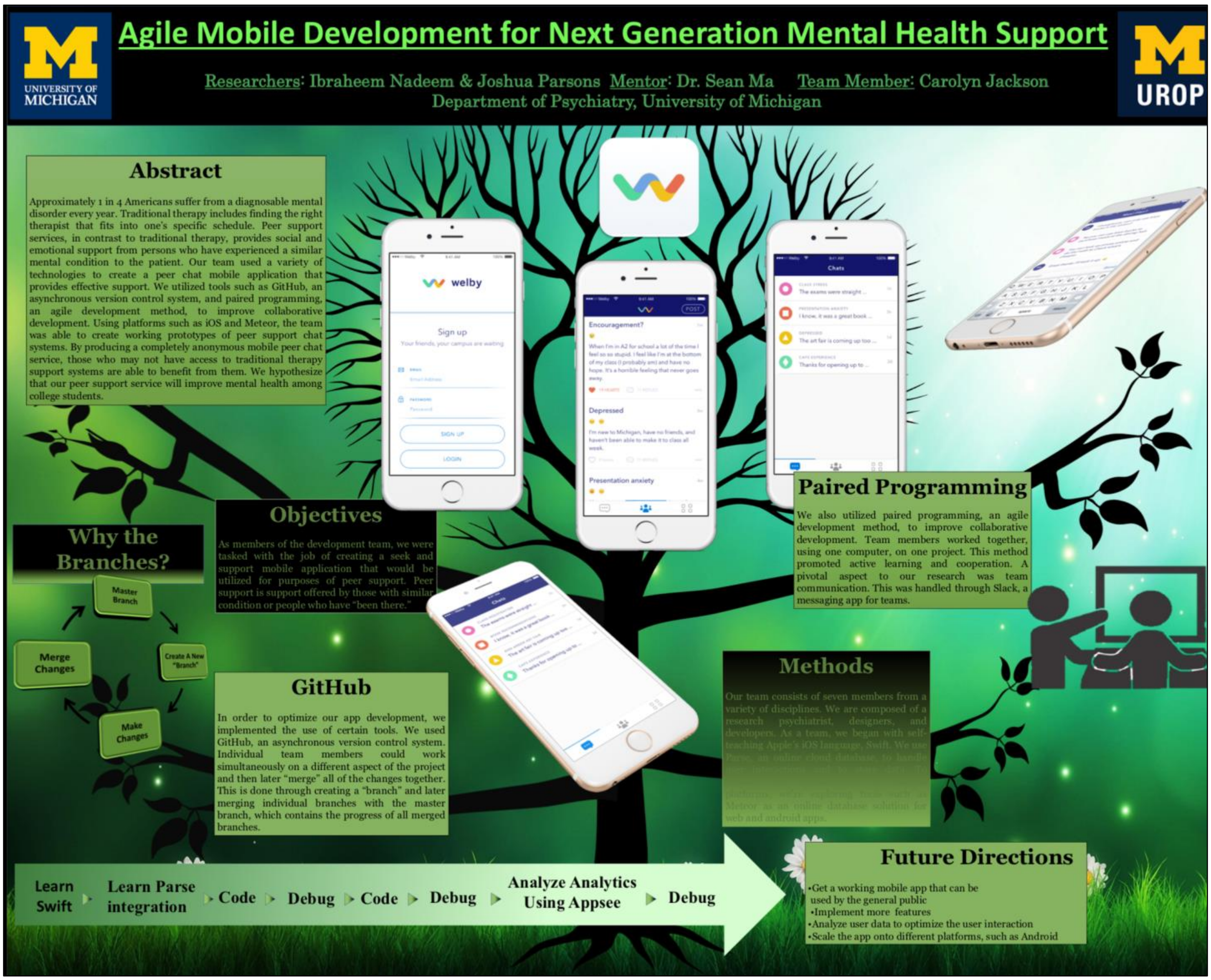
Research Portfolio (Selected)

Joshua J. Parsons
B.A. Economics ‘18
University of Michigan

Research Poster: Welby LLC

This poster showcases my contributions to Welby, a startup dedicated to supporting students' mental health. I presented this poster at the University of Michigan Undergraduate Research Opportunity Program (UROP) Summer Symposium 2015.



