Neasa Tang - 14311106

To deliver a report that considers the ways in which the software engineering process can be measured and assessed in terms of measurable data, an overview of the computational platforms available to perform this work, the algorithmic approaches available, and the ethics concerns surrounding this kind of analytics.

Software Measurement

Software measurement is vital in the software engineering discipline. [1] People who are involved in the IT field are constantly being faced with new technologies along with various other things such as competitive markets. While doing so, they also at the same time have to worry about the reliability of the product, the stability of the product, the testing of the product. This produces a need for software measurement to aid for these worries.[2]

Although software measurement by itself cannot solve these problems that the people in IT face, however, it can help elucidate and focus their understanding of problems. Additionally, when software measurement has been done correctly, sequential measurements of quality attributes of products and processes can provide an effective foundation for initiating and managing process improvement activities.

There are many different ways in which software engineering processes can be measured. In my essay, I will discuss lines of code, balance scorecard and instruction path length.

Lines of Code (LOC)

This is also know as source lines of code (SLOC). This is a software metric that is used to measure the size of a computer program by counting the number of text lines in the program's source code. SLOC is mostly used as an indicator for programmers to see how much work is left to be done for the program to be complete. This metric is also used to estimate the productivity and maintainability of the program once the software has been produced.[3]

There are two primary types of SLOC measures:
a) Physical SLOC (LOC) - It is a count of lines in the text of the program's source code which includes comment lines and blank lines unless the lines of code in a section consists of more than 25% blank lines.

---

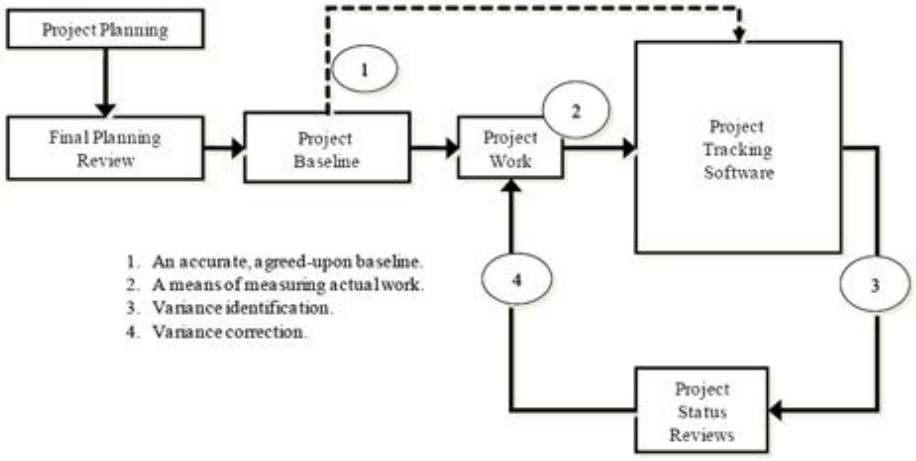
[1] https://link.springer.com/article/10.1007/BF02249053
[2] http://repository.cmu.edu/cgi/viewcontent.cgi?article=1231&context=sei
[3] https://en.wikipedia.org/wiki/Source_lines_of_code

b) Logical SLOC(LLOC) - This measures the number of “statements” however their specific definitions are related to specific computer languages i.e. a simple logical LLOC measure for C-like programming languages is the number of statement-terminating semicolons.

It’s much easier to create tools that measure physical SLOC. Furthermore, physical SLOC definitions are easier to explain that Logical SLOC. [4]

Balance Scorecard

A Balance Scorecard is an example of a cybernetic control applied to the management of the implementation of a strategy. A cybernetic control involves a self-correcting feedback loop as can be seen from the diagram below:



Instruction Path Length

The instruction path length of a program is the number of machine code instructions that are need to perform a section. The complete instruction path length is a measure of the performance of the algorithm on a specific PC hardware. The path length of a simple standard conditional instruction would generally be considered as equal to 2.

One instruction is used to perform the comparison operation and the second instruction is used to take a branch if the condition holds true/false. The amount of time that it takes to perform each instruction is generally used as an indicator of relative performance rather than in any sense absolute and definite.[5]

[4] http://www.projectcodemeter.com/cost_estimation/help/GL_sloc.htm
[5] https://www.revolvy.com/main/index.php?s=Instruction%20path%20length

Computational Programs

There are many computational programs that aid in the measurement of software engineering process such as Hackystat, Leap and PSP. These will be the platforms that I will talk about in depth.

Hackystat

Hackystat is an open source framework that is used that automatically collects data and analysises the data of the product and software engineers. The aim of the Hackystat platform is to provide a mechanism that is extendable that can thoroughly reduce the overhead associated with collection of a wide variety of software engineering data. Alongside that, the platform contains a toolkit of analyses that can make a useful report.

