# Introduction

## Overview

* Analysis of research papers can give a lot of insights about software resources and their dependency.
* In a scientific research different kinds of input resources are used. One of such input is a software.
* Used resources in a research are typically mentioned in a citation. Citation practices of formal articles in a research are matured and various citation styles exist. Even if principles for formal citation of a software has already been put out, most scientists are not properly citing resources.
* Surprisingly, sometimes researchers do not mention the type of software they used entirely or mention it with a rather vague abbreviation and just talk about the results they have obtained.
* As long as software is mentioned using formal methods, like RRID, it is possible to perform citation analysis using regular expressions which can be constructed to capture the pattern of citation.
* Though regular expression based analysis can give basic insights about the software citation it has limitations because:
  + Not so many authors use formal citation of software, like RRIDs
  + Even if scientists use formal citations, they may fail to properly follow the guidelines. For example, some authors tend to ignore the RRID-part and that creates an ambiguity by it self that it is not possible to know weather the author is actually making a software citation or it is completely something else.
  + Rule based method fails to capture context information and ignores dependencies. It is not possible to be sure about the authors intention whether or not using a software citation.
* At the same time pattern based analysis, like using regX, is not suitable to extract information about software citation, for instance the particular use of a software, especially when a software mention statement lacks any form of formality where the information is concealed in a natural language description.
* Therefore it is required to automatically extract the purpose of software use in scientific literatures. This might help to answer questions like:
  + What type of software is being frequently used for what purpose in a specific area of research? This also allows to find an answer further question like what is the most common technique researchers follow when trying to solve a given research problem in a given domain )
* Previous attempts to automatically extract information using machine learning techniques, specifically supervised machine learning technique, about the software use purpose was constrained mainly because of lack of ground truth data. But this time, with the advent of SoMeSci, it is possible to do so.

## Problem statement

## Objectives of the research

This work has the following objectives:

* List down the purpose of software usage in a research in a hierarchical manner.
* To extend SoMeSci with a manual annotation of purpose of software usage.
* To find a classifier and optimize results.

# Literature review

## The role of Software in Scientific research

### Introduction

Modern research is unthinkable without a use of software and scientific investigations in various areas of science are becoming increasingly reliant on software tools [1, 3 , 4, 14].

A software is very important asset for building a scientific knowledge and more discoveries in a research are made possible than ever by a use of software tools that automate processing of huge amount of data [9]. Typically a software is used in a research for control processes, simulation, modelling, data analysis, knowledge dissemination, etc. [2][25].

Since software is not often considered as an academic output [24][25], it is usually not cited in research papers across several fields of research [25]. To counteract against this culture, a task force that advocates about the role of software in a research, known as Research software Alliance (ReSA), has been established. The ReSa promotes the inclusion of software as a primary research out put, influences decision makers to value a research software and give credits to the developers [21]. In 2019, the task force has collected literatures, at [Zetoro group libraray](https://www.zotero.org/groups/2400609/resa/library), that evident significant roles of software in a research [22].

Scientific software is often complicated and requires specialized domain knowledge for its development [3]. Due to this increasing number of scientists are developing a software as part of their research work or directly taking part in the development process of a research software [9, 15]. This fact is clearly reinforced by a survey result from 2008 [1, 2, 6] which indicates that:

* nearly 45% of scientists spend more time developing a software as part of their research work than five years ago.
* 38% of researchers spend at least 20% of their time developing a software.
* over 90% of scientists agree that software is important for their research and
* nearly 70% claim that their research directly depends on a use of a software.

### General roles of software in a research

Software is playing various crucial roles in a research and making a shift in a research culture. For example, software tools are making most of research to be increasingly data driven i.e. insights from an in-depth analysis of large volume of data-sets form the basis of a research conclusion [1][16].

Some of the most general roles of a software in a research are:

* Software helps to explore und understand a research problem [2].
* Results from a scientific software is presented as an evidence to support a research result [15].
* *A software dictates the quality of a research outcome*[2] [23]. Outcome of a research becomes unreliable or even useless if there is an error in the software [5]. For example, several scientists retracted their scientific publications up on a retrospective discovery of a bug in their software [3, 6, 8]. A more palpable failure of a research ambition due to an error in the control-system software, for instance, is the failure of *Ariane rocket* in 1996 [17].
* A software also helps to document a research process and to *validate results of a given research* [16]. Executable cells in a Jupyter notebook is one real world example where a software can be used to validate a research result.

### Examples of software role in specific domains of research

A software is being extensively used for a research in various areas of science such as physics, chemistry, space science, life science and so on.

The physics research facility, the Large Hydron Collider at CERN, for instance uses a software with more than 5 million lines of code which is used for processing of terabytes of data generated from experiments [4].

In a nuclear research, a software is being developed increasingly to be used for experiments [19]. For example, testing a modification in a nuclear weapon can not be field tested, but instead a software that simulate the impact of modification is usually used [15]. This is because of regulations like nuclear test ban treaties and the potential disaster, to the environment and life, associated with nuclear weapons [20].

In chemistry research, a software can be used to model and simulate chemical processes that are challenging, too complex or expensive to conduct in reality. Karplus and Levitt used computer simulations for their joint-research “the development of multi-scale models for complex chemical systems” and won a Nobel prize in 2013 for their work [4, 18].