Python Chat Integration: Welby LLC

I developed this Python script to integrate the Parse platform into our Welby application. This integration allowed our team to implement a chat feature using Parse as a backend. This third-party API integration made the complex task of integrating a server-side chat feature simple.

```python
# declaring application ID and rest api key
APPLICATION_ID = "LTkPGrlxgflhb5vN6g2bSgUguo2ESeSM6x9BHuMx"
REST_API_KEY = "2gDCyfbrdNOjvvkDT7snuIAXqR8KzafR45DyL6nx"

# importing necessary modules, register to connect to parse, and Object for parseobjects
from parse_rest.connection import register
from parse_rest.datatypes import Object

# registering the company parse account
register(APPLICATION_ID, REST_API_KEY)

# creating class Messages in order to access Messages class in  Parse account
myClassName = "Messages"
myClass = Object.factory(myClassName)

# sendMessage = MessageClass(message="from sean", messageSessionID="qJkP80Lpt5", userType=False)
# sendMessage.save()

# querying all data in Parse class Messages
message_query = myClass.Query.all()

# printing first message under message header
print message_query[0].message

for m in message_query:
    print m.message

def sendMessage(msgstring):
    sendMessage = myClass(message=msgstring, messageSessionID="qJkP8OLpt5", userType=False)
    sendMessage.save()

def querymsg():
    message_query = myClass.Query.all()
    print message_query[len(message_query) - 1].message
```

Stata Analysis

In this Stata project, I incorporated non-linearities in a linear regression model by using different functional form specification. All Stata output was calculated using two separate datasets. In Part I, the dataset provides worker wage data, years of education, and years of experience. Part II uses a dataset that contains the number of car accidents in California by month. I used this data to study the relationship between unemployment, California motor vehicle laws, and car accidents.

Part I

```
. reg wage educ exper expersq

      Source |       SS           df       MS      Number of obs   =     2,061
-------------+----------------------------------   F(3, 2057)      =    139.75
       Model |  24286739.1         3  8095579.68   Prob > F        =    0.0000
    Residual |   119161332     2,057  57929.6705   R-squared       =    0.1693
-------------+----------------------------------   Adj R-squared   =    0.1681
       Total |   143448071     2,060  69634.9861   Root MSE        =    240.69

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   56.39506   2.925693    19.28   0.000     50.65743    62.13269
       exper |   62.87801   6.216213    10.12   0.000     50.68728    75.06874
     expersq |  -1.867866   .3209525    -5.82   0.000    -2.497292    -1.23844
       _cons |  -542.7231   58.24455    -9.32   0.000    -656.9475   -428.4986
------------------------------------------------------------------------------
```

- **The effect of the 1st year of working experience on log-wage**: The first year of working experience is predicted to increase someone's wage by $62.88.

- **The log-wage increase from the 11th year of working experience** is $25.48, which is less than the $62.88 wage increase from the first year of working experience.
  - ✓ 2(-1.87)(10) = 37.4
  - ✓ 62.88 – 37.4 = 25.48

- **The coefficient for exper is statistically significant** because its p-value of 0 is less than .05.

```
. test exper expersq

 ( 1)  exper = 0
 ( 2)  expersq = 0

       F(  2,  2057) =  144.06
            Prob > F =    0.0000
```

- **Since the p-value of this F-test is 0,** we reject the null at any significance level.

```
. gen count=(exper>17.6)

. tab count
```

count	Freq.	Percent	Cum.
0	2,051	99.51	99.51
1	10	0.49	100.00
Total	2,061	100.00	

- **With a turning point of 17.6 years of experience,** there are 10 people with more than 17.6 years of experience.
  - ✓ 62.878 / -3.6 = 17.467 ~ 17.6

```
. reg wage educ exper interaction
```

Source	SS	df	MS			
				Number of obs	=	2,061
				F(3, 2057)	=	140.47
Model	24391081.8	3	8130360.61	Prob > F	=	0.0000
Residual	119056989	2,057	57878.9448	R-squared	=	0.1700
				Adj R-squared	=	0.1688
Total	143448071	2,060	69634.9861	Root MSE	=	240.58

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	22.89312	5.999083	3.82	0.000	11.12821	34.65803
exper	-25.10526	9.093929	-2.76	0.006	-42.93953	-7.270992
interaction	3.926066	.6570692	5.98	0.000	2.637476	5.214657
_cons	65.74801	90.29255	0.73	0.467	-111.3263	242.8223

- Take the derivative of the conditional mean log-wage, given by: B1 +B2*educ + B3*exper + Alpha1*exper*educ, which equals B3 + Alpha1*educ
- H_o: Alpha1=0
  - ✓ Given the null, I predict that the alternative is: Alpha1 =/= 0
  - ✓ **At the 1% significance level, we fail to reject the null that return to experience does not depend on level of educ,** because the p-val for interaction is 0, which is less than .01.

```
. reg wage educ exper expersq IQ

      Source |       SS           df       MS      Number of obs   =     2,061
-------------+----------------------------------   F(4, 2056)      =    114.93
       Model |  26213134.9         4  6553283.72   Prob > F        =    0.0000
    Residual |   117234936     2,056  57020.8835   R-squared       =    0.1827
-------------+----------------------------------   Adj R-squared   =    0.1811
       Total |   143448071     2,060  69634.9861   Root MSE        =    238.79

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   49.22841   3.153675    15.61   0.000     43.04368    55.41315
       exper |   64.92916   6.177349    10.51   0.000     52.81465    77.04367
     expersq |  -1.936303   .3186427    -6.08   0.000    -2.561199   -1.311407
          IQ |   2.313721   .3980662     5.81   0.000     1.533066    3.094376
       _cons |  -691.4854   63.20014   -10.94   0.000    -815.4283   -567.5424
------------------------------------------------------------------------------
```

- The OLS estimate of the slope is 49.23, which means **that a one-year increase in education gives a $49.23 increase in log-mean wage**. It decreased from part one because this estimate includes IQ, and cov(educ, IQ) =/= 0.

