# Purpose

This is a documentation of the process undertaken to curate a catalogue of machine learning tools applied in European governments. The purpose of this catalogue is for the Policy Innovation Lab and its partners in the South African government to have a clear sense of what is possible. In order to achieve this we attempt, where possible, to focus on machine learning tools that can be applied to existing South African datasets, or datasets that we predict can be readily produced. We also prioritise those tools that can be developed easily, bearing in mind South Africa’s current digital infrastructure and technical expertise. Furthermore, we discuss potential challenges, if they exist, to deploying tools that were successful in Europe to the South African context. In summary, we catalogue the low-hanging-fruit of machine learning tools that can be deployed in the South African government, prioritising those that yield a good societal return on investment.

# The dataset

The dataset that we shall curate is the most recent version of the *Public Sector Tech Watch latest dataset of selected cases* by the European Commission, Joint Research Centre (JRC) available at

<http://data.europa.eu/89h/e8e7bddd-8510-4936-9fa6-7e1b399cbd92>

This is a catalogue of 2154 examples of emerging digital technologies, including AI and blockchain, deployed in EU governments which is published under a Creative Commons Attribution 4.0 International License. The dataset was created on 2023-10-11 and at the time of our access was last modified on 2024-05-30. We choose this data source for the following reasons:

1. It is an extensive catalogue with a substantial amount of data describing the context, development and use of the tools.
2. The EU contains a diverse group of countries in various stages of development, so although we may have additional work to do in considering the South African context, many of the tools will be applicable.
3. The EU is far ahead of the rest of the world in terms of regulating the use of AI. We therefore may assume that tools in this catalogue could be deployed in South Africa and meet its regulatory requirements as they develop.
4. The publishing license allows us to share the original material, whole or in part, whether adapted or not, indefinitely given that we do not break any terms of the license unless asked to by the licensor.

### Cleaning the data

The dataset was cleaned in two ways.

1. In the ‘Responsible Organisation category’ feature we replaced “Central-government” with “Central-Government”, since both were present.
2. The “AI keywords” feature was sparse. We updated some empty values by searching the descriptions for AI keywords that were used elsewhere, and setting those as the new values.

# Methodology

The curate the dataset using the methodology described below.

1. We filter the original catalogue:
   1. We apply the filters (in order of the columns/features) described in Table 1:

|  |  |  |
| --- | --- | --- |
| **Feature to filter** | **Exclude** | **Include** |
| Responsible Organisation category | Community-led,  Consortium,  European Institute/Agency,  Non-governmental,  Private sector | Academic-research,  Central-Government,  Local Government,  Regional Government,  Central-government |
| Status | Not in use anymore | Implemented,  In development,  Pilot,  Planned |
| Cross Border | Yes | No |
| Technology | Blockchain,  Quantum Computing,  AR/VR,  Virtual Worlds | Artificial Intelligence |
| Date-updated | \*Excludes everything after May 2024 by default. | \* Includes everything up to and including May 2024 |

Filtering was done using Python, and the code is available at [FILL]. After filtering there are [FILL] examples remaining. Note that some data cleaning was required. For example, “Central-government” and “Central-Government” were both values in the “Responsible Organisation category” feature.

* 1. We further filter the catalogue by going through each example and removing those that we consider to be either clearly irrelevant or clearly impractical to the South African context. Our judgement of irrelevancy was informed by several factors, including:
     1. that those ML tools are already used for the same purpose in South Africa,
     2. the ML tools are not specific to government use, for example, cancer detection,
     3. the tools would be prohibitively expensive to develop or maintain, or
     4. the data was clearly unavailable and not easily obtainable.

Our labelling criteria was:

0 - relevant to SA, worth keeping in the dataset  
1 - not relevant, remove from the dataset  
2 - very relevant, pay particular attention to it  
3 - relevant to current Policy Innovation Lab projects to SA  
4 – not very relevant

Upon review, both those labelled 1’s and 4’s were both removed from the database. The removal of 4’s is primarily since this is an effort to curate the dataset to obtain the examples and possibilities most worthy of our attention. A list of the PSTW IDs removed in this step is given in the appendix.

