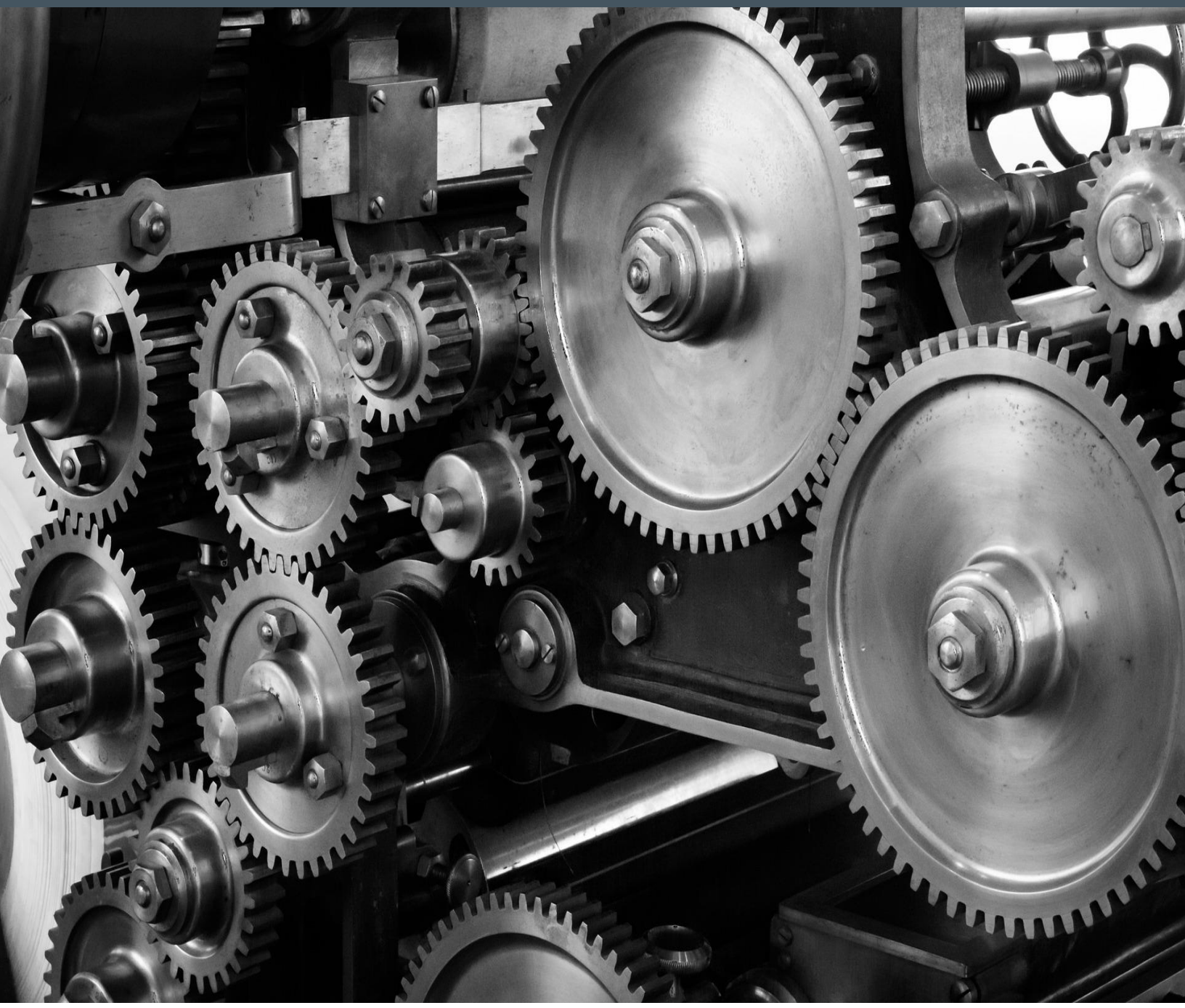
FINANCIAL RISK GROUP

Vol. 1, No. 3 May 2018

# New Machinist Journal

## The New Machinist's Process Flow

Jonathan Leonardelli

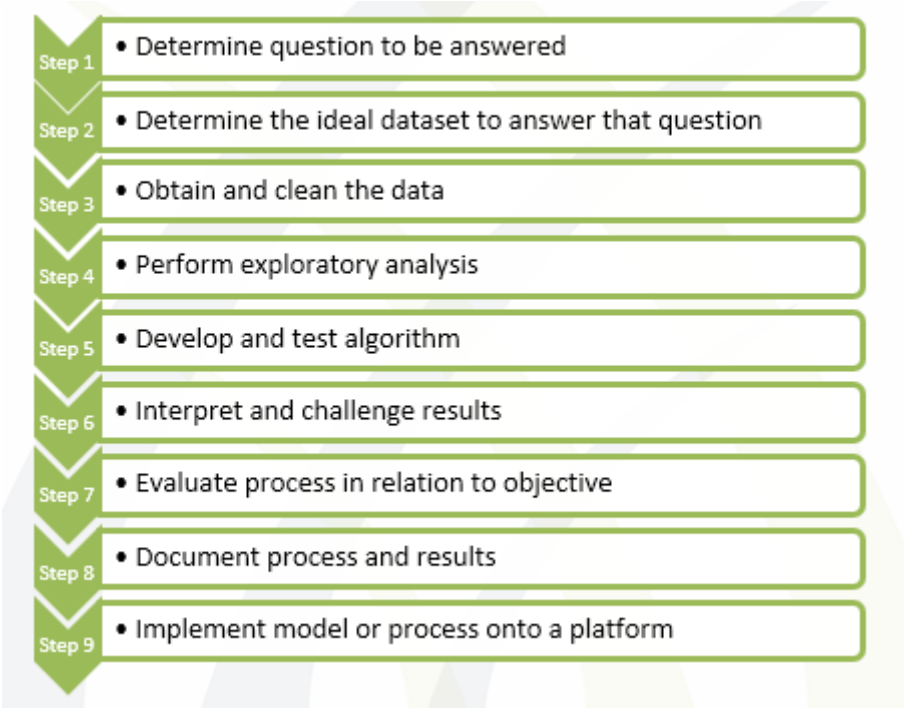


## Machinist (noun): A person who operates a machine, especially a machine tool

A machinist does not start using lathes and grinders without a goal in mind, and an articulated plan or blueprint. And neither should a new machinist. Following several previous issues that discuss some general aspects of machine learning it seems that providing a good point from which to start a new project is necessary. In this issue of the journal we provide a general process flow that a new machinist typically follows when creating a framework that uses machine learning algorithms or artificial intelligence.

### The Process Flow

Successfully creating and developing machine learning algorithms (or AI) to answer a question does not involve jumping directly to the algorithm. Instead, there is a process that should be adhered to in order to ensure the algorithm is doing what it is intended to do. The following image shows the common steps:




### The Steps (in slightly more detail)

#### Step 1 – Determine the question

Of all the steps performed in the machine learning process flow this is the most important one. It guides the remainder of the process – e.g., what data to collect, what algorithms to consider, how and where (if at all) the final algorithm gets executed. An important point to note is the lack of plurality: it is best to keep this to one question.

### Step 2 – Determine the ideal dataset

After defining the question, it is important to determine the ideal dataset. This is the dataset that is believed to answer the question posed in Step 1. It is "ideal" because typically one is unable to obtain all the data needed to satisfy the analysis.

### Step 3 – Obtain and clean the data

Given the question and the data guidelines presented from Step 2, then next step involves collecting the data. Most of the time the data will come from both structure and unstructured sources[1].

After collecting the data, it should be passed through a cleaning process to ensure the data can be used (e.g., put it into a structured format). At this stage of the process one should begin documenting changes made to the data. This is done to ensure anyone can pick up the raw data, perform the same cleaning steps, and arrive at the exact same data. This is much needed step for having a process that is repeatable.

### Step 4 – Perform exploratory analysis

During this step the data is investigated. For example, some of the common items to look at are (this list is not exhaustive):

- Whether outliers exist in the data and how they impact skewness of a variable's distribution.
- The predominance of missing data.
- Identifying relationships between variables that may indicate collinearity or be unexpected.

At the end of this process the goal is to determine if there is enough data that can be used to support or refute the hypothesis, or build a predictive model, based on the question defined in Step1.

### Step 5 – Develop and test algorithm

This is the step that is often imagined when one thinks of building a machine learning process or using artificial intelligence. If the goal of the new machinist is to test a hypothesis which does not require the development of a predictive model then this step is omitted. Otherwise, this is the step where the machine learning algorithm is built based on the available data. This step typically involves:

- Splitting data into training and testing datasets.
- Estimating parameters of models and evaluating performance on in-sample and out-of-sample data.
- Refining estimates as needed and retest.

### Step 6 – Interpret and challenge results

During this step it is best to get others involved, whether it is using a client's subject matter experts or presenting it in-house for evaluation. This step can be highly iterative and can lead to returning to Step 5. The main purpose of this step is to ensure that the model appropriately addresses the question posed.

### Step 7 – Evaluate process in relation to objective

At this step it is best to pause and assess current progress. Confirm whether the business issue/questions articulated is being addressed. Do the findings warrant implementation or should one go back to the drawing board? This can be a challenging step from the standpoint that many people do not want to look back on the work they have done and consider it wasted effort. A strongly held belief is that even work that

[1] Unstructured data is data that does not fall into a prescribed data model (e.g., think PDF files or tweets).

is not implemented is not considered wasted. There is value in knowing, for example, that the question posed cannot be adequately answered with the data available.

### Step 8 – Document process and results

The penultimate step involves documenting the process and results from the previous steps. Documentation goes beyond just ensuring the key points about the process are documented. It should also ensure that the code used to reach any of the conclusions is appropriately notated and turnkey. The requirement of it being turnkey is essential for having it be repeatable for other developers, model validation, auditors, and, perhaps, future researchers. Because of the importance of decisions being based off these results, it is critical that anyone can reproduce the exact findings.

### Step 9 – Implement model or process on platform

This is an important step that can often be overlooked when considering the entire process flow. Results from the above steps typically fall into one of two groups: either they are used to guide a decision-making process or used for predictive purposes. In the case of the latter usage, often the models need to be placed onto an existing (or new) infrastructure for execution. Transitioning development code to production-ready code, and testing its accuracy, is the final step in the machine learning framework we use.

### Conclusion

The steps outlined in this volume do not ensure that a predictive algorithm can always be built (e.g., it is possible the data that is available does not lead to a highly predictive model) or the answer to the question aligns to one's wishes. Nor does it imply that this is the only framework that might be employed. What the steps do offer is a means to thoroughly execute the process to ensure the best analysis was provided. Additionally, it offers a basis from which to work from that can be adjusted as experience is acquired, new technologies are developed, and/or environmental factors dictate change. It is important to have a plan or blueprint to insure that all of the known necessary steps are taken in the process.

For more information on FRG's research in the areas of Artificial Intelligence and Machine Learning, please visit the **website** or contact the FRG Research Institute at **Research@frgrisk.com**.