Hackystat is widely used in application areas. This can include classroom pedagogy, software engineering of high performance computing systems[6].

Leap

The Leap toolkit is created to provide Lightweight, Empirical, Anti-measurement dysfunction, and Portable approaches to software developer improvement. If software engineers use Leap, they can gather and analyze personal data regarding time, size, defects, patterns, and checklists.

In the platform Leap, there are two main activities. These are gathering primary data and performing Leap analyses. These can be increased by secondary activities of refining definitions, checklists, and patterns. Finally, these central and/or secondary activities can be directed toward individual skill acquisition and improvement or group review of work products. [7]

PSP

The Personal Software Process (PSP) is a software development process that is created to help software engineers better understand and improve their performance by keep track of their predicted and actual development of their source code.

Watts Humphrey created this platform to apply the underlying principles of the Software Engineering Institute's (SEI) Capability Maturity Model (CMM) to the software development practices of a single developer. The idea behind this platform

---

[6] https://www.researchgate.net/publication/228731372_Experiences_with_hackystat_as_a_service-oriented_architecture

[7] https://pdfs.semanticscholar.org/d0c3/5a51a50bd8f8f348a8bb2d0298b5ba83cec0.pdf

is to give software engineers the process skills necessary to work on a team software process (TSP) team.[8]

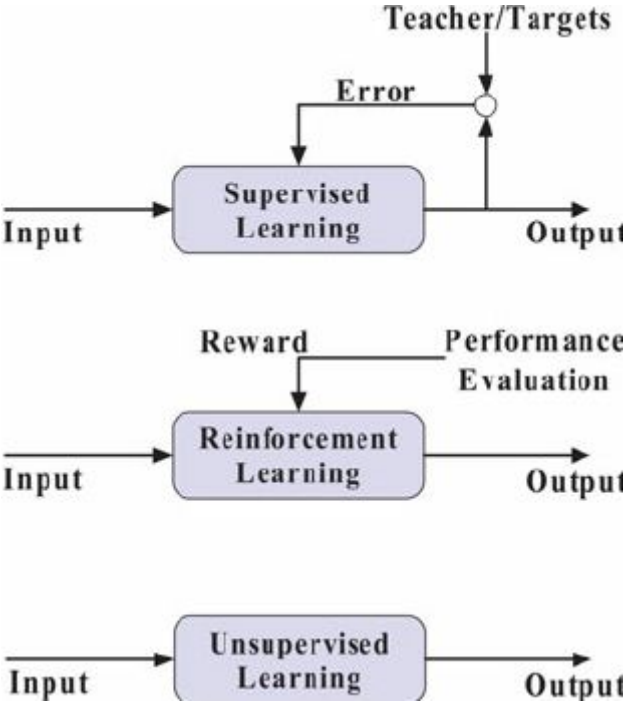
Machine Learning

What is Machine Learning? “Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.”

Machine learning focuses on getting the computer program to be able to grow and learn from themselves rather than from through programmed instructions that tells the program what to do.

Machine Learning is divided into 3 types of algorithms:

1) Unsupervised Learning
2) Supervised Learning
3) Reinforcement Learning



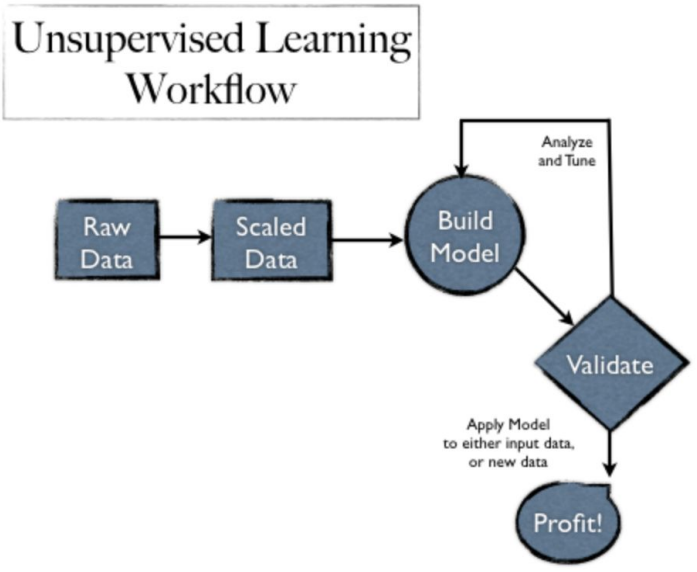
[8] https://en.wikipedia.org/wiki/Personal_software_process

Unsupervised Learning

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.[9]

Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. [10]

Cluster analysis is the most popular type of unsupervised learning. It is used to find hidden patterns and grouping from the data analysis. The clusters are modeled using a measure of similarity which is defined upon metrics such as Euclidean or probabilistic distance.



