# Introduction

## 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 the

Scientists use various kinds software tools during their scientific investigation for various purposes. Some of the most common software usage purposes are described in this section.

## Recap

* 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

# 

# The role of Software in Scientific research

## Introduction

Nowadays scientific research is unthinkable without a use of software and scientific 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 are consistent and 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 and commonly software is not cited 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 roles of software in a research, at Zetoro group library. The main aim of ReSA is to influences decision makers to 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. For example, software enables 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 result { 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 control-system 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 field tested, 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 a 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 fur more processing, helps researchers interpret results, etc{ lutz2011software }.

Software, specifically imaging software, plays a critical role to assist medical researchers for early isolation of cancer and ultimately to saving life. 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 MR 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 reasonable to say that software helps to save life.

Software plays an important role in power system planning and operation. 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

A use of software also allowed to produces better scientific discoveries and several research breakthroughs has been made possible { goble2014better }.

### 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, 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 }.

# classification of software usage purpose

## Introduction

In scientific investigation broad range of software, from simple scripts to extremely complex software with millions of lines of code behind huge projects like the LHC and the Square Kilometer Array, is being used for various purposes. Typical software purpose of software use includes simulations, computational modelling, data manipulation to drive interesting insights (goble2014better).

* Types of software ( categories )
  + Scientific software / research software vs commercial software
  + System software vs application software

The use of software ranges from execution of rudimentary tasks to computation of extremely complex tasks.

The use of software ranges from execution of rudimentary tasks to computation of extremely complex tasks.

can be used for the execution of rudimentary is used for execution of some trivial tasks like word processing and in other cases they use a software to perform critical tasks that can ultimately determine their research end result.

## Methodology

To find out the list of all possible software usage purposes in a research mainly two things have been done. First, scientific literatures has been manually analyzed manually for listing out of possible software usages. Secondly, ontologies have been analyzed for possible software purpose.

## Software purposes

Based on the analysis of scientific literatures and ontologies, overall 8 types of software usage purposes has been identified. These are ….

### Data Collection

### Data pre-processing

### Data Analysis

* Data analysis in qualitative research
* <https://en.wikipedia.org/wiki/Data_analysis>
* Types of Data analysis
* Mining Scientific Data

### Modelling

### Stimulation

### Visualization

### Programming

### Simulation

Simulations are run to improve understanding of a problem (segal2008developing).

***Question***: What are possible software usage purposes in the literatures? ( For what purpose do researchers use software in their research paper ?? )

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]. Some of the most common software usage purposes in a research are:

* Data Analysis and Processing
* Modelling
* Simulation and
* Programming

**Data Analysis**

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.

### Software usage purposes in a research

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.

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] .

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 Some of the most common software use cases in a research are:

* Data collection
* Data processing
  + Image processing
* Data Analysis
  + Mathematical Analysis
    - Statistical Data analysis
    - Numerical analysis
  + Text Analysis
    - Text mining
* Modelling
* Simulation
* Programming

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

### Software usage purpose in science and mathematics

**Data Analysis**

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 Collection
    - [Automatic indexing](https://en.wikipedia.org/wiki/Automatic_indexing)
    - [Web Crawler](https://en.wikipedia.org/wiki/Web_crawler)
  + Data Analysis
    - Mathematical analysis
      * Numerical Analysis
      * Statistical Analysis
    - Domain specific analysis
      * Densitometric Analysis
      * Voxel-based Analysis
    - Data visualization
  + Data Processing
  + Data Mining – extraction / discovery of patterns in large data sets using ML, Statistics and Data.
  + 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 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

Simulation

Flight simulation

Event simulation

Flood dynamics simulation

Numerical simulation

Simulation of vehicle schedule with any logic

Modelling

Data recording and retrieval

Data processing

Reference management

Data mining

Sequence alignment

Image processing

Image analysis ( computer vision)

Network Analysis

Statistical Analysis

Numerical Analysis

Regression Analysis

Continency Analysis in power systems reliability

Thermal modelling and characterization in chips