Contents

[1 Introduction 3](#_Toc90905373)

[1.1 Overview 3](#_Toc90905374)

[1.2 Problem statement 4](#_Toc90905375)

[1.3 Objectives of the research 4](#_Toc90905376)

[1.4 Summary 4](#_Toc90905377)

[2 The role of Software in Scientific research 5](#_Toc90905378)

[2.1 Introduction 5](#_Toc90905379)

[2.2 General roles of software in a research 6](#_Toc90905380)

[2.3 Domain specific examples 7](#_Toc90905381)

[2.4 The role of software in research breakthroughs 8](#_Toc90905382)

[2.4.1 Breakthrough - Visual Representation of a black hole 8](#_Toc90905383)

[2.4.2 Breakthrough - Visualization of gravitational waves 8](#_Toc90905384)

[2.4.3 Breakthrough - accelerated drug discovery 9](#_Toc90905385)

[3 Software usage purpose 10](#_Toc90905386)

[3.1 Brief History of Computers and software 10](#_Toc90905387)

[3.2 Research and Software Purpose 10](#_Toc90905388)

[3.3 Identification of software purposes from literatures 11](#_Toc90905389)

[3.3.1 Comparisons 11](#_Toc90905390)

[3.4 Identification of software purposes from ontologies 11](#_Toc90905391)

[3.4.1 WikiData 11](#_Toc90905392)

[3.4.2 The software ontology (SWO) 12](#_Toc90905393)

[3.4.3 OntoSoft 12](#_Toc90905394)

[3.4.4 Other sources 13](#_Toc90905395)

[3.5 Types of software purposes 13](#_Toc90905396)

[3.5.1 Data collection 13](#_Toc90905397)

[3.5.2 Data pre-processing 13](#_Toc90905398)

[3.5.3 Data Analysis 14](#_Toc90905399)

[3.5.4 Data visualization 15](#_Toc90905400)

[3.5.5 Simulation 15](#_Toc90905401)

[3.5.6 Stimulation 16](#_Toc90905402)

[3.5.7 Modelling 16](#_Toc90905403)

[3.5.8 Programing 16](#_Toc90905404)

[4 Annotation of Data 17](#_Toc90905405)

[4.1 Data set 17](#_Toc90905406)

[4.2 BRAT Annotation tool 17](#_Toc90905407)

[5 Feature selection 18](#_Toc90905408)

[5.1 Phrase 18](#_Toc90905409)

[5.2 Sentence 18](#_Toc90905410)

[5.3 Paragraph 18](#_Toc90905411)

[5.4 Dependency tree 18](#_Toc90905412)

[6 Selection of classifiers 19](#_Toc90905413)

[7 Classification and Evaluation 20](#_Toc90905414)

[8 Summary of Results and Conclusion 21](#_Toc90905415)

[8.1 Summary of Results 21](#_Toc90905416)

[8.2 Conclusion 21](#_Toc90905417)

[8.3 Limitations and future work 21](#_Toc90905418)

# Introduction

## Overview

**Why do software purpose analysis ?**

**What resources have been used?**

**What objectives are set?**

**What is the structure of the work?**

* 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 select feature for the training model
* To select a classifier and train a model.
* To evaluate and optimize results

## Summary

# 

# The role of Software in Scientific research

## Introduction

Nowadays scientific research is unthinkable without a use of software and investigations in various areas of science are becoming increasingly reliant on software tools {goble2014better, storer2017bridging, hannay2009scientists, jimenez2017four}.

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 {jimenez2017four}. Typically a software is used in a research for data processing such as data analysis, modeling, simulation, control processes, knowledge dissemination, etc. {hannay2009scientists, pan2016disciplinary}.

In modern research, a scientific software is as important as any lab-equipment {wilson2014best}. However, the development of scientific software is much more complicated and fundamentally different from an ordinary commercial software like accounting software. Scientific software requires specialized domain knowledge for its development and require direct involvement of domain expert or scientist {wilson2014best, segal2008developing}. Due to this, an 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 {jimenez2017four, kanewala2014testing}.