1. We constructed a list of South African datasets that could be used for various AI projects in government. We limited ourselves to datasets that are either public, listed on the StatsSA website, or that we expect exist within departmental databases. This list is not intended to be extensive, but to be useful in considering the ease-of-implementation of these different tools.
2. We group the remaining examples by the sub-categories in the parent categories “Improved Public Service”, “Improved Administrative Efficiency” and “Open government capability”.
3. For each sub-group we write a summary, highlighting the examples that we consider to have the greatest potential societal benefit given the magnitude of the investments. In our summaries we also consider how these tools could use data currently in existence, those that use data that can be obtained relatively easily, or those that use data that overlaps with other Policy Innovation Lab projects. We also tag examples that the Policy Innovation Lab has the resources to implement ourselves. Our opinions are educated by
   1. Our understanding of the current global literature on AI in government which includes case studies not contained in our database and literature exploring the policy and regulatory requirements necessary for the ethical use of AI
   2. Our lab expertise in South African policy,
   3. Our lab expertise understanding in the technical aspects of machine learning tools.
   4. Our labs soft knowledge of South Africa.

# Implementation

* In step 1a we went from 1620 to 967 examples.
* In step we went from 967 to 727 examples.
* In step 2 we compiled the following list of open data sources containing relevant data. These include:
  + [Open Data for Africa](https://southafrica.opendataforafrica.org/data/#menu=topic)
  + [StatsSA](https://www.statssa.gov.za/?page_id=1847)
  + [OpenUp](https://data.openup.org.za/)
  + [Open Data ZA](https://opendataza.gitbook.io/toolkit)
  + [SA Data & Policies spreadsheet](https://docs.google.com/spreadsheets/d/1asrQMHp_aJrD-LqkmW9n5yLT6Cm-K1geBEn9nLfYb3E/edit?gid=388540894#gid=388540894)
* This curated dataset is available at [FILL] and the source code at [Fill].

# Limitations

1. The original dataset is limited in the sense that it does not contain all modern technological tools deployed or planned for government use in the EU, but are those that AI Watch Europe collected.
2. A serious limitation of this document is that we do not know the performance of the tools deployed since there are very few quantitative empirical studies on the effectiveness of these AI tools in government. This is mainly because these tools were deployed too recently to fairly judge their effectiveness or because it is difficult to quantify their effectiveness. Thus our recommendations and some of the steps taken in curating the data boil down to informed opinion, and are, ironically, not necessarily data-driven decisions. However, we have balanced this with honesty and openness about our methodology.
3. Our knowledge of what data already exists is limited. This is in part due to the fact that there is no centralized repository where we could find all governments national and departmental databases.
4. There were two primary contributors to this project, and due to time constraints they did not cross-validate each others work, leading to possible inconsistencies in determining what is important.

# Notes

1. Any downstream sharing of our curated dataset must fall under the same licensing agreement as the original published dataset.
2. For the sake of explainability, reproducibility and faithful use of open data we make sure that all source code and decisions are available, either in this document or linked. If you would like access to our catalogue or more details about the decisions taken please contact the Policy Innovation Lab at Stellenbosch University.
3. The original dataset is constantly being updated. We use the version modified on May 30th 2024. Given the rapidly changing nature of these technologies, it is possible that many of the most beneficial applications of AI technologies will not be contained in the current catalogue.

