# Introduction

In the professional world a lot of systems are not built with data science in mind. To successfully gather and analyze data, data scientist must work around existing systems that businesses use to gather and store data. This tutorial will introduce how to process and analyze data from third party business management sites.

A lot of business applications can output reports so I will be focusing on how to use these reports to successfully get information about the company. This can be both simple information about the business but also much more advanced machine learning methods can be employed to have a deeper understanding of how the business runs

**Tutorial Content**

In this tutorial I will show how to capture some basic metrics about the business operations using Pandas and also more advanced metrics using machine learning libraries ENTER HERE

Data will be from [Main Event Catering](http://maineventcaterers.com/) which is stored in [CaterXpert](http://www.caterxpert.com/caterxpert/caterxpert.html" \o "CaterXpert" \t "_blank) a business management system made specifically for the catering business. Many different types of business have systems very similar to this that keep track of all sales and expenses that are relevant to the industry.

In this tutorial I will go over

* Libraries
* Exporting and processing reports
* Basic Metrics
* More advanced metrics using machine learning

# Libraries

Before getting started we will import all libraries that are needed for this project

# Loading Data

With all dependencies loaded I can now begin loading data. The data from CaterXpert is from various Excel document that tells me different parts of the business. For this I will download the reports Revenue Projection by AE & Status XLS, Tasting Analysis XLS and Post Schedules Status Report XLS. These reports show information on sales, staffing prices and tastings each a separate but equally important part of the Catering industry

How CaterXpert outputs data there is a long string of unique numbers to identify different reports like for Revenue the file name is 828251485418541212852mecreveprojbyae while for schedules it is MAINEVENT309104102017125831830PMpostschedulesstatus.net. To compensate this I use the plugin glob which uses regex to select files that match the regex. This is helpful since you may have multiple versions of the same file for different reports to capture new data so you don’t have to change the file name every time.

Another thing is because of the header you have to make room for the extra columns since it initially only makes room for 1 out of the total 75 columns for Revenue Projections which gives the error "CParserError: Error tokenizing data. C error: Expected 1 fields in line 9, saw 75" That’s why I make temporary column names to read in the data initially.

Now that the data is read in I want to get rid of the headers and change column names to be the correct information which I wrote a function that works for most of the reports.

Tasting excel is a little differently formatted so data has to be reformatted differently.

Since there is two different columns to keep track of the dates I have to combine the dates to turn it into one column with datetime objects to allow for easy manipulation of the data.

# Basic Metrics

Now that we have access to the data we can begin trying to get some information from the data. To start we will look at some simple metrics. The first thing to do is to make any columns with dates into datetime objects so we can use them to gather data on certain date ranges and make subtotal a float so it can be added.

To start we will get a simple count of all the events and the money they made within a certain time period

For a slightly more complicated metric we will look at the staff profit using two different sheets. One of the sheets holds data on how much the staffing costs for events in the time frame while the other keeps track of how much they make the company. This type of metric is important because many time different data is spread across different reports. You have to combine these reports sometime to get significant data.