According to surveys conducted in the UK and USA, 2008 and 2017 respectively, most scientists agree that software plays an important role in their research work {hettrick2014uk, nangia2017track}. Participants of the survey, in UK, were 2000 researchers working in various areas of science in roles ranging from student to senior academic staff whereas participants of the survey, in USA, were members of the US National Postdoctoral Association.

The results from of UK survey {hettrick2014uk} indicate that :

* 38% of researchers spend at least 20% of their time developing a software.
* Almost half of scientists spend more time creating software as part of their research work than five years ago .
* Over 50% of survey respondents reported that they develop their own software.
* Over 90% of scientists say software is important for their research &
* Nearly 70% claim that their research-work directly depends on use of a software.

The results from of USA survey {nangia2017track} indicate that:

* Over 90% of scientists use software.
* 63% of respondents state that their research is impossible with out using software.
* 31% of scientists say that they could do their work without using a software but more effort would require.
* Only 6% of survey respondents say that there would be no significant difference in their task if they do not use software.

Overall, results from the two surveys clearly indicate that software is pervasive in scientific investigations and many researchers use as well as develop a software for their research.

Even though software plays an important role in a modern research, usually the contributions of software is understated. This can be seen from the poor citation practice of software in research papers across several fields of research {yang2018important, pan2016disciplinary}. In attempt to promote the recognition of the roles of scientific software in a research, the ReSA has collected literatures that evident the roles software play in a research, at Zetoro group library. The main aim of ReSA is to influences decision makers to properly attribute contributions of a research software and give credits to its developers.

The next section presents more details about the role of software in general, in specific domains, and some examples of research breakthroughs.

## General roles of software in a research

Software is playing crucial roles in a research and making a shift in a research culture in terms of enabling automation of analysis pipelines, creation of new ways of analysis via computational models, supporting sophisticated analysis of large volume of data, documentation of a research, etc. {jay2020software}.

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

* Software helps to explore und understand a research problem {hannay2009scientists}.
* Results from a scientific software is presented as an evidence to support a research conclusion {kanewala2014testing}.
* *A software dictates the quality of a research outcome* {hannay2009scientists}. Outcome of a research becomes unreliable or even useless if there is an error in the software {soergel2014rampant}. For example, several scientists retracted their scientific publications up on a retrospective discovery of a bug in their software {wilson2014best,merali2010computational,miller2006scientist}. A more palpable failure of a research ambition due to an error in the software, for instance, is the failure of *Ariane rocket* in 1996 {enwiki:1054482061}.
* A software also helps to document a research process and to *validate results of a given research* {jay2020software}. Executable cells in a Jupyter notebook is one real world example where a software can be used to validate a research result.
* Software allows experiments to be made beyond constrains of the physical world. This is because experiments that run on a computer are not limited by processes that occur in nature but only by the laws imbedded in the computer code {wolfram1984computer}.

## Domain specific examples

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 {storer2017bridging}.

In a nuclear research, a software is being developed increasingly to be used for experiments {yan2017case}. For example, testing a modification in a nuclear weapon can not be done on filed, but instead a software that simulate the impact of modification is usually used {kanewala2014testing}. This is because of regulations like nuclear test ban treaties and the potential disaster, to the environment and life, associated with nuclear weapons {enwiki:1053274189}.

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 {storer2017bridging, andre2014nobel}.

In a climate and environmental studies, software is used to make predictions about climate changes. For example historical temperature data can be integrated to make predictions about future temperature variations {storer2017bridging}.

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 for further processing, helps researchers interpret results, etc.{lutz2011software}.

In medical research and diagnosis, imaging software plays a critical role to assist medical researchers for early isolation of cancer. The main reason for low chance of survival from cancer is mainly due to late detection of cancer cells in the body and once cancer spreads throughout the body it is difficult to treat. This makes a diagnosis of cancer to be a time critical task and early identification of cancer implies curability of a disease and a higher chance of survival {wagner2004challenges}. Especially on the early stages, it is not straight forward to determine which cells are likely to develop a cancer. For this reason, medical scientists use different types of software to identify cancer cell or to decide weather a tumor is malignant or not. Using a software, they could perform various kinds of analysis and processing on imageries obtained from scans such as MRI or CT Scan {al2012lung}. An example of software that is used for cancer imaging research is DMRI. Such software is extensively used by many researchers, more than 75,000 downloads every year {norton2017slicerdmri}. Therefore, it is clear that software plays a critical role in medicine, to diagnose diseases and ultimately to save life.