In a climate and environmental studies, software is used to make predictions about climate changes. For example a historical temperature data can be integrated to make predictions about future temperature variations [4].

In a space science, space probes heavily rely on software. In this case a software navigates space crafts to other planets, processes and transmits scientific data back to Earth fur more processing, helps researchers interpret results, etc[28].

### The role of software in research breakthroughs

A use of software also allowed to produces better scientific discoveries and several research breakthroughs has been made possible[1].

One of the research breakthroughs is creation of the very first visual representation of a black hole using an open source software NumFOCUS. To observe a black hole that is 55 million light years away, it would have required to build a huge telescope of size of planet earth. But instead of building one giant telescope, hundreds of scientists spent decades of years creating a global network of telescopes, known as Event Horizon Telescope (EHT) [29], synchronized precisely using atomic clocks. The EHT gathered a huge amount of data for years. However there was a lot of noise in data the collected data because :

* The EHT was a network of non-similar telescopes.
* The radio signals were coming through attenuated due to atmospheric effect like water vapor, clouds, turbulence … etc.

Therefore the scientists had to use various algorithms and data analysis pipelines. The resulting image from various data processing was compared to ensure the integrity of the result. This huge scientific breakthrough in a space research, can be attributed to mainly the use of powerful data processing software.

Other scientific breakthroughs that can be attributed to role of software in a research include:

* The detection and visualization of gravitational waves for the first time, using a LIGO software [30][31].
* Software accelerates drug discovery [32].

## Literature review on classification of software usage purpose

### Introduction

In a research various scientific laws are used to design software algorithms and then a software is created by implementing those algorithms. Often times there are scientific methods endemic to only some areas of science. In this case a more domain specific research software would be the end result.

* Sometimes scientists use several software together in their research[12].

This because that each software has one or more purpose that serves to explore the research problem and help to find an answer. Example: a scientist might use a data analysis software together with modelling software.

### Software usage purposes in a research

In a research a scientific software is used for various purposes. In a data driven science, one of the most important software use case is for Data analysis. Usually a huge amount of data is analyzed using mathematical or statistical methods. Further more, domain specific data analysis techniques exist.

The data to be analyzed using a scientific software also has a broader range. Data could be something obtained from a sensor, an image retained from a microscope, a data generated from a random walk, data generated from simulation of a model. … etc

* A software can be used for various purposes during a research. The main purposes of use of software in a research are:
  + Data Analysis
    - Data visualization
    - Mathematical analysis
      * Numerical Analysis
      * Statistical Analysis
    - Domain specific analysis
      * Densitometric Analysis
      * Voxel-based Analysis
  + Data Processing
  + Data Mining – extraction / discovery of patterns in large data sets using ML, Statistics and Data.
  + Data Collection
    - [Automatic indexing](https://en.wikipedia.org/wiki/Automatic_indexing)
    - [Web Crawler](https://en.wikipedia.org/wiki/Web_crawler)
  + Simulation
    - [Physics engine](https://en.wikipedia.org/wiki/Physics_engine)
  + Modelling
    - [Graphics software](https://en.wikipedia.org/wiki/Graphics_software)
    - animation software

For example: to perform Data analysis (might refer to inspecting, cleaning, transformation, and modelling a data). The main purpose of data analysis is to extract meaningful information from a data that will help to make businesses operate more effectively [13].

### Categories of research software

* System software
  + [OS](https://en.wikipedia.org/wiki/Operating_system)
    - [Firmware](https://en.wikipedia.org/wiki/Firmware)
    - [Middleware](https://en.wikipedia.org/wiki/Middleware)
  + Drivers
* [Application software](https://en.wikipedia.org/wiki/Application_software)
  + Web browser
  + [Word processor](https://en.wikipedia.org/wiki/Word_processor)
  + [IDE](https://en.wikipedia.org/wiki/Integrated_development_environment)
    - Compiler
  + [CASE Tools](https://en.wikipedia.org/wiki/Computer-aided_software_engineering)
  + [Scorewriter](https://en.wikipedia.org/wiki/Scorewriter)
  + Text Editor
  + Social software
    - Email
    - Social ntk apps
* Programming software
* Science software
  + [genealogy](https://en.wikipedia.org/wiki/Genealogy) software

### Taxonomy of software purpose

Taxonomy based on [licensing](https://en.wikipedia.org/wiki/Software_license) and distribution

* Commercial software
  + [Free](https://en.wikipedia.org/wiki/Free_software) software
  + [Proprietary](https://en.wikipedia.org/wiki/Proprietary_software) software
    - Freeware
  + Open-source

### Other types of software

* [Business software](https://en.wikipedia.org/wiki/Business_software)
  + project management software
  + accounting software
  + Banking software
  + [Decision-making software](https://en.wikipedia.org/wiki/Decision-making_software)
* [Web server](https://en.wikipedia.org/wiki/Web_server)
* Packet Analyzer
* Antivirus software
* [Content management system](https://en.wikipedia.org/wiki/Content_management_system)
* Computer Program
  + Software [Library](https://en.wikipedia.org/wiki/Library_(computing))
  + Spell checker
  + Autoresponder
  + Computer virus
    - Malware
    - Spyware
    - Rootkit

### Systems

* [Integrated library system](https://en.wikipedia.org/wiki/Integrated_library_system)
* database management system
* Network security system
  + Firewall system