# OSEMN methodology
# Excerpts from How To Work Through A Problem Like A Data Scientist By Jason Brownlee[1] and other sources

Stepan Oskin

July 19, 2019

**Abstract**

In a 2010 post Hilary Mason and Chris Wiggins described the OSEMN process as a taxonomy of tasks that a data scientist should feel comfortable working on.

The title of the post was *A Taxonomy of Data Science* on the now defunct dataists blog[2]. This process has also been used as the structure of a recent book, specifically *Data Science at the Command Line: Facing the Future with Time-Tested Tools* by Jeroen Janssens published by OReilly[3].

In this document, we take a closer look at the OSEMN process for working through a data problem.

## 1 OSEMN process

OSEMN is an acronym that rhymes with "possum" or "awesome" and stands for **Obtain, Scrub, Explore, Model, and iNterpret**.

It is a list of tasks a data scientist should be familiar and comfortable working on. Although, the authors point out that no data scientist will be an expert at all of them.

In addition to a list of tasks, OSEMN can be used as a blueprint for working on data problems using machine learning tools.

From the process, the authors point out that data hacking fits into the "O" and "S" tasks and machine learning fits into the "E" and "M" tasks, and that data science requires a combination of all elements.

## 1.1 Obtain Data

The authors point out that manual processes of data collection do not scale and that you must learn how to automatically obtain the data you need for a given problem.

They point to manual processes like pointing and clicking with a mouse and copy and pasting data from documents.

The authors suggest that you adopt a range of tools and use the one most suitable for the job at hand. They point to unix command line tools, SQL in databases, web scraping and scripting using Python and shell scripts.

Finally, the authors point to the importance of using APIs to access data, where an API may be public or internal to your organization. Often data is presented in JSON and scripting languages like Python can make data retrieval a lot easier.

## 1.2 Scrub Data

The data that you obtain will be messy.

Real data can have inconsistencies, missing values and various other forms of corruption. If it was scraped from a difficult data source, it may require tripping and cleaning up. Even clean data may require post-processing to make it uniform and consistent.

Data cleaning or scrubbing requires "command line fu" and simple scripting.

The authors point out that data cleaning is the least sexy part of working on data problems but good data cleaning may provide the most benefits in terms of the results that you can achieve.

**A simple analysis of clean data can be more productive than a complex analysis of noisy and irregular data.**

The authors point to simple command line tools such as sed, awk, grep and scripting languages like Python and Perl.

For more information, see Data Preparation Process.

## 1.3 Explore Data

Explore in this case refers to exploratory data analysis.

This is where there is no hypothesis that is being tested and no predictions that are being evaluated.

Data exploration is useful for getting to know your data, for building an intuition for it's form and for getting ideas for data transforms and even predictive models to use later on in the process.

The authors list a number of methods that may be helpful in this task:

- Command Line Tools for inspecting the data like more, less, head, tail or whatever.
- Histograms to summarize the distribution of individual data attributes.
- Pairwise Histograms to plot attributes against each other and highlight relationships and outliers.
- Dimensionality Reduction methods for creating lower dimensional plots and models of the data.
- Clustering to expose natural groupings in the data

For more information, see Exploratory Data Analysis.

### 1.4 Model Data

Model accuracy is often the ultimate goal for a given data problem. This means that the most predictive model is the filter by which a model is chosen.

Often the "best" model is the most predictive model.

Generally, the goal is to use a model to predict and interpret. Prediction can be evaluated quantitatively, whereas interpretation is softer and qualitative.

A model's predictive accuracy can be evaluated by how well it performs on unseen data. It can be estimated using methods such as cross validation.

The algorithms that you try and your biases determine the hypothesis space of possible models that can be constructed for the problem. Choose wisely.

For more information, see Modeling.

### 1.5 Interpret Results

> "The purpose of computing is insight, not numbers."
>
> —Richard Hamming

The authors use the example of handwritten digit recognition. They point out that a model for this problem does not have a theory of each number, rather it is a mechanism to discriminate between numbers.

This example highlights that the concerns of predicting may not be the same as model interpretation. In fact, they may conflict. A complex model

may be highly predictive, but the number of terms or data transforms performed may make understanding why specific predictions are made in the context of the domain nearly impossible.

The predictive power of a model is determined by its ability to generalize. The authors suggest that the interpretative power of a model are its abilities to suggest the most interesting experiments to perform next. It gives insights into the problem and the domain.

The authors point to three key concerns when choosing a model to balance predictive and interpretability of a model:

- Choose a good representation, the form of the data that you obtain, most data is messy.
- Choose good features, the attributes of the data that you select to model
- Choose a good hypothesis space, constrained by the models and data transforms you select.

## 2 Summary

In this document we've discovered the OSEMN proposed by Hilary Mason and Chris Wiggins.

OSEMN stands for Obtain, Scrub, Explore, Model, and iNterpret.

Like Knowledge Discovery in Databases and the applied machine learning process, this process can be used to work a machine learning problem.

## References


[1] J. Brownlee, "How To Work Through A Problem Like A Data Scientist," 2014.

[2] H. Mason and C. Wiggins, "A Taxonomy of Data Science," 2010.

[3] J. Janssens, *Data Science at the Command Line*. Sebastopol, CA: O'Reilly Media, Inc., first ed., 2014.