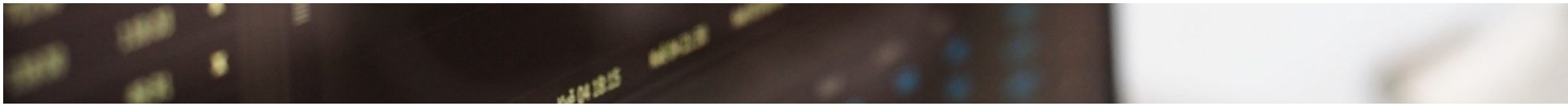
Manav Sehgal Follow
Data Science and Machine Learning Expert
Dec 21, 2016 · 3 min read

7 Stages For A Reliable Data Science Solutions Workflow


Bid
Ask
Auto
GBPUSD M15
1.4505
1.00
1.4508
SL/TP



This data science solutions workflow provides a repeatable, robust, and reliable framework to apply the right-fit workflows, strategies, tools, APIs, and domain for your data science projects.

The new book Data Science Solutions explains the seven stages of this workflow with real-world projects, code, and tools pipeline. Sample chapter available online.

Question. Problem. Solution.

Before starting a data science project we must ask relevant questions specific to our project domain and datasets. We may answer or solve these during the course of our project. Think of these questions-solutions as the key requirements for our data science project. Here are some templates that can be used to frame questions for our data science projects.

- Can we classify an entity based on given features if our data science model is trained on certain number of samples with similar features related to specific classes? Example: Image recognition.
- Do the samples, in a given dataset, cluster in specific classes based on similar or correlated features? Example: Customer segmentation.

- Can our machine learning model recognise and classify new inputs based on prior training on a sample of similar inputs? Example: Spam detection.
- Can we analyse the sentiment of a given sample? Example: Twitter stream sentiment analysis.

Acquire. Search. Create.

This stage involves data acquisition strategies including searching for datasets on popular data sources or internally within your organisation. We may also create a dataset based on external or internal data sources.

The acquire stage may also include data scraping and crawling to create new datasets from online content. Preparation of data catalogs and metadata definition about your project data starts during acquire stage.

The acquire stage may feedback to the question stage, refining our problem and solution definition based on the constraints and characteristics of the acquired datasets.

Wrangle. Prepare. Cleanse.

The data wrangle stage prepares and cleanses our datasets for our project goals. This workflow stage starts by importing a dataset, exploring the dataset for its features and available samples, preparing the dataset using appropriate data types and data structures, and optionally cleansing the data set for creating model training and solution testing samples.

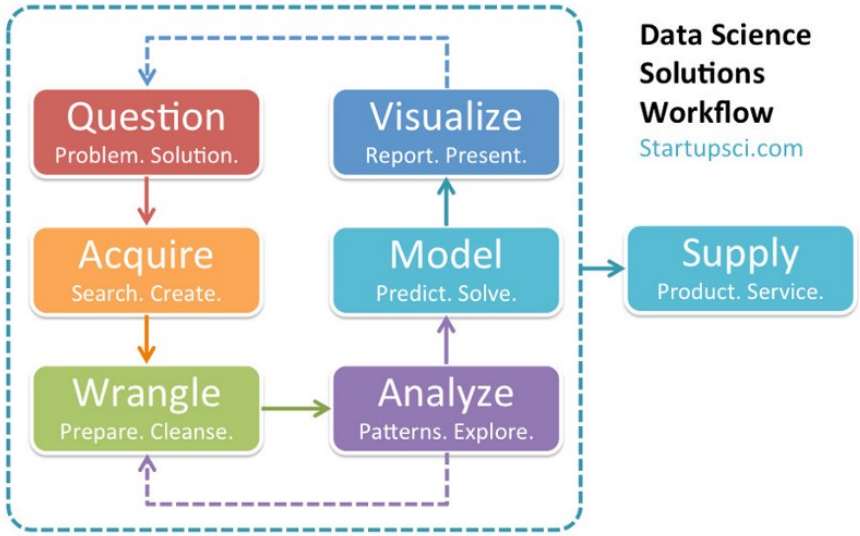
Preparation of knowledge graphs or rules and heuristics for your target domain and data starts during wrangle stage.

The wrangle stage may circle back to the acquire stage to identify complementary datasets to combine and complete the existing dataset.

Analyze. Patterns. Explore.

The analyze stage explores the given datasets to determine patterns, correlations, classification, and nature of the dataset. This helps determine choice of model algorithms and strategies that may work best on the dataset.

The analyze stage may also visualize the dataset to determine such patterns.



Model. Predict. Solve.

The model stage uses prediction and solution algorithms to train on a given dataset and apply this training to solve for a given problem. Most data science projects may not create new algorithms.

Many projects will reuse well established libraries including Python Scikit-learn and frameworks like Google Tensorflow. Models can also build on data science and machine learning APIs offered by IBM Watson, Google, Amazon, and Microsoft.

Visualize. Report. Present.

The visualization stage can help data wrangling, analysis, and modelling stages. Data can be visualized using charts and plots suiting the characteristics of the dataset and the desired results.

Visualization stage may help as early as the analysis stage also provide the outputs required for the supply stage.

Supply. Products. Services.

Once we are ready to monetise our data science solution or derive further return on investment from our projects, we need to think about distribution and data supply chain. This stage circles back to the acquisition stage. In fact we are acquiring data from someone else's data supply chain.