Software plays an important role in power system planning and operation as well. One of the major activities in power system operation is contingency analysis. During contingency analysis, engineers determine violations of power grid operation conditions, such as overloading, which might occur when outage of a transmission line or a power generation unit occurs. Contingency analysis helps to understand power system behavior after outages and gives an opportunity to take preventative actions {mishra2012contingency}. Power grids are extremely complex and such kind of analysis tasks are unimaginable with out a use of software. An example of software that is used to perform contingency analysis in the power system operation is Power World software {powerworld.com}.

## The role of software in research breakthroughs

The use of software enabled scientists to produce better scientific discoveries and achieve research breakthroughs {goble2014better}. Although it is not possible to list down all the research breakthroughs, some examples have been presented as follows.

### Breakthrough - Visual Representation of a black hole

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, which is not possible any way, hundreds of scientists spent decades of years creating a global network of telescopes, known as Event Horizon Telescope (EHT) {enwiki:1052167868}, synchronized precisely using atomic clocks. The EHT gathered a huge amount of data for years. However there was a lot of noise in 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.

### Breakthrough - Visualization of gravitational waves

The other scientific breakthrough that can be attributed to role of software in a research is the detection and visualization of gravitational waves for the first time, using a LIGO software {enwiki:1047100294, mukherji2017report}.

### Breakthrough - accelerated drug discovery

Software accelerates drug discovery {bhati2021pandemic}.

# Software usage purpose

## Brief History of Computers and software

If we say that computers are for computation, then an idea of using artifacts for computation goes back to 2400 B.C. to the ancient Babylon where abacus was used for arithmetic tasks. From the 11th century and onwards, such computation artifacts evolved to became mechanical devices and were used for solving problems and to perform studies, for instance, astronomy. Though so many artifacts have been developed, it was on early 19th century that the first analytical general purpose *“mechanical computer”* was made by Charles Babbage whom is considered to be the “*father of computer*”. Parts of the machine include ALU, control flow with conditional branching and loops, and even integrated memory. The machine was programmable via data provided in the form of punched cards. Outputs of the machine, analogous to the modern computers printer and speaker, were curve plotter and a bell respectively. After that computers transitioned from mechanical to being electromechanical and soon later to purely electronic devices.

In 1936, the first modern computer known as a universal Turning Machine was proposed and brought new concepts like a computer program ( execution of instructions) and a concept of memory where the instructions are stored. Then sooner transistors and then integrated circuits have emerged giving rise to a modern computer that we use today.

From the history of computers, it is clear that computers have been used for reckoning and computation from day one. However, memory and data storage capability of computers emerged on the later part of the computer history. Thus, any modern computer can be considered as a data processing and storage device.

Based on a context that “computers are data processing and storage” devices, the functionality of computers as data processing machines can be referred to software where as storage of data implies to hardware functionality.

## Research and Software Purpose

In scientific investigations broad range of software is in use. In terms of size, software ranges from simple scripts to extremely complex software with millions of lines of code. In terms task, a software is used for execution of rudimentary tasks to computation of extremely complex ones.

Scientists also use several software together in their research [12]. This is because that each software provides a set of functionalities and has one or more purpose. Example: a scientist might use a data analysis software together with modelling software.

Typical software purpose of software use includes simulations, computational modelling, data manipulation to drive interesting insights (goble2014better).

To be able to classify a software usage purpose in scientific publications, first the list of most important software usage purposes has been identified. To list out the most important software usage purposes, scientific publications in SoMeSci data set has been analyzed to identify the list of all possible software purposes. In addition, possible software usage purposes has also been sought from software ontologies such as WikiData, DBPedia, Ontosoft, …etc.

This section presents, how various software usage purposes has been identified and lists down the most important software usage purposes. After the list of important software purposes has been identified, it is used to annotate a SoMeSci data set using BRAT annotation tool to use it for fitting a model that will automatically identify purpose of a software from context.