Part II

- When **beltlaw** changed from 0 to 1 in the dataset is when California introduced the Belt Law.
  - ✓ Jan 1986 is when the Belt Law was introduced.

. reg totacc year feb mar apr may jun jul aug sep oct nov dec

Source	SS	df	MS
Model	1.8088e+09	12	150732516
Residual	463535721	95	4879323.38
Total	2.2723e+09	107	21236690.8

Number of obs	=	108
F(12, 95)	=	30.89
Prob > F	=	0.0000
R-squared	=	0.7960
Adj R-squared	=	0.7702
Root MSE	=	2208.9

totacc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
year	1400.111	82.32155	17.01	0.000	1236.682	1563.54
feb	-1525.222	1041.294	-1.46	0.146	-3592.453	542.0086
mar	3589.333	1041.294	3.45	0.001	1522.103	5656.564
apr	1125.889	1041.294	1.08	0.282	-941.3419	3193.12
may	1835.556	1041.294	1.76	0.081	-231.6753	3902.786
jun	1465.556	1041.294	1.41	0.163	-601.6753	3532.786
jul	2336.444	1041.294	2.24	0.027	269.2136	4403.675
aug	3142.556	1041.294	3.02	0.003	1075.325	5209.786
sep	2750.222	1041.294	2.64	0.010	682.9914	4817.453
oct	4622.667	1041.294	4.44	0.000	2555.436	6689.897
nov	4215.111	1041.294	4.05	0.000	2147.88	6282.342
dec	5538.333	1041.294	5.32	0.000	3471.103	7605.564
_cons	-2738814	163409.9	-16.76	0.000	-3063224	-2414404

- **There is seasonality in accidents;** if there was no seasonality in accidents, the coefficients would be zero.

. reg totacc unem year feb mar apr may jun jul aug sep oct nov dec

Source	SS	df	MS
Model	1.9280e+09	13	148308444
Residual	344316137	94	3662937.63
Total	2.2723e+09	107	21236690.8

Number of obs	=	108
F(13, 94)	=	40.49
Prob > F	=	0.0000
R-squared	=	0.8485
Adj R-squared	=	0.8275
Root MSE	=	1913.9

totacc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
unem	-994.0684	174.2438	-5.71	0.000	-1340.033	-648.1033
year	870.2175	117.1085	7.43	0.000	637.6957	1102.739
feb	-1447.906	902.3136	-1.60	0.112	-3239.47	343.6591
mar	3213.796	904.6099	3.55	0.001	1417.672	5009.921
apr	396.9054	911.2154	0.44	0.664	-1412.334	2206.145
may	819.3968	919.6257	0.89	0.375	-1006.542	2645.335
jun	725.527	911.4889	0.80	0.428	-1084.256	2535.31
jul	2137.631	902.8846	2.37	0.020	344.9323	3930.329
aug	2380.437	912.048	2.61	0.011	569.5438	4191.329
sep	1700.928	920.7682	1.85	0.068	-127.279	3529.135
oct	3595.463	920.0026	3.91	0.000	1768.776	5422.15
nov	3287.314	916.7518	3.59	0.001	1467.082	5107.546
dec	4477.994	921.1569	4.86	0.000	2649.015	6306.973
_cons	-1679166	233548.4	-7.19	0.000	-2142882	-1215450

- Lower unemployment means that more people must drive to work, and the lower unemployment also means higher economic activity. Therefore, the negative coefficient on unem makes sense. **A 1% increase in unemployment rate reduces the number of total accidents by about 994, controlling for seasonality. This is statistically significant at the 1% level because the p value of zero is less than .01.**

. reg totacc feb mar apr may jun jul aug sep oct nov dec beltlaw spdlaw unem

Source	SS	df	MS			
Model	2.0300e+09	14	144999088	Number of obs	=	108
Residual	242338686	93	2605792.32	F(14, 93)	=	55.64
Total	2.2723e+09	107	21236690.8	Prob > F	=	0.0000
				R-squared	=	0.8934
				Adj R-squared	=	0.8773
				Root MSE	=	1614.2

