Thesis Proposal:

The Role of Embeddings in Data-Driven Augmentation

Federico Dassereto

September 28, 2020

Abstract

A-o meu neuo gh'é neue nae neue: a ciù neua de neue nae neue a n'eu anâ.

1 Motivation and Description

In data science, it is increasingly the case that the main challenge is not in integrating known data, rather it is in finding the right data to solve a given data science problem. Today, data is a mass (uncountable) noun like dust, and data surrounds us like dust, even lovely structured data. Data is so cheap and easy to obtain that it is no longer important to always get the integration right and integrations are not static things. Data integration research has embraced and prospered by using approximation and machine learning. The uncontrolled nature of data manifests in large repositories of data (data lakes), in which both structured and unstructured data are stored. The peculiarity of data lakes lies in the fact that there is uncertainty about the presence of metadata describing the data themself. Furthermore, it is common the situation in which there is a lack of schemas, making traditional database approaches to integrating or querying data difficult to pursuit or even infeasible. Along with the uncertainty regarding the quality of the data, data Volume makes it infeasible the traditional human-in-the-loop framework, since hand labeling or manual rating of very large amounts of data is extremely expensive.

2 Reference Area and Relevance of Goals

3 State of the Art

4 Goals and Preliminaries

5 Research Plan

Acknowledgements

So long, and thanks for all the fish.