

# Process Manual

**From ML to M&A:  
Ten M&A Target Predictions  
through a Machine Learning Model**

December 16, 2019



**WISCONSIN**  
SCHOOL OF BUSINESS

UNIVERSITY OF WISCONSIN-MADISON

TOGETHER FORWARD®

NICHOLAS CENTER *for*  
CORPORATE FINANCE &  
INVESTMENT BANKING

# Contents

Project Summary and Overview .....	4
Purpose.....	4
Problem Specificity .....	4
What is Machine Learning and Deep Learning? .....	4
Why Machine Learning? .....	4
Building a Machine Learning model.....	5
How is this manual structured?.....	5
Variable Selection.....	6
Choosing Variables and Categorizing .....	6
Comprehensive Variable List .....	7
Variables by Data Source .....	8
Exclusions and Limitations.....	8
Bias .....	8
Availability.....	9
Time Variance.....	9
Specificity .....	9
Time Series .....	9
Data Collection .....	10
Bloomberg Terminal and Bloomberg API.....	10
Variables Pulled in Bloomberg.....	10
Bloomberg API Tutorial .....	10
Capital IQ .....	11
Variables Pulled in Capital IQ.....	11
Capital IQ Tutorial.....	11
Yahoo! Finance API .....	12
Variables Pulled from Yahoo! Finance API.....	12
Compustat.....	13
Variables Pulled from Compustat .....	13
Compustat Tutorial.....	14
Execucomp .....	16
Variables Pulled from Execucomp .....	16

Execucomp Tutorial.....	16
Edgar Company Filings (SEC) .....	17
Variables Pulled from EDGAR.....	17
Edgar Tutorial.....	18
Federal Research and Economic Database (FRED) .....	19
Variables Pulled from FRED .....	19
FRED Tutorial.....	19
US Patent and Trademark Office (USPTO).....	20
Technological Drift .....	20
Technological Momentum.....	20
Variables Pulled from USPTO.....	21
USPTO Tutorial .....	21
Processing Raw Data .....	22
Raw Data Processing Tutorial .....	22
Error Checking .....	22
Error Checking Tutorial .....	22
Training and Testing .....	24
Validating Results.....	25
Integrity of Results.....	25
Identify Common Factors .....	25
Human Judgement on Results .....	25

# Project Summary and Overview

## Purpose

We wanted to find a question relevant to investment managers which could be answered with the help of machine learning. After discussing the types of events we wanted our model to predict, we decided on the following problem statement:

---

**Which public companies are most likely to be M&A targets in the next 12 months?**

---

## Problem Specificity

To define our problem statement more specifically, we defined this statement with the following parameters:

Project Parameters	
<b>Problem Statement:</b>	Predict the most likely publicly traded M&A targets to be acquired within the next year
<b>Acquisitions:</b>	We targeted acquisition announcements in the following 12 months. Based on date of announcement, not date of closure
<b>Time Frame:</b>	Quarterly, historical data from 1990 to Q2 2019
<b>Acquirer Stake:</b>	To maintain a large sample size, we included majority and minority stake acquisitions that fit the equity method definition
<b>Companies:</b>	All equities traded on major U.S. exchanges (NYSE, NASDAQ, AMEX)
<b>Excluded Sectors:</b>	Finance and Real Estate
<b>Model Selection:</b>	Three model types were used. A random forest model, a neural network model, and an ensemble model of the two.
<b>Training and Testing:</b>	The model was trained on historical data from 1990-2013 and tested from 2014 - Q2 2019
<b>Prediction Tools:</b>	To form our predictions, we used machine learning algorithms in SQL and Python
<b>Data Collection:</b>	Data was collected from widely available sources (Bloomberg, CapIQ, Compustat, EDGAR, FactSet, FRED, USPTO, and Yfinance)
<b>Error Checking:</b>	All datasets had to be error checked for accuracy. The model can only be as good as the data inputs.

In this document, we outline our process of building a machine learning model to solve this problem. Our goal is for the process to be replicable and adaptable for other uses.

## What is Machine Learning and Deep Learning?

Machine learning is an application of artificial intelligence in which data is input into specialized algorithms, which are able learn from the data and generate models that can be used to make out-of-sample predictions.

Within machine learning there are a variety of algorithms which can be used, such as random forests and neural networks.

Deep learning is a subset of machine learning which utilizes neural networks to find relevant connections among and between data points.

## Why Machine Learning?

Machine learning is ideal when attempting to analyze very large sets of complex data. It is particularly well-suited for problems where there are not clear linear correlations between individual variables and outcomes. Because of the large amount of financial information available to us, and the fact that countless variables can affect M&A outcomes (often in inter-

related and non-linear ways), we determined that machine learning models could be a valuable tool for predicting M&A events.

## **Building a Machine Learning model**

The basic steps of building the model include choosing variables, pulling data, error checking the data, writing the code, running the model, and analyzing the results. The timeline for our project was about twelve weeks.

## **How is this manual structured?**

This manual outlines our process in building the machine learning model. We go into detail of each step along the way in order for our process to be utilized by other professionals hoping to conduct a similar type of analysis.