totacc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
feb	-1446.774	761.0514	-1.90	0.060	-2958.072	64.52302
mar	3208.301	763.0423	4.20	0.000	1693.05	4723.552
apr	386.2376	768.7683	0.50	0.617	-1140.384	1912.859
may	943.6156	774.2991	1.22	0.226	-593.9889	2481.22
jun	853.7867	768.3369	1.11	0.269	-671.9782	2379.551
jul	2273.811	763.0553	2.98	0.004	758.534	3789.087
aug	2508.373	768.7336	3.26	0.002	981.8204	4034.926
sep	1824.662	775.16	2.35	0.021	285.3477	3363.976
oct	3719.52	774.5826	4.80	0.000	2181.352	5257.688
nov	3412.826	772.1549	4.42	0.000	1879.479	4946.172
dec	4601.566	775.4538	5.93	0.000	3061.669	6141.464
beltlaw	5661.487	534.7462	10.59	0.000	4599.587	6723.387
spdlaw	-1251.802	534.1831	-2.34	0.021	-2312.584	-191.0205
unem	-1008.615	149.0033	-6.77	0.000	-1304.506	-712.7244
_cons	46091.75	1446.066	31.87	0.000	43220.15	48963.35

- The coefficient on spdlaw is negative, which means **a higher speed limit decreased traffic accidents**.
  - ✓ People may have become more cautious after the increase in speed limit.

. reg totacc feb mar apr may jun jul aug sep oct nov dec beltlaw spdlaw unem year

Source	SS	df	MS			
Model	2.0644e+09	15	137624059	Number of obs	=	108
Residual	207965033	92	2260489.49	F(15, 92)	=	60.88
Total	2.2723e+09	107	21236690.8	Prob > F	=	0.0000
				R-squared	=	0.9085
				Adj R-squared	=	0.8936
				Root MSE	=	1503.5

totacc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
feb	-1459.42	708.8427	-2.06	0.042	-2867.243	-51.59754
mar	3269.725	710.8641	4.60	0.000	1857.887	4681.562
apr	505.4724	716.6753	0.71	0.482	-917.9066	1928.851
may	1197.747	724.1126	1.65	0.102	-240.4036	2635.897
jun	1062.753	717.6245	1.48	0.142	-362.5114	2488.017
jul	2394.254	711.3725	3.37	0.001	981.4067	3807.101
aug	2720.952	718.0627	3.79	0.000	1294.818	4147.087
sep	2084.213	725.0375	2.87	0.005	644.2255	3524.2
oct	3975.458	724.4175	5.49	0.000	2536.702	5414.213
nov	3652.504	721.7986	5.06	0.000	2218.95	5086.059
dec	4862.924	725.3527	6.70	0.000	3422.311	6303.537
beltlaw	4308.809	606.9502	7.10	0.000	3103.353	5514.264
spdlaw	-2043.126	537.3253	-3.80	0.000	-3110.301	-975.9515
unem	-846.0225	144.9084	-5.84	0.000	-1133.823	-558.222
year	514.7668	132.0076	3.90	0.000	252.5883	776.9454
_cons	-976220.7	262166.8	-3.72	0.000	-1496907	-455534.7

- When controlling for year, the coefficient on spdlaw is –2043, which is less than the coefficient when not controlling for year. Since spdlaw has a p-value of 0, which is less than .05, we can conclude that it is statistically significant at the 5% level.

- **The OLS estimate of spdlaw decreased once we added year to the regression because of omitted variable bias.** This is because we didn't consider the time trend in total accidents. The first regression of the OLS estimate of the coefficient on the spdlaw was biased up. The bias was positive because corr(year, totacc) > 0: there was a positive time trend in total accidents; and corr(year, spdlaw) > 0.

```
. reg prcfat feb mar apr may jun jul aug sep oct nov dec beltlaw spdlaw unem

      Source |       SS           df       MS      Number of obs   =       108
-------------+----------------------------------   F(14, 93)       =     11.30
       Model |  .670841928        14  .047917281   Prob > F        =    0.0000
    Residual |  .394406228        93  .004240927   R-squared       =    0.6298
-------------+----------------------------------   Adj R-squared   =    0.5740
       Total |  1.06524816       107  .00995559    Root MSE        =    .06512

------------------------------------------------------------------------------
      prcfat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         feb |  -.0030758   .0307026    -0.10   0.920     -.064045    .0578934
         mar |  -.0015291   .0307829    -0.05   0.960    -.0626578    .0595996
         apr |   .0572332   .0310139     1.85   0.068    -.0043542    .1188206
         may |   .0757831    .031237     2.43   0.017     .0137526    .1378136
         jun |   .1003331   .0309965     3.24   0.002     .0387802    .1618859
         jul |   .1693277   .0307834     5.50   0.000      .108198    .2304574
         aug |   .1878873   .0310125     6.06   0.000     .1263026    .2494719
         sep |   .1550841   .0312717     4.96   0.000     .0929847    .2171836
         oct |   .0940974   .0312484     3.01   0.003     .0320442    .1561506
         nov |    .003512   .0311505     0.11   0.910    -.0583467    .0653707
         dec |  -.0020653   .0312836    -0.07   0.948    -.0641882    .0600577
     beltlaw |  -.1000755   .0215729    -4.64   0.000    -.1429149    -.057236
      spdlaw |   .0259248   .0215502     1.20   0.232    -.0168696    .0687192
        unem |  -.0069638   .0060111    -1.16   0.250    -.0189007    .0049731
       _cons |   .9028632   .0583376    15.48   0.000     .7870162     1.01871
------------------------------------------------------------------------------
```