## Identification of software purposes from literatures

In a research work, scientists follow scientific method to discover knowledge. Typically, scientists begin with a question and attempt to answer questions through a research and propose hypothetical answers for their questions. Then, they test the proposed hypothesis by conducting various experiments. Although all scientists do not follow the exact same step, the over all idea remains the same {enwiki:1061107378- Scientific method}.

Analysis of literatures aimed at answering a question of “for what purpose scientists are using a software” in their research work. To answer that question, subsets of literature from SoMeSci data set has been manually analyzed and some software purposes has been identified as follows.

### Comparisons

In a research scientists often compare experimental groups.

## Identification of software purposes from ontologies

A given software can be used for various purposes. To list down a software usage purposes, software ontologies has been analyzed and the results from each ontology has been consolidated to create a list of possible software usage purposes. The ontologies, that has been considered in this work are WikiData, DBpedia, SWO - the software ontology, and OntoSoft.

### WikiData

Wikidata is a multilingual knowledge graph that is curated collaboratively by a Wikimedia community{enwiki:1060408581}. It was created by Wikimedia foundation and serves as a freely available common source of data for everyone {enwiki:1060114687}. Wikidata comes with a bunch of tools like, SPARQL end point, query builder, data visualization tools, etc.

Although wikidata has a tremendous amount of data in it, there was no information about software purposes. However, it was possible to obtain classes of software categories and, therefore, an indirect approach has been taken to list down list of possible software purposes from software categories by assuming each software category has essentially a software purpose associated to it.

Accordingly a SPARQL query has been formulated and a group of software categories has been identified. Take a look at the SPARQL query used on the Appendix. From the SPARQL query, the list of software categories obtained was well over 3000 software categories. To reduce this into a manageable and more general group of software, a network analysis has been done on 3000 groups of software categories using Gephi software. Using Gephi software, clustering of related software categories and filtering has been made to have a more generalized software categories.

According to the network analysis, all software tools fall into two main groups as system software and application software. The relation between the software categories has been sketched and summerized as follows:

### The software ontology (SWO)

The other ontology analyzed was SWO (the software ontology). Unlike wikidata, from this ontology, a list of possible software purpose were found directly in under “planned process” section. Navigate to the software purpose list in SWO as follows: “items”> “entity “>”concurrent process”> “planned process”. The software ontologies presented in the SWO ontology has been categorized in to two main groups as “data transformation” and “data visualization”. Under data transformation, 40 sub -types of data transformation software purposes is listed.

### OntoSoft

Ontosoft is a distributed semantic registry for scientific software and it is structured in 6 dimensions to provide information about a given software. Dimensions and associated information for the dimension is described on the table below:

|  |  |
| --- | --- |
| Dimension Name | Information |
| Identify | * Name of software, abbreviation of the software, etc. |
| Understand | * Creator of the software, publisher of the software, * *domain specific key words* |
| Execute | * URL for downloading the software, license, system requirements …etc. |
| Do Research | * Input / output file formats, preferred citation information |
| Get support | * Contact details, possible support included, etc. |
| Update | * Version, developer community, software development process , maintenance, etc. |

From the set of information provided in those 6 dimensions, particularly the “understand” dimension has nearly 400 domain specific key words that would potentially indicate software purposes. Therefore, those has been retrieved and condensed into a more general software purposes list of 60.

### Other sources

The other sources of information analyzed for the software purpose are DBpedia and Sci-crunch.

## Types of software purposes

Based on the analysis results form the software ontologies and scientific literatures in SoMeSci, a list of software purposes has been identified and consolidated into a group of 8 main types of software purposes as follows:

1. Data Collection
2. Data pre-processing
3. Data Analysis
4. Data visualization
5. Simulation
6. Stimulation
7. Modelling
8. Programing

### Data collection

Scientists collect data for their research using various means from using paper and to sophisticated software to collect data from various sources.