# Variable Selection

## Choosing Variables and Categorizing

Knowing that variable selection was the foundational key to our project, we vetted variables in many ways.

- We reviewed key metrics commonly used by analysts and investment managers to value businesses.
- We reviewed proxy statements and press releases from mergers for stated factors or metrics.
- We reviewed publications describing key variables and considerations in M&A deals.
- We continued this process with several other sources until we felt comfortable that our list was comprehensive and contained minimal bias.

Once we had a large list of variables, we grouped them into distinct categories. This enabled us to compare variables and ensure that our list was comprehensive. The following is our finalized list of categories along with rationale on why we believed the category was relevant.

1. Profitability & Growth
  - Is this an attractive business?
  - Do trends in growth/profit affect targeting of companies?
2. Returns
  - Has the company generated strong returns historically? Is it a laggard?
  - Is this an efficient capital allocator?
3. Capital Structure/Liquidity
  - How does capital structure affect the attractiveness of a potential acquisition?
4. Trading Multiples
  - How is this business valued relative to peers?
5. Executive Compensation/Demographics
  - Do incentives for key decision makers affect M&A activity?
6. Market/M&A Trends
  - Is the macro environment favorable for M&A?
7. Board Ownership/Share Classes
  - Does ownership structure facilitate or impede M&A?
8. Patent Profile
  - Can intellectual property cause business synergies and lead to M&A?
9. Industry Trends
  - Academic research reflects industry waves in M&A.
10. Security Pricing/Volume
  - How do fluctuations in price and trading activity affect M&A?
11. Predicted Variable
  - Was this company the target of an acquisition announcement in the following 12 months?

## Comprehensive Variable List

After trimming our initial variable list based on several factors that will be detailed later, we settled on this final variable list:

Category	Variable	Category	Variable	Category	Variable
Top Line Growth	Revenue (\$ value) Revenue growth (%) Growth trend (3 year CAGR) Growth trend (5 year CAGR)	Market/M&A Trends	LTM acquisitions by sector LTM acquisitions in total market Cross sectional median of fundamental variables Michigan consumer sentiment index VIX Federal Funds Rate 1-year Treasury Rate 5-Year Treasury Rate 10-Year Treasury Rate 20-Year Treasury Rate CPI LIBOR Rates Corporate tax rate SPY 1-5 year price change Personal consumption expenditure Average hourly earnings MZM Money stock	Patent Profile Board	Technological Drift Technological Momentum Chairman and CEO same person? Chairman tenure Chairman age Is the Board Staggered? Board average age Board size Board % female Board average tenure % of independent directors on board Year-over-year change in annual meetings
Profitability and growth	Gross Profit (\$ value) Gross Profit growth (%) Gross Profit margin (%) Gross Profit margin growth (%) SG&A (\$ value) SG&A growth (%) SG&A margin (%) SG&A margin growth (%) R&D (\$ value) R&D growth (%) R&D margin (%) R&D margin growth (%) EBIT (\$ value) EBIT growth (%) EBIT margin (%) EBIT margin growth (%) NOPAT (\$ value) NOPAT growth (%) NOPAT margin (%) NOPAT margin growth (%) EBITDA (\$ value) EBITDA growth (%) EBITDA margin growth (%) EBITDA margin growth (\$ value) Restructuring Expense (\$ value) Restructuring Expense Growth (%) Asset Impairment (\$ value) Asset Impairment growth (%) Share-based payment Expense (\$ value) Share-based payment Expense growth (%) Net Income (\$ value) Net Income growth (%) Net Income margin (%) Net Income margin growth (%) Basic EPS (\$ value) Basic EPS growth (%) Diluted EPS (\$ value) Diluted EPS growth (%) Days sales outstanding (value) Days sales outstanding growth (%) Days inventory outstanding (value) Days inventory outstanding growth (%) Days payables outstanding (value) Days payables outstanding growth (%) Cash conversion cycle (value) Cash conversion cycle growth (%) CAPEX (\$ value) CAPEX growth (%) CAPEX as a % of sales (%) CAPEX as a % of sales growth (%) CAPEX as a % of PP&E (%) CAPEX as a % of PP&E growth (%) Levered Free Cash Flow (\$ value) Levered Free Cash Flow growth (%) Levered Free Cash Flow margin (%) Levered Free Cash Flow margin growth (%)	Industry Trends	Industry classification (dummy variable) Industry 1-5 year index price change Industry cross sectional median of fundamental variables Industry momentum Industry volatility Industry beta Concentration Ratio (top 4 sales by total industry)	Demographics	% institutional owned Single Class Stock System Dual Class Stock System
Executive Compensation	Base Salaries Base as % of total compensation Average Executive Total Compensation Average Director Stock Based Compensation Average Board Total Compensation Total CEO Compensation Total CEO Bonuses Highest Bonus Amount Paid % of options awarded Value of options awards	Security Pricing / Volume	Net share issues Market Beta Insider ownership Institutional ownership 1-5 year variance of daily returns 1-5 year price change Distance from 52 week high / low Distance from 3 year high / low Dividend Yield SMA 200 SMA 50 Relative volume Average volume	Textual Analysis (10K & 10Q)	"Competition" "Test" "Improvements" "Unrecognized" "Testing" "Accounted" "Allocated" "Component" "Strategy" "Longlived" "Yield" "Next" "Combined" "Then" "Trends" "Retirement" "Earned" "Environment" "Protection" "Final" "Investing" "Cumulative" "Low" "Investing" "Out" "Treasury" "Performed" "Impaired" "Leased" "Assumed" "Goods" "FASB" "Early" "Adoption" "Entitled" "Evaluate" "Standard" "Transfer" "GAAP" "Strategic" "Variable" "Timing" "Proprietary" "Name" "Recognize" "Major" "Proxy" "Ending" "Unless" "Dependent"
Predicted Returns	Acquired in following 12 months ROIC (%) ROIC growth (%) Invested Capital (\$ value) Invested Capital growth (%) Return on Equity (%) Return on Equity growth (%) Return on Assets (%) Return on Assets growth (%) EBITDA Depreciation Factor (value) EBITDA Depreciation Factor growth (%)	Capital Structure / Liquidity	Market Capitalization (\$ value) Market Capitalization growth (%) Enterprise Value (\$ value) Enterprise Value growth (%) Cash / Market Capitalization (%) Cash / Market Capitalization growth (%) Enterprise Value / LTM EBITDA (x) Enterprise Value / LTM EBITDA growth (%) Share Price / LTM EPS (x) Share Price / LTM EPS growth (%) Market Capitalization / Book Value of Equity (x) Market Capitalization / Book Value of Equity growth (%) Total Debt (\$ value) Total Debt growth (%) Book value of Equity (\$ value) Book value of Equity growth (%) Total Debt As % of Total Capitalization (%) Total Debt As % of Total Capitalization growth (%) Leverage (multiple) Leverage growth (%) Total Assets (\$ value) Total Assets growth (%) Non-controlling interests (\$ value) Non-controlling interests growth (%) Interest coverage ratio (value) Interest coverage ratio growth (%) Quick ratio (value) Quick ratio growth (%) Current ratio (value) Current ratio growth (%) Asset turnover (value) Asset turnover growth (%) Goodwill + Intangibles (\$ value) Goodwill + Intangibles growth (%) Goodwill + Intangibles as a % of Total Assets (value) Goodwill + Intangibles as a % of Total Assets growth (%)		Current Year (Ex: "2019") *Previous 10 Years* *Next 10 Years*

