What is Representation Learning?

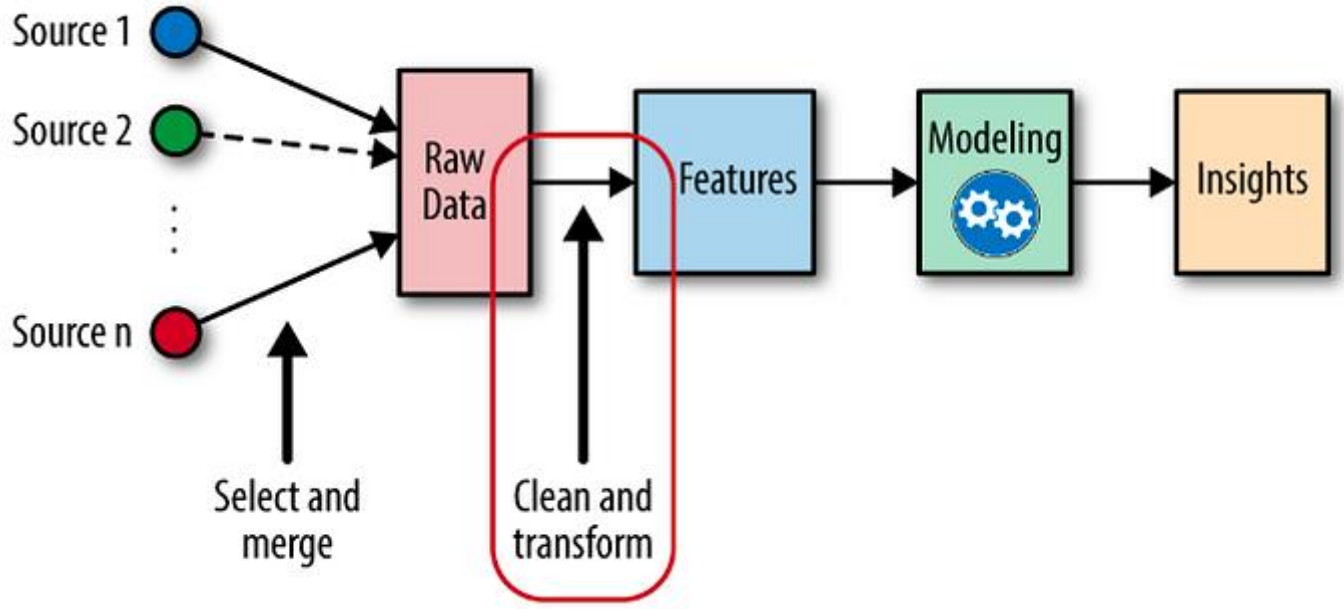
Representation learning is a method of training a machine learning model to discover and learn the most useful representations of input data automatically. These representations, often known as “features”, are internal states of the model that effectively summarize the input data, which help the algorithm to understand the underlying patters of this data better.

Representation learning marks a significant shift from traditional, manual feature engineering, instead entrusting the model to automatically distill complex and abundant input data into simpler, meaningful forms. This approach particularly shines with intricate data types like images or text, where manually identifying relevant features becomes overwhelming. By autonomously identifying and encoding these patterns, the model simplifies the data and ensures that the essential information is kept. So summarised, representation learning offers a way for machines to autonomously grasp and condense the information stored in large datasets, making the steps in machine learning that follow after more informed and efficient.

From Hand-Crafted to Automated: The Shift in Feature Engineering

In the early days of machine learning, feature engineering was akin to sculpting by hand. It required engineers to manually identify, extract, and craft features from

raw data, a process heavily reliant on domain expertise and intuition. Imagine trying to predict car prices. Beyond the obvious features like make, model, and mileage, one would have to speculate: Does the color matter? Or the month of sale? The process was tedious and often restrictive, bounded by the limits of human insights.



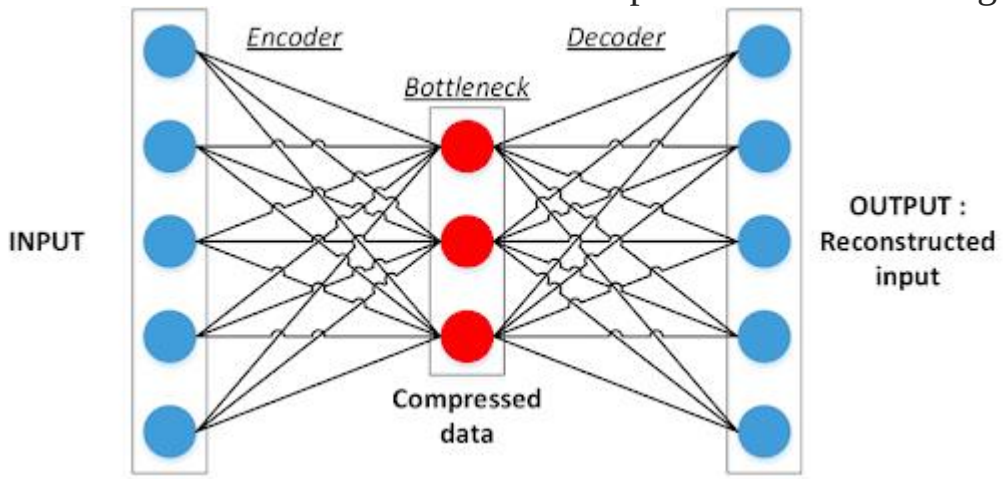
Feature engineering in a machine learning pipeline.

Then came representation learning, a revolutionary approach that lets the model learn what the most informative features are. It is possible to do this without a concrete task in mind ("what is a car") or tailored to a specific task ("price probably will matter"). So, while traditional feature engineering laid the foundation, representation learning streamlines and deepens our exploration of data, showing a new era of efficiency and adaptability.

Self-Supervised Learning: Autoencoders and the Embedding

Space

Self-supervised learning, a subset of unsupervised learning, is a powerful way to learn representations of data. Among the popular approaches in this category is the autoencoder. An autoencoder is a type of neural network that learns to encode input data into a lower-dimensional, and thus more compact, form. The network then uses this encoded form to reconstruct the original input. The encoding process discovers and extracts essential features in the data, while the decoding process ensures that the extracted features are representative of the original data.



Representation of an autoencoder's architecture.

An essential concept to grasp in self-supervised learning is the idea of an embedding space. This space represents the features or characteristics the autoencoder (or any other self-supervised model) has learned. In an effectively trained model, similar data instances will be close to each other in this space, forming clusters. For example, a model trained on a dataset of images might form distinct clusters for different categories of images, like birds, clothing, or food. The

distance and direction between these clusters can often provide valuable insights into the relationships between different categories of data.