11

Supervised Learning

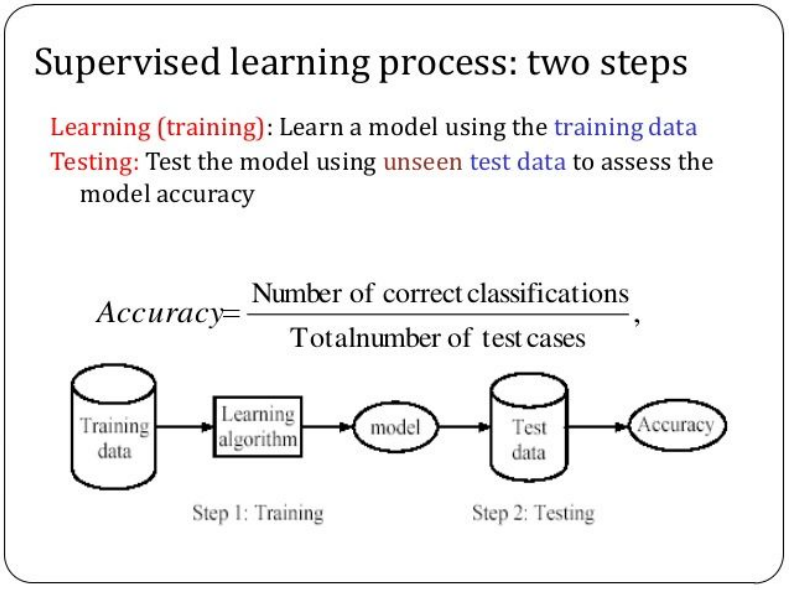
Supervised learning is the most common one that is used in the practical machine learning. How supervised learning works is that you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output i.e. Y = f(X).

The goal of supervised learning is to estimate the mapping function accurately that when there is a new input data (x), you can predict the output variables (Y) for that data.

---

[9] https://www.mathworks.com/discovery/unsupervised-learning.html
[10] http://www.expertsystem.com/machine-learning-definition/
[11] https://careermemo.wordpress.com/tag/data-analysis/

It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher overlooking and supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. The learning process ends when the algorithm reaches to a satisfactory level in performance.[12]



[13]

However as a result of this type of machine learning, there are two types of problems that can occur:

1) Classification Problem

What this involves is taking the input data and deciding which class/category the input data belongs to based on the training data. This deals with instances when the predicted output is a category, e.g red, blue, etc.
For example: an algorithm that sorts vehicles by colour or by type.

2) Regression Problem

Regression deals with estimating the relationship between variables. It is used in instances when the expected output is a real value e.g. weight, currency etc.

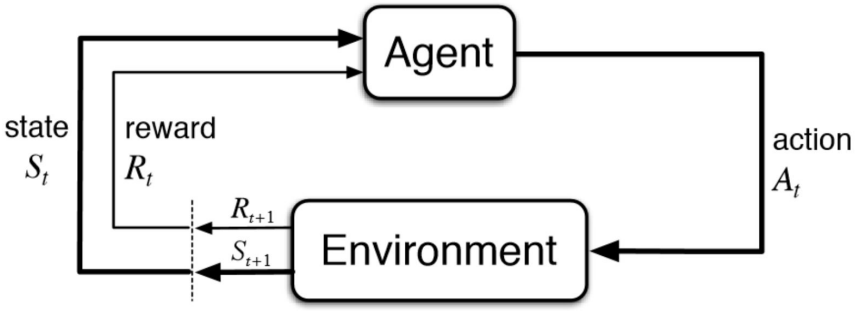
For example: an algorithm that determines the energy consumption used in a particular period of time for a particular country.

[12] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
[13] https://www.slideshare.net/tonmoybhagawati/presentation-on-supervised-learning

The main reason why a lot of people claim that supervised learning is deemed to be not so great is that it's limited in terms of scalability. Not only that but it is very costly to get it to a near perfect system.[14]

Reinforcement Learning

Reinforcement Learning is a type of machine learning algorithm that allows the machine to learn its behaviour based on feedback from the environment. This behaviour can be learnt once and for all, or can carry on adapting as time goes by. If the problem is created carefully, some Reinforcement Learning algorithms can converge to the global optimum; this is the ideal behaviour that maximises the reward.



15

This automated learning scheme means that there is little need for a human expert who knows about the domain of application. An advantage of this type is that there is less time spend on designing the model, since there is no need for hand-crafting complex sets of rules. Not only that, all that is necessary for this is for someone to be proficient with Reinforcement Learning.[16]

Ethics

Ethics is the study of value concepts whether it's good or bad. With regards to ethics, I will discuss among the ethics of privacy and accuracy.