## Variables by Data Source

Taking the comprehensive list, we tied each variable to a specific data source that we felt would contain the most complete and accurate data. We tried to keep our number of data sources low while still grasping all points of data that we felt were necessary to the project. We used the following sources to capture the variables listed under each.

1. Bloomberg
2. Capital IQ
3. Compustat
4. SEC EDGAR
5. ExecuComp
6. FRED
7. US Patent and Trade Office (USPTO)
8. Yahoo! Finance

## Exclusions and Limitations

We excluded two sectors from our project: Financial Institutions and Real Estate. Due to the structure of businesses within these industries, we felt that excluding them removed outliers that would have skewed our overall analysis. We also excluded certain variables when data was limited, or when we determined they were redundant or not applicable to the scope of the project. The following tables contain some variables that we excluded due to these factors, and we recognize that there is a much larger list of variables that never made it to our initial list. We believe that further iterations of this project could greatly expand the variable list used.

Category	Excluded Variable	Category	Excluded Variable
Demographics	CEO ethnicity/race	Security Pricing / Volume	Long-term reversal
Board	Chairman ethnicity/race		Average age of debt
Ownership	% institutional owned growth (%)	Executive Compensation	Long-term incentive plans
	% activist owned		Deferred Compensation
	% activist owned growth (%)		Strike Prices
	% hedge fund owned		Expiration Dates
	% hedge fund owned growth (%)		Value of in-the-money granted options
	% insider owned		Value of out-of-the-money granted options
	% insider owned growth (%)		Health benefits upon retirement
	% founder owned		Golden parachutes
	% founder owned growth (%)		Private Jet
Industry Trends	Concentration Ratio (top 4 sales by total industry)		Travel Reimbursements

## Bias

Because we are eliminating two sectors and have chosen the variables we will use in the project, our data carries some human bias. We have attempted to remove as much bias as possible in regard to the inputs of our model but do believe a small amount of human judgment is necessary in order to obtain the best results.

## Availability

Due to the lack of comprehensive data sets, some variables that we would have liked to include will not be used in our analysis. When training a model on this many companies spanning 30 years of data, some datasets will not be available or reliable, thus limiting the usable inputs.

## Time Variance

Because all variables are not reported or analyzed on the same time intervals, we needed to reconcile these differences in order to standardize our inputs. We have done so by taking the most recent values for every variable at the start of each calendar quarter.