- The slope on spdlaw is .026, which implies that accidents increased by .026 on average once the speed law was introduced, accounting for seasonal influence and the unemployment rate. This is not statistically significant at a significance level of 1% because .232 > 0.

```
. reg prcfat feb mar apr may jun jul aug sep oct nov dec beltlaw spdlaw unem year

      Source |       SS           df       MS      Number of obs   =       108
-------------+----------------------------------   F(15, 92)       =     15.57
       Model |  .764194266        15  .050946284   Prob > F        =    0.0000
    Residual |   .30105389        92  .003272325   R-squared       =    0.7174
-------------+----------------------------------   Adj R-squared   =    0.6713
       Total |  1.06524816       107  .00995559    Root MSE        =     .0572

------------------------------------------------------------------------------
      prcfat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         feb |  -.0024167   .0269697    -0.09   0.929     -.055981    .0511475
         mar |  -.0047301   .0270467    -0.17   0.862    -.0584471    .0489869
         apr |   .0510194   .0272678     1.87   0.065    -.0031367    .1051756
         may |   .0625394   .0275507     2.27   0.026     .0078213    .1172576
         jun |   .0894431   .0273039     3.28   0.001     .0352153     .143671
         jul |    .163051    .027066     6.02   0.000     .1092956    .2168064
         aug |    .176809   .0273205     6.47   0.000     .1225481      .23107
         sep |   .1415581   .0275859     5.13   0.000       .08677    .1963461
         oct |   .0807596   .0275623     2.93   0.004     .0260184    .1355008
         nov |  -.0089784   .0274627    -0.33   0.744    -.0635217    .0455649
         dec |  -.0156855   .0275979    -0.57   0.571    -.0704973    .0391263
     beltlaw |  -.0295827    .023093    -1.28   0.203    -.0754474    .0162819
      spdlaw |   .0671634   .0204439     3.29   0.001       .02656    .1077668
        unem |  -.0154371   .0055134    -2.80   0.006    -.0263872    -.004487
        year |  -.0268263   .0050226    -5.34   0.000    -.0368015    -.016851
       _cons |   54.17908   9.974812     5.43   0.000     34.36824    73.98992
------------------------------------------------------------------------------
```

- When we add variable *year* to the regression, the slope on spdlaw increases
  - ✓ cov(spdlaw, year) =/=0.
- After adding variable *year*, the higher speed limit does not have a statistically significant effect on the percent of fatal accidents at a 1% significance level

