Data Science Process Alliance (https://www.datascience-pm.com/)

OSEMN Data Science Life Cycle

BY NICK HOTZ (HTTPS://WWW.DATASCIENCE-PM.COM/AUTHOR/NICK/) |
LAST UPDATED: JANUARY 19, 2023 (HTTPS://WWW.DATASCIENCE-PM.COM/OSEMN/) |
LIFE CYCLE (HTTPS://WWW.DATASCIENCE-PM.COM/CATEGORY/LIFE-CYCLE/)

Different data scientists have different processes for conducting their projects. And different types of projects require different steps. However, most data science projects flow through a similar workflow. One popular representation of this workflow is called OSEMN (pronounced “awesome”).

Whether you use this or another life cycle, understanding the basic data science project life cycle (https://www.datascience-pm.com/data-science-life-cycle/) can help you directly execute or collaborate on data science projects. So in this post, we’ll explore

- What is OSEMN?
- Should you use OSEMN?
- What are some alternative frameworks?

What is OSEMN?

Hilary Mason and Chris Wiggins created OSEMN in their 2010 post called “A Taxonomy of Data Science” on a now-defunct website (visit an archived version (https://web.archive.org/web/20160220042455/dataists.com/2010/09/a-taxonomy-of-data-science/) of this post). OSEMN is a somewhat clever acronym with each letter representing a phase of a data science project:

- **O**btain
- **S**crub
- **E**xplore
- **M**odel
- i**N**terpret

Let’s break this down.

Obtain

> “pointing and clicking does not scale”
>
> – MASON AND WIGGINS

Without data, you don’t have a data analytics or data science project. Therefore, obtaining data is the first step of OSEMN.

A successful project typically requires more than just an ad hoc manual process for gathering data. Rather, you need to understand where to find the data:

- Do you already have access to the data? Then query it.
- Can you get access to the data? Request access.
- Is the data available for purchase? Conduct a value-cost tradeoff and (if deemed appropriate) purchase it.
- Can you create the data? Set up systems to capture the data.

Regardless of the approaches you take, ensure that you obtain and use the data responsibly (https://www.datascience-pm.com/achieving-responsible-ai/).

Scrub

> "the world is a messy place"
>
> -MASON AND WIGGINS

Just because you have data, doesn't make it useful. Indeed, data rarely arrives in the format desired for evaluation. Rather, it often contains many inconsistent, erroneous, or extraneous data points. In essence, a great deal of what is collected is simply "noise."

Thus, OSEMN's second phase is to *scrub* the data to convert the raw data into a usable format. The exact steps of *scrubbing* vary based on the project intent and the data set. However, common steps include:

- Assessing data quality
- Removing irrelevant or duplicate data points
- Imputing missing values (guessing what the value should be)
- Combining different data sets (often using SQL joins)
- Load the data into a new target location (e.g. an AWS Bucket)
- Documenting the transformations taken (you or a colleague might want to revisit this later)

Many consider this phase of a data science project to be the lengthiest. But without it, your project might fall victim to the adage: "garbage in, garbage out", whereby you produce bad results based on bad data. Indeed, Mason and Wiggins comment that "scrubbing data is the least sexy part of the analysis process, but often one that yields the greatest benefits." Moreover, Andrew Ng (Coursera founder) evangelizes data-centric (https://landing.ai/data-centric-ai/) approaches to projects over model-centric, which in part re-emphasize the importance of *scrubbing* and maintaining high-quality data.

Explore

> “you can see a lot by looking”
>
> -MASON AND WIGGINS

In OSEMN’s third phase, you are not yet testing any hypotheses or making predictions. Rather, you are exploring the data using various techniques to better understand the data and its story.

There are several data exploration techniques. The ones that Mason and Wiggins call out are:

- Look at the first few rows of your data sets
- Create histograms to visualize distributions
- Produce scatter plots to assess the relationship between variables
- Conduct dimensionality reduction to simplify the data set
- Cluster the data to identify groups of data with similar characteristics

Additionally, you might want to provide descriptive statistics on the data and create dashboards to facilitate ad hoc data exploration for you and your stakeholders. Regardless, typically your objective in this phase should not be to understand every minor detail but rather to get a good sense of the data before proceeding to the following phases.

Model

> “always bad, sometimes ugly”
>
> -MASON AND WIGGINS

The *Model* phase is what data scientists are most famous for.