Privacy:

Privacy is one of the major ethics involved in software engineering. The reasoning behind this is due to the improper access to personal information and the misuse of information.

a) Location Privacy

[14] https://dev.to/overridedeveloper/an-intro-to-supervised-learning-the-good-the-bad-and-the-ugly-1oo
[15] http://web.stanford.edu/class/cs234/index.html
[16] http://reinforcementlearning.ai-depot.com

Location-based tracking systems (LTSs) use a wide range of different technologies to record the locations of objects and in particular, users. Using LTS affects the security and privacy of users.

When using an App that accesses your location, you don't know what other Apps or companies also know about your location, hence increasing the issue with security and privacy of users. As a result, they can use your data, and then can recommend you ads based off your location.

For example, with the development of mobile apps that track a user's location in real time such as Snapchat, the ads on the App can even be adapted to the user's current location. [17]

b) Public Information

When using many software or services these days, it involves taking in our data. Some data can be personal and very sensitive and as a result, data abuse exists. This is a reason why security is such an important area in IT. However, because it takes in our data, it is never too clear as to what they are doing with our data.

For example, marketing companies thrive from personal information. They collect user's data and as a result, use that data to predict people's purchasing preferences. This helps to target and adapt their advertising. This information that they retrieve can be purchased from other sources to help companies with their sales. Privacy policies exist however, in many cases, the user doesn't have to read all of it and if they do not accept the terms, they cannot avail of the service. As a result, they end up handing over their personal information for a service, not knowing what that company is actually doing with their data. There are also companies such as facebook who change their privacy policies without giving the users adequate notice e.g. Facebook is a company who has done this a number of times. [18]

Accuracy

Accuracy is another focal point of ethics in software engineering. Although it is a broad topic, it has many associated ethical issues.

a) Software Accuracy

The analyst of a system is supposed to know and be able to predict all states for complex systems however this leads to various ethical issues regarding software

---

[17] http://ieeexplore.ieee.org/document/5155910/?reload=true

[18] https://www.scu.edu/ethics/focus-areas/internet-ethics/resources/unauthorized-transmission-and-use-of-personal-data/

accuracy. An example of a way to approach these ethical issues is performing the software through system validation and verification.

At first sight, it would appear that a system developer would be ethically bound to correct all system errors. However, if they deal with errors it can raise ethical dilemmas where 15-20% of attempts to remove program errors generally introduce a couple of more errors.

For example, if fixing an error in a program that has more than 100,000 lines of code, the chances of introducing a more severe error is so high that it is sometimes better to keep the error and work around it rather than trying to fix it.

b) Language and Culture

Language and terminology used to frame a question can significantly influence the accuracy of the information elicited. This is true for any system in which the system user is forced to converse with software using concepts unfamiliar to them.

In Chinese culture for example, there are many superstitions that involve around the number “4” and “8”. This is because 4 means “death” and 8 sounds like “wealth”. As a result, if an app has a survey and has a list of numbers and asks for the user to tap on their favourite number, 4 will have a very low number and 8 will have a very high number. However, in other countries, these superstitions don’t exist, or other superstitions involving other numbers might exist, and as a result, the data will not be as accurate as the people who are running the app would like it to be. As a result, knowing about different cultures and languages is important regarding the ethics of data analytics.

References

1. https://link.springer.com/article/10.1007/BF02249053
2. http://repository.cmu.edu/cgi/viewcontent.cgi?article=1231&context=sei
3. https://en.wikipedia.org/wiki/Source_lines_of_code
4. http://www.projectcodemeter.com/cost_estimation/help/GL_sloc.htm
5. https://www.revolvy.com/main/index.php?s=Instruction%20path%20length
6. https://www.researchgate.net/publication/228731372_Experiences_with_hackystat_as_a_service-oriented_architecture
7. https://pdfs.semanticscholar.org/d0c3/5a51a50bd8f8f348a8bb2d0298b5ba83cec0.pdf
8. https://en.wikipedia.org/wiki/Personal_software_process
9. https://www.mathworks.com/discovery/unsupervised-learning.html
10. http://www.expertsystem.com/machine-learning-definition/
11. https://careermemo.wordpress.com/tag/data-analysis/
12. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/
13. https://www.slideshare.net/tonmoybhagawati/presentation-on-supervised-learning
14. https://dev.to/overridedeveloper/an-intro-to-supervised-learning-the-good-the-bad-and-the-ugly-1oo
15. http://web.stanford.edu/class/cs234/index.html
16. http://reinforcementlearning.ai-depot.com
17. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.8164&rep=rep1&type=pdf
18. http://ieeexplore.ieee.org/document/5155910/?reload=true
19. https://www.scu.edu/ethics/focus-areas/internet-ethics/resources/unauthorized-transmission-and-use-of-personal-data/