## Specificity

Some industries see a higher frequency of M&A activity. In our analysis, we have included variables that apply to all companies in our dataset. It is possible that by focusing on specific sectors and adding industry specific variables, the model could provide much better insight to potential M&A activity. However, we found that focusing on a single industry reduced our prediction sample by too much to be reliable.

## Time Series

We attempted to be all-inclusive with the times series considerations. Depending on the variables, we will be constructing data around the following time series:

1. Quarterly (Q)
2. Yearly (Y)
3. Quarter-over-quarter (QoQ)
4. Year-over-year (YoY)
5. Last 4 calendar quarters (CQ-4)
6. Last 12 months (LTM)
7. Last 12 months, year ago (LTM-1)
8. Calendar Year (CY+1)
9. Calendar Year, year ago (CY+1-1)

## Data Collection

After we determined the variables to include in our data set, we evaluated potential sources of that information. The top priorities of data sourcing were accuracy and reliability. In addition, we needed to source the data in a way that allowed exportation of large data sets in a machine-readable format.

To ensure integrity of data we chose to draw from sources commonly used in the finance industry. Once all target variables were collected, we compiled the data into central, aggregated SQL files using Python. Below we go over the various sources of data and the steps required for collection.

### Bloomberg Terminal and Bloomberg API

The Bloomberg Terminal is a widely used source for real-time financial information. Many finance professionals have access to Bloomberg and are familiar with its interface.

The Bloomberg Open API (Application Programming Interface) allows applications like Microsoft Excel to access Bloomberg data through the Terminal. Excel add-ins from the Bloomberg Terminal facilitate building spreadsheets with this market data. The Bloomberg API allows integration of real-time or delayed data, historical data, reference data and intraday data.

### Variables Pulled in Bloomberg

Category	Variable	Category	Variable
Demographics	Mgmt % female	Market/M&A Trends	Michigan consumer sentiment index
	CEO Tenure		VIX
	CEO Age		Corporate tax rate
	CEO Founder?	Security Pricing / Volume	Insider ownership
	CEO gender		Institutional ownership
Board	Chairman and CEO same person?		Dividend Yield
	Chairman tenure	Ownership	% institutional owned
	Chairman age	Share Classes	Single Class Stock System
	Is the Board Staggered?		Dual Class Stock System
	Board average age	Market/M&A Trends	LTM acquisitions by sector
	Board size		LTM acquisitions in total market
	Board % female		
	Board average tenure		
	% of independent directors on board		
	Year-over-year change in annual meetings		

### Bloomberg API Tutorial

For a general tutorial on pulling data using Bloomberg's API, see the Nicholas Center's API Guide:

<https://github.com/UW-Nicholas-Center/Adage/blob/master/Bloomberg%20API%20Guide.pdf>

See source code at:

<https://github.com/UW-Nicholas-Center/Adage/tree/master/Corporate%20Governance>

## Capital IQ

Capital IQ (CAPIQ) has a large database of historical information, including SEC filings, for public companies and allows for easy exportation of data sets into CSV files. There are many filters available for data mining which can make the interface seem complex; but once one understands which parameters they need and where they are located on the website, it is fairly intuitive. Using CapIQ requires a paid subscription.

### Variables Pulled in Capital IQ

Category	Variable
Predicted Variables	Acquired in following 12 months
Ownership	% institutional owned
Share Classes	Single Class Stock System Dual Class Stock System
Market/M&A	LTM acquisitions by sector
Trends	LTM acquisitions in total market

### Capital IQ Tutorial

1. Login to [Capital IQ](#)
2. At the top of the screen click on "Screening"
3. On the "Screening and Analytics" screen click "Transaction Screening"
4. Parameter selection on the "Transaction Screening" page
  - a. Under header "M & A Details" click on "Dates"
    - i. Under "Time Frame" select "From" and input the date 01/01/1990
    - ii. Click on "Add Criteria" and the parameter should show as criteria #1
  - b. Under "Equity Details" click on "Exchanges"
    - i. Select "Major US Exchanges" from "Available Items" column
    - ii. Click on right arrow to add Major US Exchanges to "Selected Items" column
    - iii. Click on "Add Criteria" and parameter should show as criteria #2
  - c. Under "General Transaction Details" click on "Types"
    - i. Under "Available Items" column select "Merger/Acquisition"
    - ii. Click on right arrow to add Merger/Acquisition to "Selected Items" column
    - iii. Click on "Add Criteria" and parameter should show as criteria #3
5. Once all parameters have loaded click on the excel icon at the top of page to export data to an excel file

## Yahoo! Finance API

Yahoo! Finance is a free, online resource which provides financial news and data such as stock quotes, press releases, financial reports, and original content. The Yahoo! Finance API (Application Programming Interface) allows the user to scrape data in bulk from the Yahoo! Finance website. In our project, we used this API to pull daily stock prices and trading volumes.