In this phase, you are trying to find the algorithm that best describes how known input data can predict unknown output values. A simplified set of steps in this process is:

- Identify how you will evaluate the accuracy of a model (identify a “cost” function)
- Select the data fields (known as the “features”) that you will use in your model
- Split your data into different data sets (at least “training” and “validation” data)
- Use the training data to train different algorithms to predict the target unknown value based on the known input features
- Evaluate the performance of the model with the “unseen” validation data
- Select the model that “best” accomplishes the intended modeling goal. This often is the model which minimizes the cost function. However, you should consider other factors such as model training time, model complexity, future data availability, and model explainability.

iNterpret

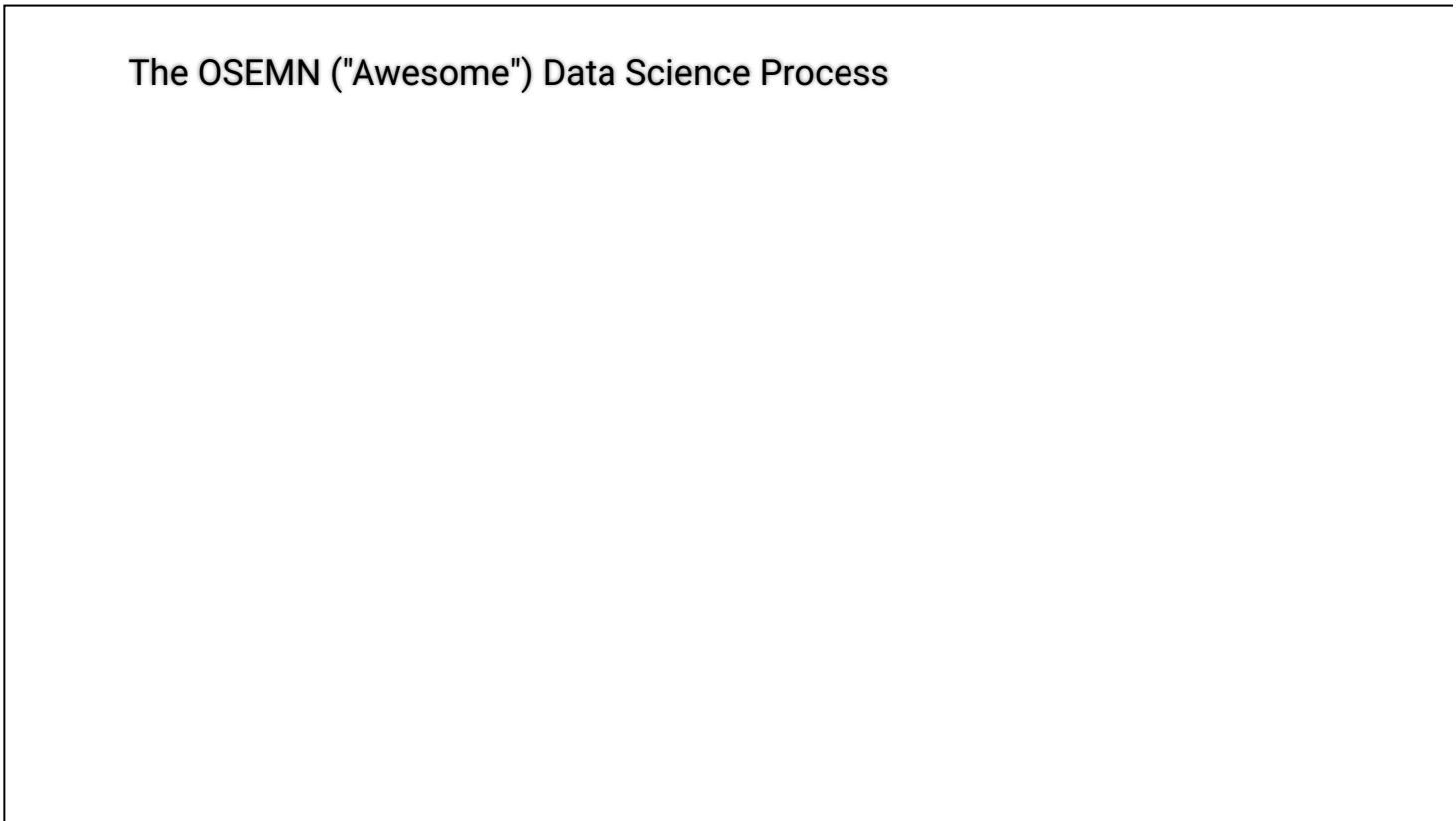
> “The purpose of computing is insight, not numbers.”
>
> -MASON AND WIGGINS

So you’ve obtained your data, scrubbed it, explored it, and modeled it. But what does all this actually mean? What insights did you gain? How can your analysis and model drive real-world outcomes?