# Appendix

## List of PSTW IDs deleted in step 1b:

PSTW-8,PSTW-11,PSTW-16,PSTW-17,PSTW-19,PSTW-31,PSTW-34,PSTW-37,PSTW-39,PSTW-43,PSTW-51,PSTW-59,PSTW-61,PSTW-64,PSTW-68,PSTW-86,PSTW-90,PSTW-96,PSTW-97,PSTW-98,PSTW-100,PSTW-102,PSTW-103,PSTW-111,PSTW-115,PSTW-116,PSTW-119,PSTW-120,PSTW-121,PSTW-140,PSTW-141,PSTW-142,PSTW-148,PSTW-152,PSTW-153,PSTW-155,PSTW-161,PSTW-163,PSTW-164,PSTW-167,PSTW-168,PSTW-169,PSTW-171,PSTW-172,PSTW-173,PSTW-175,PSTW-179,PSTW-183,PSTW-184,PSTW-192,PSTW-196,PSTW-199,PSTW-222,PSTW-232,PSTW-233,PSTW-235,PSTW-236,PSTW-239,PSTW-240,PSTW-243,PSTW-245,PSTW-246,PSTW-256,PSTW-259,PSTW-260,PSTW-264,PSTW-271,PSTW-275,PSTW-276,PSTW-279,PSTW-292,PSTW-293,PSTW-312,PSTW-319,PSTW-323,PSTW-394,PSTW-441,PSTW-443,PSTW-453,PSTW-460,PSTW-461,PSTW-469,PSTW-531,PSTW-543,PSTW-544,PSTW-545,PSTW-563,PSTW-628,PSTW-629,PSTW-631,PSTW-634,PSTW-640,PSTW-645,PSTW-662,PSTW-663,PSTW-675,PSTW-683,PSTW-990,PSTW-1016,PSTW-1018,PSTW-1053,PSTW-1058,PSTW-1068,PSTW-1070,PSTW-1153,PSTW-1155,PSTW-1161,PSTW-1162,PSTW-1167,PSTW-1243,PSTW-1329,PSTW-1338,PSTW-1340,PSTW-1345,PSTW-1346,PSTW-1355,PSTW-1366,PSTW-1376,PSTW-1381,PSTW-1383,PSTW-1388,PSTW-1462,PSTW-1484,PSTW-1502,PSTW-1503,PSTW-1514,PSTW-1525,PSTW-1561,PSTW-1566,PSTW-1571,PSTW-1575,PSTW-1578,PSTW-1584,PSTW-1587,PSTW-1591,PSTW-1604,PSTW-1626,PSTW-1643,PSTW-1646,PSTW-1647,PSTW-1653,PSTW-1654,PSTW-1655,PSTW-1656,PSTW-1657,PSTW-1658,PSTW-1660,PSTW-1664,PSTW-1665,PSTW-1672,PSTW-1673,PSTW-1677,PSTW-1679,PSTW-1687,PSTW-1688,PSTW-1691,PSTW-1692,PSTW-1693,PSTW-1694,PSTW-1702,PSTW-1721,PSTW-1722,PSTW-1723,PSTW-1724,PSTW-1725,PSTW-1726,PSTW-1727,PSTW-1728,PSTW-1729,PSTW-1731,PSTW-1734,PSTW-1736,PSTW-1738,PSTW-1739,PSTW-1740,PSTW-1741,PSTW-1745,PSTW-1749,PSTW-1758,PSTW-1760,PSTW-1761,PSTW-1774,PSTW-1777,PSTW-1778,PSTW-1801,PSTW-1804,PSTW-1805,PSTW-1806,PSTW-1808,PSTW-1809,PSTW-1810,PSTW-1812,PSTW-1813,PSTW-1814,PSTW-1850,PSTW-1851,PSTW-1870,PSTW-1875,PSTW-1886,PSTW-1888,PSTW-1889,PSTW-1915,PSTW-1917,PSTW-1932,PSTW-1937,PSTW-1975,PSTW-1979,PSTW-1980,PSTW-1981,PSTW-1987,PSTW-1990,PSTW-1991,PSTW-1992,PSTW-1995,PSTW-1996,PSTW-1998,PSTW-2001,PSTW-2004,PSTW-2006,PSTW-2065,PSTW-2072,PSTW-2073,PSTW-2084,PSTW-2085,PSTW-2088,PSTW-2089,PSTW-2094,PSTW-2099,PSTW-2125,PSTW-2126,PSTW-2154,PSTW-1063,PSTW-1406,PSTW-1450,PSTW-1457,PSTW-1491,PSTW-1500,PSTW-1501,PSTW-1504,PSTW-1524.