### Variables Pulled from Yahoo! Finance API

Category	Variable
Security Pricing / Volume	Net share issues
	Market Beta
	1-5 year variance of daily returns
	1-5 year price change
	Distance from 52 week high / low
	Distance from 3 year high / low
	SMA 200
	SMA 50
	Relative volume
	Average volume
Industry Trends	Industry 1-5 year index price change
	Industry momentum
	Industry volatility
	Industry beta

## Yahoo! Finance API Tutorial

See source code at: <https://github.com/UW-Nicholas-Center/Adage/tree/master/Market>

## Compustat

Compustat is a widely used database produced by Standard & Poor's which provides comprehensive financial and market information on active and inactive companies, indices and industries around the world. Through our university, we are able to access Compustat through Wharton Research Data Service (WRDS.)

### Variables Pulled from Compustat

Category	Variable	Category	Variable
<b>Top Line Growth</b>	Revenue (\$ value) Revenue growth (%) Growth trend (3 year CAGR) Growth trend (5 year CAGR)	<b>Returns</b>	ROIC (%) ROIC growth (%) Invested Capital (\$ value) Invested Capital growth (%)
<b>Profitability and growth</b>	Gross Profit (\$ value) Gross Profit growth (%) Gross Profit margin (%) Gross Profit margin growth (%) SG&A (\$ value) SG&A growth (%) SG&A margin (%) SG&A margin growth (%) R&D (\$ value) R&D growth (%) R&D margin (%) R&D margin growth (%)  EBIT (\$ value) EBIT growth (%) EBIT margin (%) EBIT margin growth (%) NOPAT (\$ value) NOPAT growth (%) NOPAT margin (%) NOPAT margin growth (%) EBITDA (\$ value) EBITDA growth (%) EBITDA margin (%) EBITDA margin growth (%) Restructuring Expense (\$ value) Restructuring Expense Growth (%) Asset Impairment (\$ value) Asset Impairment growth (%) Share-based payment Expense (\$ value)  Share-based payment Expense growth (%)  Net Income (\$ value) Net Income growth (%) Net Income margin (%)	<b>Capital Structure / Liquidity</b>	Return on Equity (%) Return on Equity growth (%) Return on Assets (%) Return on Assets growth (%) EBITDA Depreciation Factor (value) EBITDA Depreciation Factor growth (%)  Total Debt (\$ value) Total Debt growth (%) Book value of Equity (\$ value) Book value of Equity growth (%) Total Debt As % of Total Capitalization (%) Total Debt As % of Total Capitalization growth (%) Leverage (multiple) Leverage growth (%) Total Assets (\$ value) Total Assets growth (%) Non-controlling interests (\$ value) Non-controlling interests growth (%) Interest coverage ratio (value) Interest coverage ratio growth (%) Quick ratio (value) Quick ratio growth (%) Current ratio (value) Current ratio growth (%) Asset turnover (value) Asset turnover growth (%) Goodwill + Intangibles (\$ value) Goodwill + Intangibles growth (%) Goodwill + Intangibles as a % of Total Assets (value) Goodwill + Intangibles as a % of Total Assets growth (%)  <b>Trading Multiples</b> Market Capitalization (\$ value) Market Capitalization growth (%) Enterprise Value (\$ value)

Net Income margin growth (%)	Enterprise Value growth (%)
Basic EPS (\$ value)	Cash / Market Capitalization (%)
Basic EPS growth (%)	Cash / Market Capitalization growth (%)
Diluted EPS (\$ value)	Enterprise Value / LTM EBITDA (x)
Diluted EPS growth (%)	Enterprise Value / LTM EBITDA growth (%)
Days sales outstanding (value)	Share Price / LTM EPS (x)
Days sales outstanding growth (%)	Share Price / LTM EPS growth (%)
Days inventory outstanding (value)	Market Capitalization / Book Value of Equity (x)
Days inventory outstanding growth (%)	Market Capitalization / Book Value of Equity growth (%)
Days payables outstanding (value)	Cross sectional median of fundamental variables
Days payables outstanding growth (%)	Industry classification (dummy variable)
Cash conversion cycle (value)	Industry cross sectional median of fundamental variables
Cash conversion cycle growth (%)	Concentration Ratio (top 4 sales by total industry)
CAPEX (\$ value)	
CAPEX growth (%)	
CAPEX as a % of sales (%)	
CAPEX as a % of sales growth (%)	
CAPEX as a % of PP&E (%)	
CAPEX as a % of PP&E growth (%)	
Levered Free Cash Flow (\$ value)	
Levered Free Cash Flow growth (%)	
Levered Free Cash Flow margin (%)	
Levered Free Cash Flow margin growth (%)	

## Compustat Tutorial

1. Go to <https://wrds-www.wharton.upenn.edu/> and create an account
  - a. On the home page, you have an option to choose a “Subscription.” Choose “Compustat – Capital IQ.”
  - b. From this page, choose “North America – Daily” under the Compustat header.
  - c. On the next page, choose “Fundamentals Quarterly”
  - d. You are brought to the data selection page.
    - i. Step 1: Choose Your Date Range
      1. We chose data from Jan 1990 – (current month) 2019
    - ii. Step 2: Apply Your Company Codes
      1. Leave the “TIC” bubble filled in
      2. Fill in the bubble next to the “Browse” box. We have a file called “tickers” which we used, which contains the tickers of all companies we want to include in our project.
      3. Screening Variables: don’t change anything here
    - iii. Step 3: Choose Variable Types
      1. First box “Select Variable Types:” don’t change anything here.