* https://en.wikipedia.org/wiki/Data\_collection
* Using electronic surveys to collect data: experiences from the field
* Preparing and conducting interviews to collect data
* Recording, Archiving, and Using WSR-88D Data
* Incremental Organization for Data Recording and Warehousing
* [Automatic indexing](https://en.wikipedia.org/wiki/Automatic_indexing)
* [Web Crawler](https://en.wikipedia.org/wiki/Web_crawler)

### Data pre-processing

* <https://en.wikipedia.org/wiki/Data_pre-processing>
* Sequence alignment

Often times after a data is collected, it has to go through a process of processing before it is actually used. This is

### Data Analysis

In a modern research, where a research is increasingly relying on processing of huge amount of data, the most common purpose of software usage purpose is to perform data analysis.

When it comes to the application of data analysis in actual research works, various kinds of data analysis techniques exit. Some of the data analysis techniques can be more general where as others are more domain [specific](https://pubs.acs.org/doi/pdf/10.1021/ac00238a008) .

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

Data analysis is a broad term which can refer to inspecting, cleaning, transforming, modelling data, etc. with a particular goal of discovering a meaningful information from the data which can be used to make conclusions or decisions [13] .

Modern research is increasingly data driven and software is being used extensively to analyze large amount of data. Using a data analysis software, scientists usually inspect, clean, transform and model data in search for meaningful information from the data which will in turn will support conclusions for a research [13].

Diverse types of data analysis techniques exist, some of which are more general and applicable to any field of research and science. But there are also some data analysis techniques which are endemic to only some domains. For Example, densitometric analysis. More general data analysis techniques are usually mathematical, like statistical data analysis, numerical analysis and so on.

The use of software for data analysis has several advantages. One is, a data analysis software gives more efficient and effective work. It allows analysis of a large volume of data. The other is, data analysis software gives insights hidden in a data, correlation between variables, etc.

We have diverse types of data analysis software. However, possible two main categories of data analysis software are Qualitative Data analysis software and Quantitative Data analysis software. Qualitative data refers to any from of data obtained by a researcher by direct observation, interviewing, recordings, etc. (<https://en.wikipedia.org/wiki/Qualitative_research> ). Qualitative research software is particularly popular in social science research where most of the data is qualitative by its nature. Qualitative data analysis software are collectively known as Computer Assisted Qualitative Data Analysis software ([CAQDAS](https://en.wikipedia.org/wiki/Qualitative_research#Data_analysis)) .

The other type of data analysis is quantitative which is based on numbers.

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].

* <https://en.wikipedia.org/wiki/Data_analysis>
* Data analysis in qualitative research
* Types of Data analysis
* Mining Scientific Data
* Network Analysis
* Statistical Analysis
* Numerical Analysis
* Data Mining – extraction / discovery of patterns in large data sets using ML, Statistics and Data.
* Regression Analysis
* Mathematical analysis
* Numerical Analysis
* Statistical Analysis
* Domain specific analysis
* Densitometric Analysis
* Voxel-based Analysis
* Continency Analysis in power systems reliability
* Image analysis ( computer vision)
* Data mining

### Data visualization

* https://en.wikipedia.org/wiki/Data\_visualization

### Simulation

* <https://en.wikipedia.org/wiki/Simulation>
* Simulation
* Flight simulation
* Event simulation
* Flood dynamics simulation
* Numerical simulation
* Simulations are run to improve understanding of a problem (segal2008developing).
* [Physics engine](https://en.wikipedia.org/wiki/Physics_engine)
* Simulation of vehicle schedule with any logic

### Stimulation

* https://en.wikipedia.org/wiki/Stimulation

### Modelling

* <https://en.wikipedia.org/wiki/Scientific_modelling>
* Thermal modelling and characterization in chips
* Modelling
  + [Graphics software](https://en.wikipedia.org/wiki/Graphics_software)
  + animation software

### Programing

* https://en.wikipedia.org/wiki/Computer\_programming

In a scientific investigation scientists use software for various purposes. The use cases of software in a research ranges from execution of some trivial tasks to execution of more critical tasks that will determine a research result [38].

# Annotation of Data

## Data set

## BRAT Annotation tool

# Feature selection

## Phrase

## Sentence

## Paragraph

## Dependency tree

# Selection of classifiers

# Classification and Evaluation

# Summary of Results and Conclusion

## Summary of Results

## Conclusion

## Limitations and future work