In this fifth and final OSEMN phase, you reflect on the questions that prompted the data process to occur in the first place, as well as the actionable value that comes from the process of investigating and modeling the data. The entire point of the process is to offer insight to the organization or stakeholders. If the OSEMN process produces no such thing, then conducting the project was essentially a waste of time and resources.

Moreover, you need to make your analysis understandable and digestible by the interested parties. Therefore, as essential as technical skills are to the process, being able to outline findings in a logical, sensible, actionable presentation to an audience is crucial. Without efficient communication

of the results, such as presenting technical data to a non-technical audience, you undermine the whole project. After all, all that work is worth it only if the results yield something of value for the interested parties.



Hilary Mason reflects on OSEMN in 2022 (12 years after its introduction)

Should you use OSEMN?

Benefits

OSEMN offers many benefits. Specifically, it is...

- Simple – It distills the complex process of a data science project into five clear steps. This is especially noteworthy given that this general process and the modern concept of data science were still new when Mason and Wiggins created OSEMN in 2010.
- Catchy – OSEMN is Awesome!
- Makes sense – The steps presented have a logical flow representative of the general data science life cycle.
- Provides a shared understanding – OSEMN creates a taxonomy to help define how a data science project progresses.

Shortcomings

Although its beauty lies in its simplicity and catchy taxonomy, OSEMN has several shortcomings.

- Misses business understanding – The framework starts with *Obtain* which ignores the key base questions that should come first, namely: "Should I invest time on this project?" and "What outcome am I trying to drive?"
- Doesn't consider deployment – OSEMN implicitly assumes that you are delivering a one-time output. In reality, you often need to deploy a model in a production system so that it continues to provide value over time.
- Ignores teamwork – Data science is increasingly a team sport. Yet, OSEMN ignores the broader team aspect of modern projects.
- It's linear – OSEMN proceeds in a waterfall-like manner (https://www.datascience-pm.com/waterfall/) with each phase following the other. In reality, you often switch back and forth between phases as needed. Moreover, you will want frequent decision points where you re-assess and adjust your plan based on recent learnings.

So...should you use it?

OSEMN is awesome for what it is – A simple, catchy, easy-to-understand representation of the data science life cycle. Thus, it is great as an introduction to data science projects. Indeed, many college courses, online courses, blog posts, and even books use OSEMN to teach newcomers to the field.

However, OSEMN is not sufficient as a framework for most real-world projects. It lacks a lot of key details you need to execute a project. In fact, I added most of the bullets in the process explained above to detail the otherwise very high-level process.

If you explicitly use OSEMN, you should add details on top of the base framework. Moreover, given its lack of consideration for teamwork, combine OSEMN with a collaboration framework such as Kanban (https://www.datascience-pm.com/kanban/), Scrum (https://www.datascience-pm.com/scrum/), or Data Driven Scrum (https://www.datascience-pm.com/data-driven-scrum/).

On the one hand, you probably don't want to set out on a major data science project explicitly using OSEMN (At least I haven't met a data professional who says they directly use OSMEN). On the other hand, you'll implicitly use OSEMN – at least at a conceptual level – as you flow through your projects.

What are some alternatives to OSEMN?

OSEMN is one of the countless data science project life cycles (also known as “frameworks” or “workflows”). Indeed, I come across a “new” life cycle every few months. Most of these life cycles essentially communicate the same thing (namely, the steps you take in a data science project). However, they vary in aspects such as

- Level of detail (OSEMN being one of the higher-level approaches)
- Emphasis of project initialization at the project start and deployment – or even further to operations – at the end (OSEMN being a more myopic framework)
- Inclusiveness of teamwork (which OSEMN ignores)
- Modern infrastructure (which OSEMN ignores)

Alternative frameworks include:

- CRISP-DM (https://www.datascience-pm.com/crisp-dm-2/): The most famous (and perhaps most detailed) framework.
- SEMMA (https://www.datascience-pm.com/semma/): Similar to CRISP-DM but more myopic ignoring Business Understanding and Deployment.
- Microsoft Team Data Science Process (https://www.datascience-pm.com/tdsp/): Combines a base life cycle with a modern Agile collaboration framework.
- Domino Data Labs Life Cycle (https://www.datascience-pm.com/domino-data-science-life-cycle/): Represents a data science project as a flow diagram and includes operational aspects.
- Other workflows (https://www.datascience-pm.com/data-science-workflow/): Each having its own angle on the steps of a data science project.

These frameworks can help make your project life cycles repeatable and consistent and to help ensure quality into your project. Perhaps more important than *which* framework you use, is that you actually *use* one.

OSEMN might not be the best framework for implementing modern data science projects. But it is a great conceptual option if you’re starting out and executing a small project.

Related Posts