2. Second box: click the bubbles of each variable you want to include in your search. Be sure to include "Ticker Symbol." Easiest to search by variables in the search field.
    - a. Note that each variable has a code in front of it. If the code ends in Y, it's a yearly variable. If the code ends in Q, it's a quarterly variable.
  3. Third box: Conditional Statements: leave blank
- iv. Step 4: Select Query Output
1. Choose comma-delimited text (.csv)
- e. Click Submit and save your file

## Execucomp

Execucomp is a database which is part of Compustat / Capital IQ and is offered through Wharton Research Data Service. It provides comprehensive data on executive compensation for companies traded on major US exchanges.

### Variables Pulled from Execucomp

Category	Variable
Executive Compensation	Base Salaries
	Base as % of total compensation
	Average Executive Total Compensation
	Average Director Stock Based Compensation
	Average Board Total Compensation
	Total CEO Compensation
	Total CEO Bonuses
	Highest Bonus Amount Paid
	% of options awarded
	Value of options awards

### Execucomp Tutorial

1. Go to <https://wrds-www.wharton.upenn.edu/>
  - a. Under subscriptions, choose “Compustat – Capital IQ”
  - b. Under “Compustat,” choose “Execucomp – Monthly Updates”
  - c. On the next page, select the appropriate variable (i.e., annual compensation)
  - d. Step 1: Choose Your Date Range
    - i. We chose data from Jan 1990 – (current month) 2019
  - e. Step 2: Apply Your Company Codes
    - i. Leave the “Ticker” bubble filled in
    - ii. Fill in the bubble next to the “Browse” box. We have a file called “tickers” which we used, which contains the tickers of all companies we want to include in our project.
  - f. Step 3: Query Variables
    - i. Click the bubbles of each variable you want to include in your search. Be sure to include “Ticker Symbol.” Easiest to search by variables in the search field.
  - g. Step 4: Select Query Output
    - i. Choose comma-delimited text (.csv)
  - h. Click Submit and save your file

## Edgar Company Filings (SEC)

We wanted our model to consider management's analysis of the current state of a company. We turned to the Management Discussion and Analysis (MD&A) section of companies' 10-K filings for this data. We used Python, this time writing a script that scrapes these filings from EDGAR and parses their contents to count the frequency of different words and phrases. This can be done for several decades of filings from thousands of different companies in a matter of hours.

While this is a relatively simple method of language processing, a deep learning model will be able to pick up on relationships between words, as well as between the contents of the MD&A and the other variables in our dataset.

### Variables Pulled from EDGAR

Category	Variable	Category	Variable
<b>Textual Analysis (10K &amp; 10Q)</b>	"Competition"	<b>Textual Analysis (10K &amp; 10Q)</b>	"Impaired"
	"Test"		"Leased"
	"Improvements"		"Assumed"
	"Unrecognized"		"Goods"
	"Testing"		"FASB"
	"Accounted"		"Early"
	"Allocated"		"Adoption"
	"Component"		"Entitled"
	"Strategy"		"Evaluate"
	"Longlived"		"Standard"
	"Yield"		"Transfer"
	"Next"		"GAAP"
	"Combined"		"Strategic"
	"Then"		"Variable"
	"Trends"		"Timing"
	"Retirement"		"Proprietary"
	"Earned"		"Name"
	"Environment"		"Recognize"
	"Protection"		"Major"
	"Final"		"Proxy"
	"Investing"		"Ending"
	"Cumulative"		"Unless"
	"Low"		"Dependent"
	"Investing"		Current Year (Ex: "2019")
	"Out"		*Previous 10 Years*
	"Treasury"		*Next 10 Years*
	"Performed"		

## Edgar Tutorial

1. To access data manually, go to <https://www.sec.gov/edgar/searchedgar/accessible-edgar-data.htm>
2. Because we needed to pull data for such a large number of companies over a period of several decades, we wrote a Python script to automatically scrape 10-K filings from Edgar. The code we used to do this can be found at <https://github.com/UW-Nicholas-Center/Adage/blob/master/Other/parEdgar.py>

## Federal Research and Economic Database (FRED)

The Federal Reserve Economic Database is a widely used database maintained by the research division of the Federal Reserve Bank of St. Louis. It has 500,000+ economic time series from 87 sources, covering such areas as banking, business, consumer price indices, employment, population, exchange rates, GDP, interest rates, and more. The time series are compiled by the Federal Reserve and are collected from government agencies such as the U.S. Census and the Bureau of Labor Statistics.

