# Application for Ethics Approval

Source: *https://turingcomplete.topdesk.net/tas/public/ssp/content/serviceflow?unid=2e579708ee364171b58eb4c8*

## 1. Project Goal and Purpose

> Please introduce your project. Tell us, among other things: - the purpose and goal of your project - the intended benefits of your project - how this research project aligns with the Turing's goals, challenges, and/or objectives

### Background

Machine learning techniques are effective for building predictive models because they are good at identifying patterns in large datasets. However, the development of a model for complex real life problems often stops at the point of publication or proof of concept.

The practice of maintenance of predictive models is crucial for safe and effective long term use. A model developed using retrospective data in the medical domain risks becoming obsolete as medical treatments advance.

> Clinical data is highly dependent on the landscape of clinical practice as well as underlying population demographics and comorbidities, all of which vary over time. The complete utility of a healthcare model can be nearly impossible to ascertain unless one accounts for the inevitable effect of temporal dataset drift
>
> [https://arxiv.org/pdf/1908.00690.pdf]

In the case of cystic fibrosis, new treatments mean that patients can survive longer without the need for a lung transplant; this means that an algorithm that predicts when a patient requires a transplant will be gradually be making predictions about older patients with different comorbidities. Breast cancer risk assessment models which were trained on data collected from patients ten years ago may be using features such as smoking, when smoking as a trend amongst women has changed over the years. Another example is when new treatments are introduced or new data types such as genomic profiles and imaging need to be incorporated into prediction models.

Retraining a single model can be computationally expensive. Retraining a model which comprises of an ensemble of dependent models is both expensive and complex. This is especially true if models provide absolute class decisions instead of probabilistic values, because absolute class decisions are based on carefully selected thresholds. Thresholds are selected to maximise metrics such as sensitivity or specificity, according to purposes of applications. Reselection of multiple thresholds to rebalance system performance is both complex and brittle.

For all these reasons, decisions to retrain models are difficult decisions. In addition resulting changes in model behaviour can be a source of uncertainty for users making high impact or high risk decisions. If reasons for making changes are not available or communicated clearly, this can result in user distrust and resistance.

### The purpose and goal of this project

The purpose of Learning Machines is to enable continuous appraisal of the performance of an algorithm, as new data is collected over time.

The goal of LM project is to determine, design and build software infrastructure neccessary to support the continuous appraisal of an algorithm performance, for the long term. In this document, we will refer to this infrastructure that supports long term monitoring of an algorithm, the intelligence infrastructure (II).

There are many ways to build II. This project will prioritise approaches and components which enables good research engineering practices with minimal overheads to researchers themselves. Examples of modules to be used in the Learning Machines Intelligence Infrastructure (LM II) are:

1. Pre-specification of statistical measures of the dataset used to train a model leading to automated report generation of those measures when new data is available (eg. when the algorithm is used)
2. Version control of datasets
3. Sotware packages for enabling reproducibility checks
4. Version control of algorithms
5. Pre-specification of algorithm performance measures, including acceptability threshholds
6. Automated testing leading to generation of algorithm performance reports
7. Good documentation.

### How this research project aligns with the Turing's goals, challenges, and/or objectives

'Learning Machines - Best Practices for Models in Production' addresses the issues about keeping models updated, in order to reflect data changes in trends and changes in the current environment.

Learning Machines aligns with Turing goals because at the Turing we are interested in developing best practices for research and software engineering. The intelligence infrastructure proposed will support long term maintenance of prediction models built using machine learning techniques.

## 2. Data and Research Methods Description

Please provide a description / overview of the data that you will use for your project, if any, as well as the research methods you will use.

> Please cover: - the type of data - the amount of data in the sample, e.g. number of people, items, variables, months, years covered in the data - the sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from - if known, where and how the data to be used was collected - a brief description of the research to be carried out and the methods that will be used

### The sources of the data, e.g. the datasets that will be used, who they were created by, where they are accessible from

We propose to obtain datasets from the Surveillance, Epidemiology, and End Results (SEER) Program. The SEER program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS).

We have identified SEER as the source for data as one of its goals are as below:

> Describe temporal changes in cancer incidence, mortality, extent of disease at diagnosis (stage), therapy, and patient survival as they may relate to the impact of cancer prevention and control interventions.
>
> https://seer.cancer.gov/about/goals.html

There are two SEER products: - SEER Research, which excludes geography (county, state/registry), month in dates (e.g. month at diagnosis), and a few other demographic fields. and - SEE Research Plus, which include geography, months in dates, and other demographic fields, as well as information on radiation therapy and chemotherapy given as part of the first-course treatment.

Learning Machines will apply to access SEER Research, therefore accordingly each member of the Learning Machines team will sign a SEER Research Data Use Agreement (DUA) to request for SEER data. A copy of the DUA is found here and also included as supplementary information with this application.

As per accessibility protocol for SEER Research, the SEER Stat program will be used to perform cohort selection neccessary for the project and to download SEER data.

### The type of data

We propose to obtain cancer incidence data; these data were collected by the SEER program and amalgamated from multiple population-based cancer registries. Data types include patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment, follow ups with patients for vital status.

### The amount of data

The dataset we are considering contains 5,347,692 records of incidences of tumours, collected from up to 9 cancer registries in the United States, from years 1975-2017

There are approximately 141 patient features in the the SEER Research database; examples of these which are relevant to Learning Machines are summarised in the table below.

| Name | Data Type (Category vs Numerical vs Date) | Description |
|---|---|---|
| Age recode with <1 year olds | Category | This recode has 19 age groups in the age recode variable (< 1 year, 1-4 years, 5-9 years, . . . , 85+ years). |

| Name | Data Type (Category vs Numerical vs Date) | Description |
|---|---|---|
| Race recode (White, Black, Other) | Category | Race recode is based on the race variables and the American Indian/Native American IHS link variable. |
| Year of diagnosis | Date (Year) | Values are 1973-2014 but may be a subset. |

| Name | Data Type (Category vs Numerical vs Date) | Description |
|---|---|---|
| Primary Site | Category | Codes are found in the Topography section of the International Classification of Diseases for Oncology (ICD-O) 3rd edition |

| Name | Data Type (Category vs Numerical vs Date) | Description |
|---|---|---|
| Grade | Category | Grade Based on grade codes in ICD-O-3 |

| Name | Data Type (Category vs Numerical vs Date) | Description |
|---|---|---|
| Total number of in situ/malignant tumors for patient | Numerical | Based on maximum sequence number of any malignant/in situ tumors in SEER through the last released year of diagnosis. This value is the same across all tumors for a person |

| Name | Data Type (Category vs Numerical vs Date) | Description |
|---|---|---|

The full list of patient features are listed as a data dictionary here

### A brief description of the research to be carried out and the methods that will be used

The first step for LM is to automate the generation of quality control checks and descriptive statistics for the dataset. This will be developed using Python Unit Testing framework. We will be using both single value measures (eg. distribution or mean) and also more complicated methods such as autoencoders to quantify how values are changing over time and whether a retraining is necessary.

The algorithm or analysis performed here on the dataset will be simplistic 'toy models' or have been published with open source code. We will be looking at deploying well established machine learning methods (e.g. random forest) to predict disease outcomes.

It is important to note that the algorithm or analysis itself is not the focal point of the project, instead it is features of datasets changing over time as new data is accumulated, and how this affects algorithm performance.