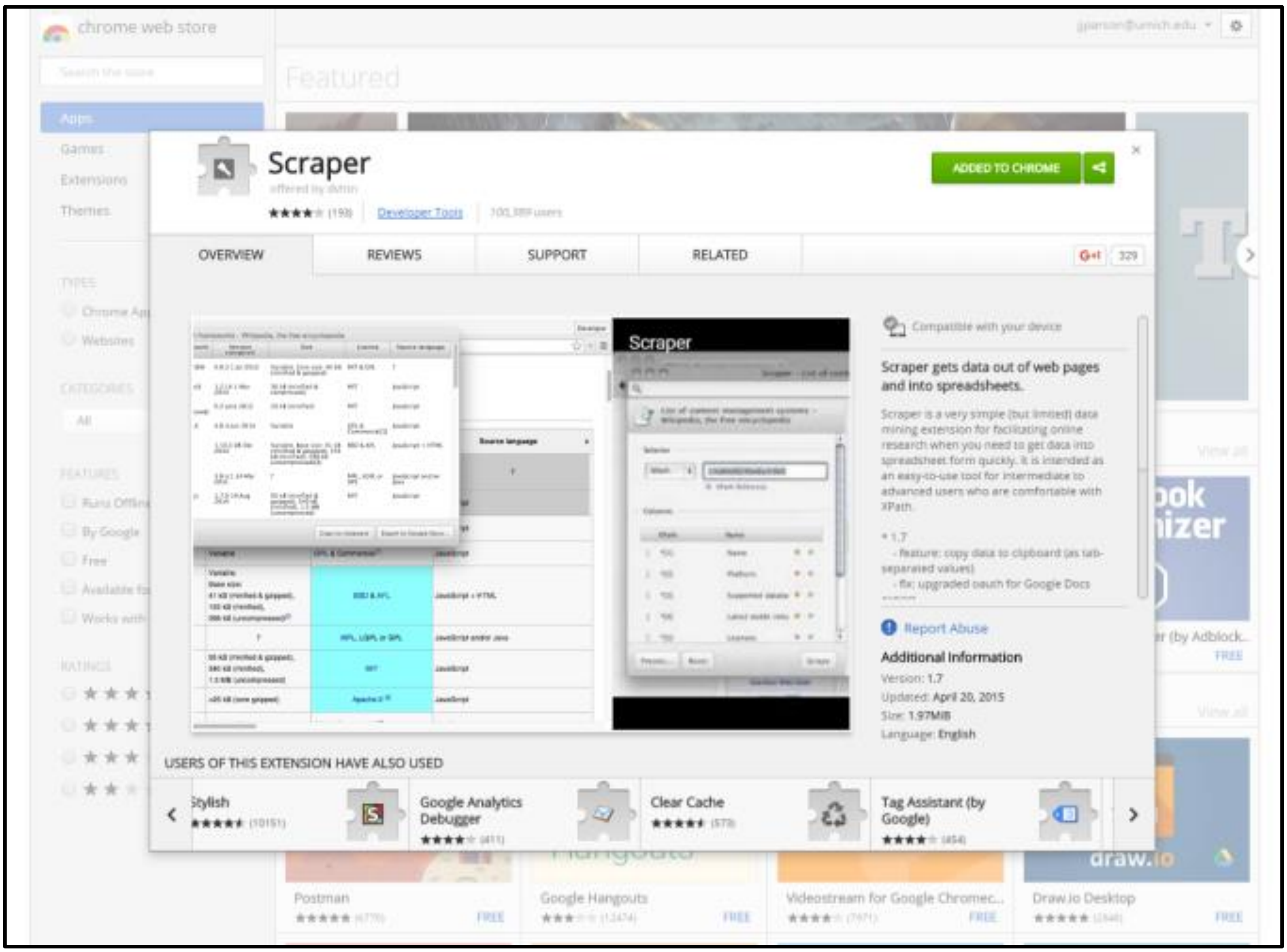
Web Scraper Demonstration

This is a tutorial I created for the office during my time at SVN Stewart Commercial Group to improve office efficiency by automating data collection and organization. My post was featured on the University of Michigan College of Literature, Science, and the Arts (LSA) Summer Vocation 2016 blog.

Web Scraping for Collecting Data and Acquiring Contacts

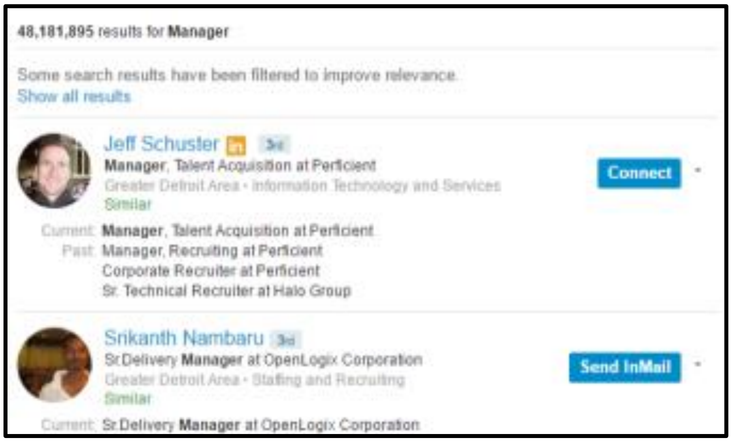
Web scraping is a software method for obtaining data from web pages. The method outlined in this tutorial shows a program that allows a user to extract data that is in the form of a list, such as a search result or table.

To download, search "Chrome scraper" in Google and open the first search result (it should be a link to the Chrome webstore.) Now, you should see a screen similar to this:

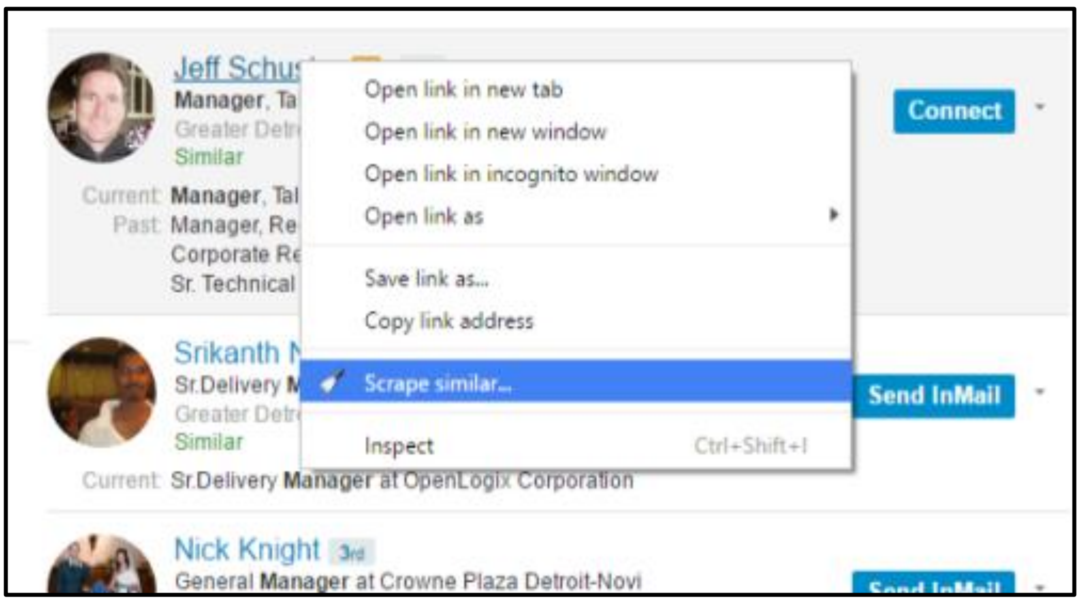


In the top right corner of the Scraper window, there is a button that says "Add to Chrome." In my screen above, this button says "Added to Chrome," since I have already added this extension.

Now that we added the extension to Chrome, we can now start scraping data. For example, if I wanted to scrape contacts from a LinkedIn Search for "Manager":

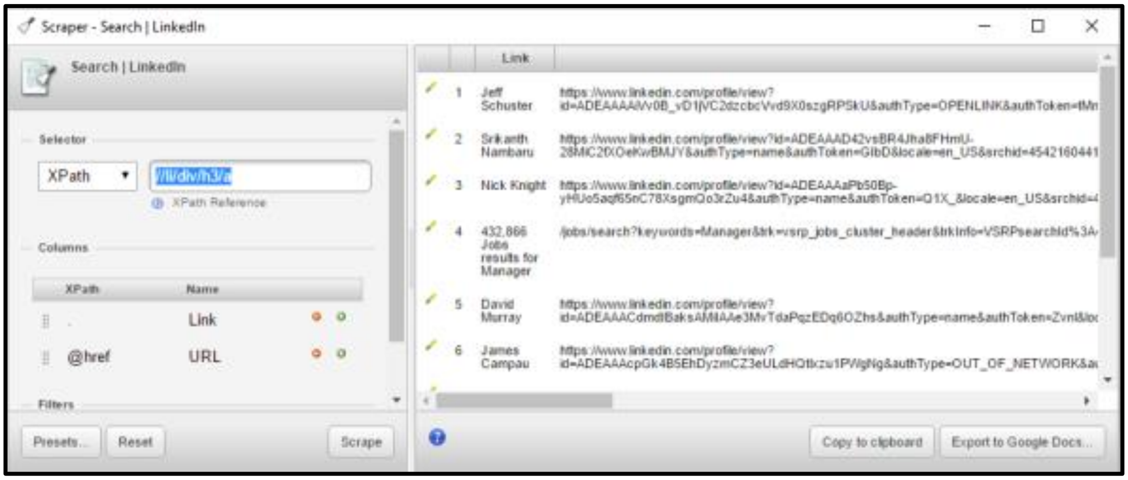


I would highlight then right click one of the names on the screen, which gives us this menu:

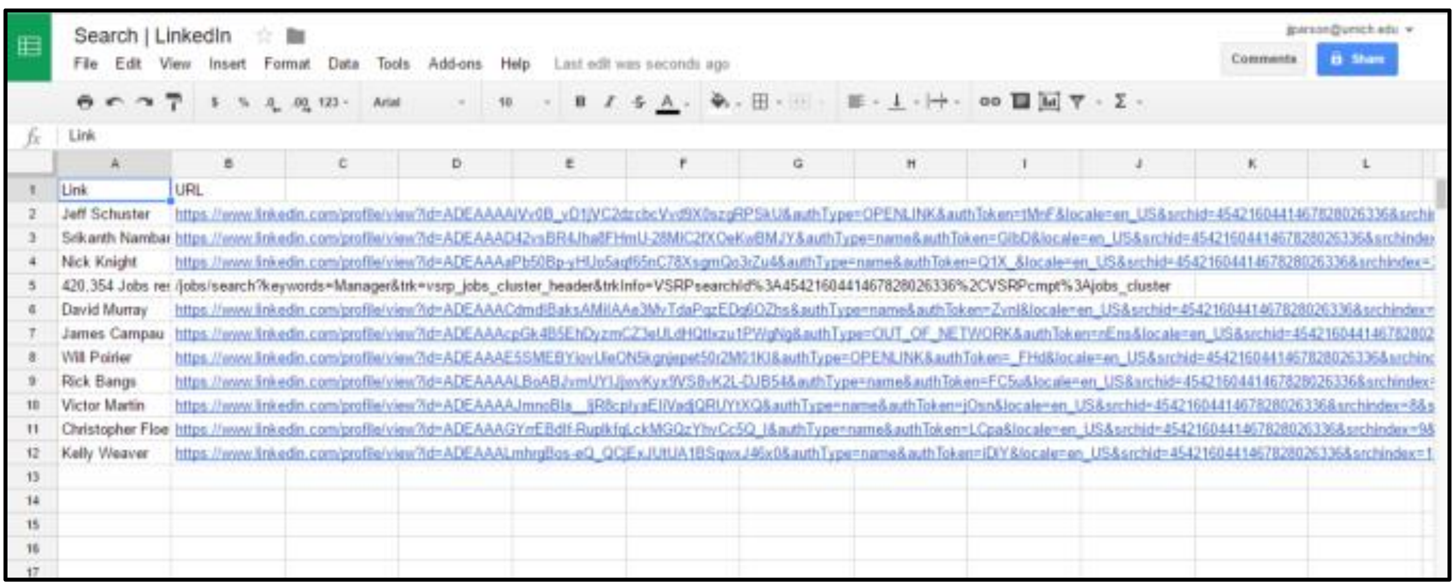


As you see, we have a new option under the right-click menu that says "Scrape similar…." This becomes an option after we downloaded the Scraper extension.

Click on Scrape similar. Now we have this menu that has the names we highlighted and the link to that person's profile. The reason why we have names and links in this data set is because we're scraping search results that are hyperlinks.



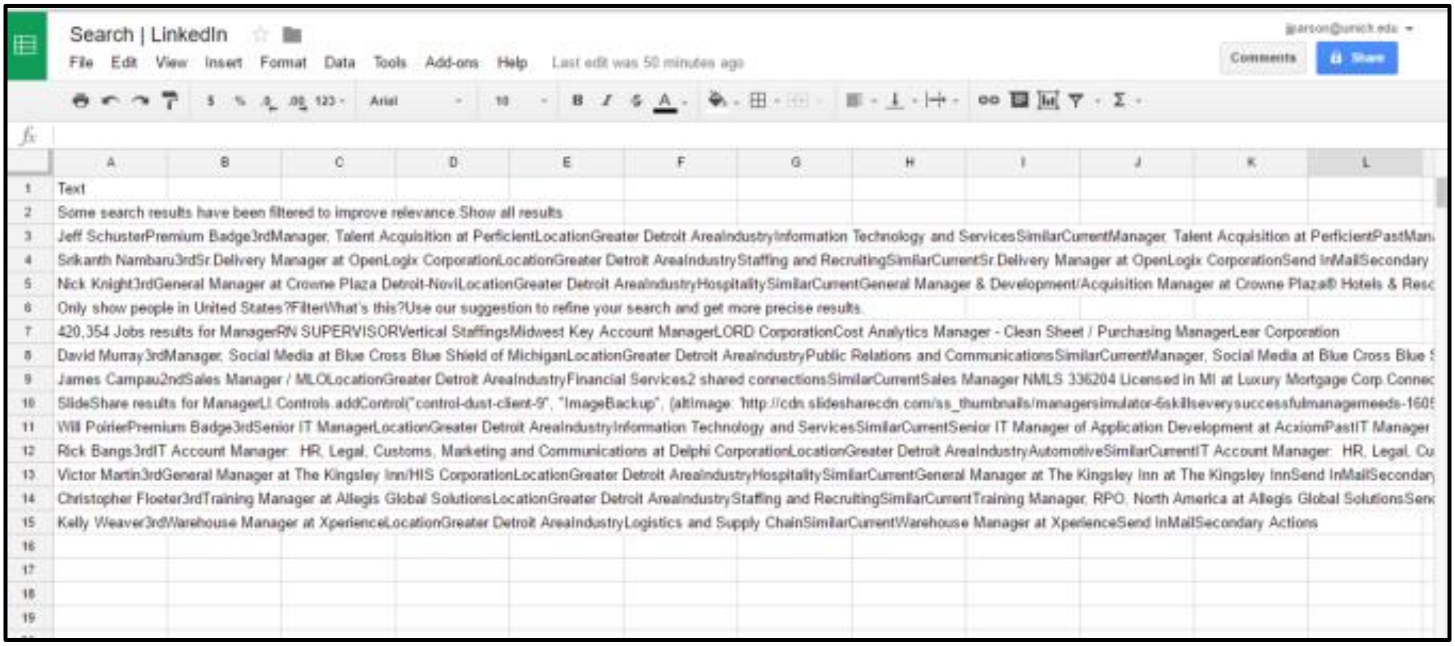
Now that we have the data we want, we can export this information to Google Docs by clicking the "Export to Google Docs" button in the bottom right hand corner of the Scraper window, which will open a window with a populated Google Document like this:



As you can see, this pulled all of the names and links to profiles on the page. It also pulled an advertisement from the page; we can remove this ad by right clicking the row and selecting "Delete Row."

We can pull results from multiple pages by going to the pages that we want to scrape and repeating this process. When we export each page to a Google Document, it will appear on a separate document. To put the information across multiple documents on to a single document, copy and paste the results from subsequent documents on to the first document.

We also have some control over what we can scrape from the web page. For example, if we highlight a name with the current and previous job titles and select "Scrape similar," our Google document will look like this:



However, we have little control over how the program displays data: we can remedy this through manual manipulation of the data or exporting the Google Document to Excel and modifying the display settings.