### Variables Pulled from FRED

Category	Variable
Market/M&A	Federal Funds Rate
Trends	1-year Treasury Rate
	5-Year Treasury Rate
	10-Year Treasury Rate
	20-year Treasury Rate
	CPI
	LIBOR Rates
	Personal consumption expenditure
	Average hourly earnings
	MZM Money stock

### FRED Tutorial

Go to <https://fred.stlouisfed.org/>

1. Create an account.
2. Type your search criteria in the search field.
3. On the next page, filter “concepts” in the box on the left side of the page.
4. Select any relevant data that you would like to include.
5. Click on the “add to data list” button.
6. Download and save the data.

## US Patent and Trademark Office (USPTO)

Companies are often acquired for their intellectual property. Although it is common practice for companies to disclose how much they are spending on research and development, there are no widely available metrics that contain information about the products of this spending.

To fix this problem we turned to the US Patent and Trademark Office's online public database, which contains information on every patent granted since 1976. We downloaded this data in bulk (over 4 million patents in total) and ran a textual analysis algorithm to identify which patents belonged to companies traded on major US exchanges, along with which of the 427 USPTO technological classes the patent belonged.

We used this data to systematically construct "technological profiles" for each company at every quarter throughout their existence. These profiles are essentially a snapshot of which technological classes a firm has been granted patents in within the last five years.

We used these technological profiles as the basis for two metrics that quantify different aspects of a company's intellectual property: technological drift and momentum.

### Technological Drift

Companies that are exploring new sectors of research, rather than focusing on existing areas of expertise, are likely in a transitory period of their business. We believe that this could be an important indicator of both a company's internal strategy and how they are viewed by other members of their industry, both of which could impact acquisition likelihood.

In order to test these beliefs, we created a metric called "technological drift" to quantify the year-over-year change in a company's technological profile. We use the formula

$$TECHDRIFT_{it} = \frac{1}{2} * \sum_{c=0}^n |PATENTPROP_{ic,t} - PATENTPROP_{ic,t-12}|$$

to find the drift index for company  $i$  in month  $t$ , where  $PATENTPROP_{ic,t}$  is equal to the proportion of Company  $i$ 's patents belonging to patent class  $c$  that were granted in the five years prior to month  $t$ . This formula returns values between 0 and 100%, with higher numbers corresponding to a higher amount of technological drift.

Our data confirms the relevance of this factor, showing that companies with high amounts of internal technological drift are significantly less likely to be acquired in the following year.

### Technological Momentum

Companies with closely aligned technological expertise often have similar outside factors that affect their firm's value, even if they operate in different industries. Knowing how a company is

performing relative to its technological peers can be just as important as understanding its positioning among industry competitors. If a firm is performing stronger than its technological peers, it may be a sign that it has been able to uncover unique value within its field of expertise.

We created a metric called “technological momentum” that attempts to capture the recent performance of a company’s tech peers. We calculate the technological similarity between two companies by using the formula

$$TECH_{ijt} = \frac{(T_{it} T'_{jt})}{(T_{it} T'_{it})^{1/2} (T_{jt} T'_{jt})^{1/2}}$$

where  $T_{it}$  is the vector of firm  $i$ ’s proportional share of patents across the USPTO classes over the last 5 years as of time  $t$ . This gives us a value ranging from 0 to 1 for every pair of firms, with higher values for more technologically similar firms.

After finding the similarity between each pair of firms, we can calculate a technology momentum index for each firm by weighting the previous quarter’s stock returns of all other firms in the market based on their technological similarity. Mathematically, this technological momentum factor return is defined as

$$TECHRET_{it} = \frac{\sum_{j \neq i} TECH_{ijt} * RET_{jt}}{\sum_{j \neq i} TECH_{ijt}}$$

Preliminary analysis indicates that companies with a negative technological momentum factor are 22.6% more likely to be acquired in the following twelve months than those with positive momentum. This factor may also have more complex correlations with other variables in our dataset.

### Variables Pulled from USPTO

Category	Variable
Patent	Technological Drift
Profile	Technological Momentum

### USPTO Tutorial

See source code at: <https://github.com/UW-Nicholas-Center/Adage/tree/master/Patents>

## Processing Raw Data

Once we collected all of the raw data for our model, we needed to standardize formatting and aggregate all of the variables into a single database. To do this, we labeled each datapoint with the ticker of its company and the first calendar quarter following when it was originally available. We saved each individual, labelled dataset into SQL tables, and then joined the SQL tables on ticker and date.

### Raw Data Processing Tutorial

The Python code we used for processing can be found here:

<https://github.com/UW-Nicholas-Center/Adage>

## Error Checking

When manually pulling large amounts of data from several different sources, there is a chance of user error in mapping and aligning the data. To ensure the integrity of our data set, we extracted random points of data from each set that we pulled, and manually cross-referenced them for accuracy.

### Error Checking Tutorial

1. Download DB Browser for SQLite
  - a. Use the following link: <https://sqlitebrowser.org/dl/>
  - b. Select the option that matches your computer specs
    - i. You may need to research your computer's specs to make correct decision
2. Create organizational system to track progress and potential errors
  - a. Create excel file (or other preferred application)
  - b. Have the following columns:
    - i. SQL Dataset – Name of the dataset you are checking
    - ii. Date Check – Track when you performed check
    - iii. Original Source – Where dataset is pulling originally
    - iv. Verification Source – Dataset you are using to check
      1. Original & Verification sources should be different
    - v. Notes – Section to put comments if needed
    - vi. Owner – Team member responsible for checking
  - c. Fill out file as errors are checked
3. Create file archiving system
  - a. You should create the file archive before downloading the dataset, as this will keep you organized
  - b. Files will need to include:
    - i. .db files you are error checking – Allows others to see source
    - ii. .csv file of dataset
    - iii. .xlsx file – This is the file which you will actually manipulate
4. Creating reviewable files
  - a. Download .db files that you need to error check

- b. Open the DB Browser for SQLite, click “Open Database”, and select the .db file that you downloaded
  - c. Click on the “Browse data tab”, selecting the “Save” button and choose Export to CSV
  - d. Save CSV file into corresponding file within your archive
  - e. Save As the CSV file into Excel, so you will be able to review
5. Reviewing individual datasets
  - a. Open the excel file of dataset
    - i. You want to use the excel file so your modifications save
  - b. Select between 15-20 individual rows to review
    - i. Be sure to review rows in different time periods, not just recent time periods
    - ii. Aside from managing the time period, selection of rows should be random
  - c. Now check rows against data source different from the original source
    - i. If an error is found be sure to highlight the cell and label it
6. Drill-down on the individual errors
  - a. Review specific errors to see what could be causing them
  - b. Some errors might due to how the database is set-up, in which case you would need to check with the data provider
  - c. All errors should have some sort of explanation

## Training and Testing

Hyperparameters shape the architecture of a machine learning model. Hyperparameters are not the parameters of the model (or how the model transforms inputs into outputs), but rather how we want the model to be structured. We provide hyperparameters laying out the floorplan and allow the model to fill the space with its own learned parameters.

Part of the selection of hyperparameters is experimentation to determine what works “best,” including runtime, accuracy, computation capacity, and proper use of data.

1. Follow this link to see the code we used for creating our models:  
<https://github.com/jKard1210/Mergers-and-Acquisitions-ML/tree/master/Modeling>
2. Each type of model requires a different set of hyperparameters. After experimenting and tweaking the default parameters, we used the following:
  - a. Neural Network HP
    - i. Model was created using SciKit Learn’s MLP Classifier package and trained incrementally with Dask
    - ii. Alpha: 0.0001
      1. The conservative versus aggressive weighting of inputs
      2. In our situation, we needed a very low alpha due to sensitivity of our predicted variables
    - iii. Activation Function: RELU
      1. Defines the output of each node given the set of inputs
      2. RELU is a Rectified Linear Unit, using summed weights of inputs at each node to determine output
    - iv. Hidden Layer Sizes: 100, 20, 5
      1. Each layer is a set of nodes summing to provide predicted variables
      2. Three layers is the standard. 100, 20, 5 chosen for runtime
    - v. Solver: Adam
      1. Chosen for runtime, deep learning sensitivity
  - b. Random Forest HP
    - i. Model was created using SciKit Learn’s Random Forest package
    - ii. Number of Estimators: 100
      1. The number of decision trees being constructed in our model and aggregated to final output
      2. ROC AUC didn’t improve significantly when we tried adding additional estimators
    - iii. Max Depth: 4
      1. Max number of layers in a decision tree, chosen to avoid overfitting and maintain most impactful variables

## Validating Results

After receiving preliminary results from the model, we needed to validate results using the following tests:

1. Integrity of results
2. Identify common factors between hits and misses
3. Find current predictions and make rational judgements on the results

### Integrity of Results

The results need to be validated by reviewing samples of the data for where the predictions occurred. In our situation, validating the initial results resolved the following concerns:

1. Minority/immaterial stake acquisitions
2. Double counting predicted vs. actual hits
3. False positives (predicting an acquisition of a company that has already been acquired)
4. Predictions accounting for proper gap in quarters

This review required the following research by our team:

1. Review model output data
2. Cross-reference CapitalIQ
3. Recent news on predicted companies
4. Filings on previous announced acquisitions (actual hits)

### Identify Common Factors

The comparison of hits and misses should show common threads among those firms selected. While the threads will not be identical, the model should be weighting and valuing similar items across all companies if functioning correctly. This can be validated by retrieving common company metrics or ratios and comparing for patterns or similarities.

### Human Judgement on Results

After testing the integrity and viability of the results, we used human judgement to test the soundness of the predictions by applying the following fundamental analysis techniques:

- Reviewed news stories, press releases, and company filings
- Noted stock performance and movement
- Compared financial metrics and ratios to industry benchmarks
- Looked at value-driving financials such as growth and free cash flow
- Analyzed industry and macroeconomic factors impacting the firm
- Looked into history of M&A activity at the company
- Reviewed past performance and experience of senior leadership

The fundamental analysis allows us to vet the results and compare them to what an analyst would traditionally look for when screening targets. For our purposes, we were able to key in on a few of the predictions we found most interesting and most likely to occur.