WS1: Idea to Valid Solution - Theme 5: Knowledge Transfer

Final Report

Author: Michael Reusens (Statistics Flanders)

Collaborators: Christophe Bontemps (United Nations ESCAP), Carmel Colohan (Northern Ireland Statistics and Research Agency), Tim Linehan (CSO, Ireland), Sollange Correa-Onel (ONS, UK)

Date: 23/11/2021

Introduction

It is challenging to identify use cases where ML has significant added value in the creation of official statistics. One straightforward way to do this is to see in which cases ML has had added value in other NSI's, and then replicating these successful projects in your own country or domain. This has a clear benefit for the replicating NSI but is also beneficial for the NSI doing the original project. They can get their findings challenged and confirmed. Furthermore, the replicating project might lead to further improvements that can be shared with the original country.
This activity aimed at shedding light on how often statistical organisations are replicating each other's ML work, what the enablers and blockers of replication are, and if they are open to having their ML work replicated in the future.
To this end, a questionnaire to elicit current replication activity in statistical organizations was sent out to all participants of the ML2021 project. In this report we will report on its core results.

Results

Terminology

The *target organisation* is the organisation that is doing the replication of another organisation's work.
The organisation whose work is being replicated is called the *original organisation*.
If an organisation is either a target- or original organisation, it is a *replicating organisation*.

Response

49 people from 29 statistical organizations in 26 countries responded. Some organisations had multiple respondents. The results from respondents from the same organisation were aggregated using the mode (with a common-sense tie breaker) whenever there were questions about the organisation as a whole.

40 respondents had a technical role within their organisation, such as methodologist, statistician, data scientist and economist. The other respondents had a mix of roles, including project leader, team leader, deputy director and business analyst.

Current replication activity

Figure 1: Amount of replication in organisations

Figure 1 shows how frequently statistical organisations are replicating other organisation's ML work or having their ML work replicated by other organisations. The responses from the same organisation were aggregated using the mode with ['yes' > 'no' > 'I do not know'] as tie breaker. Out of 29 organisations, only 5 report to have replicated ML work, while 16 report not to have replicated ML work. Appendix 1 gives a brief description of the ML projects that have been replicated.

Note: For Figures 2-5, only the responses of the 9 respondents of replicating organisations were selected.

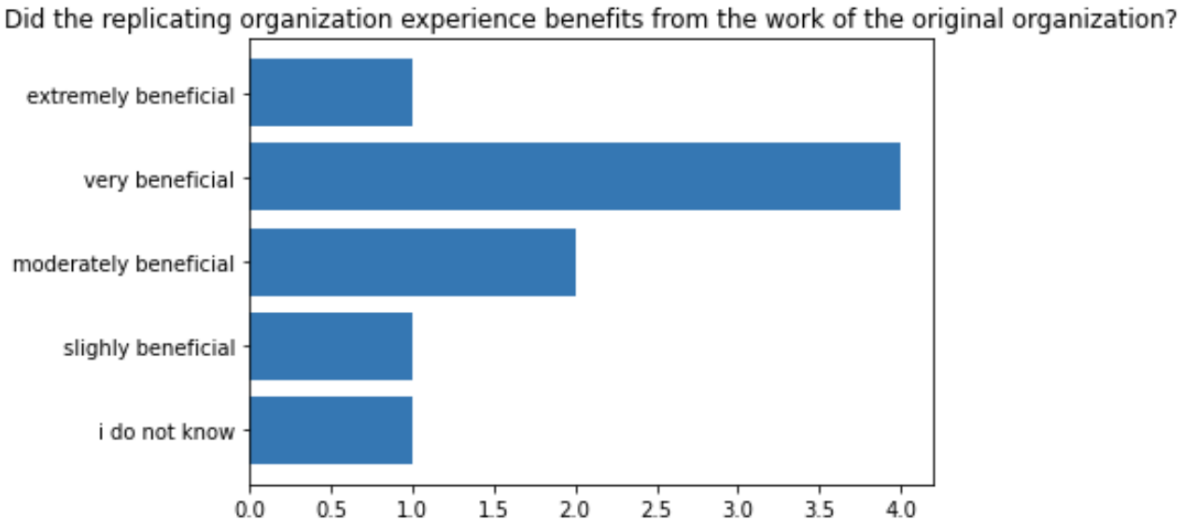

Figure 2: Is replicating beneficial for the replicating organization

Figure 2 shows the degree to which the replicating organisation perceived a benefit from the replication. 6 out of 9 respondents report the replication to have been 'very beneficial' of 'moderately beneficial'.

The respondents reported the following benefits:

- The original project can be inspiring.
- The original project can be used as a pedagogical tool to teach about the ML tools and techniques.
- Starting from an existing method leads to quick results. Especially an existing data pre-processing and model selection methodology were reported to save a lot of time.
- The time saved during the replication can be spent experimenting with potential method improvements.

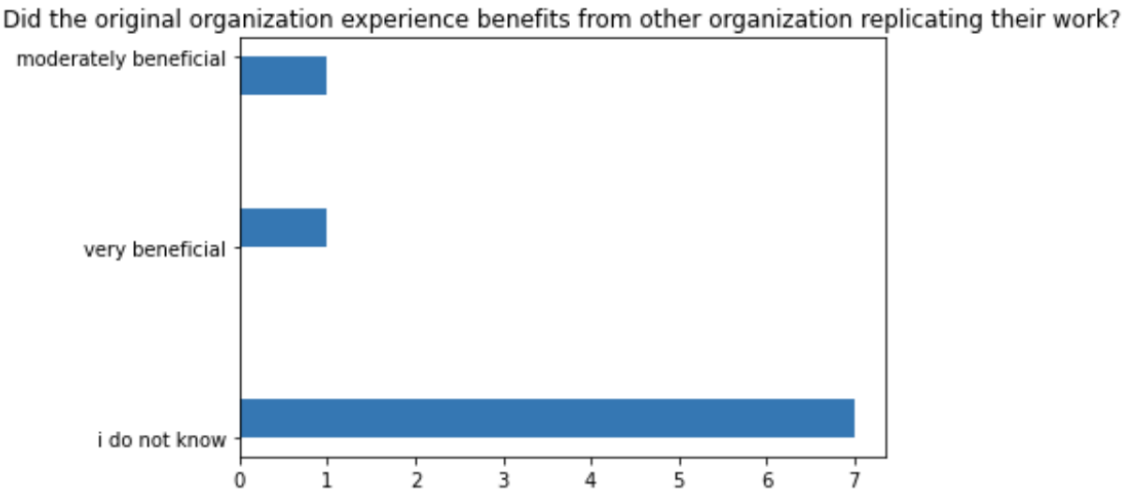

Figure 3: Is replicating beneficial for the original organisation

Figure 3 shows how beneficial the replication activity is for the original organisation. Those that have experienced a benefit for the original organisation report the following benefits:

- The original work was challenged and validated.
- The replicating organisation provided avenues for improvement to the original project.
- The developers of the original work could use the successful replication as argument to convince sceptical people within their own organisation of the validity of their work.

Figure 4: collaboration style during replication

Figure 4 shows the types of collaboration between the target and original organisation. In 2 cases the original organisation was not aware of the replication. 5 respondents report on having at least notified the original organisation of the replication activity. In 3 cases, the original organisation was kept up to date of the replication progress and results. In 3 cases the original organisation provided feedback to the replication process.

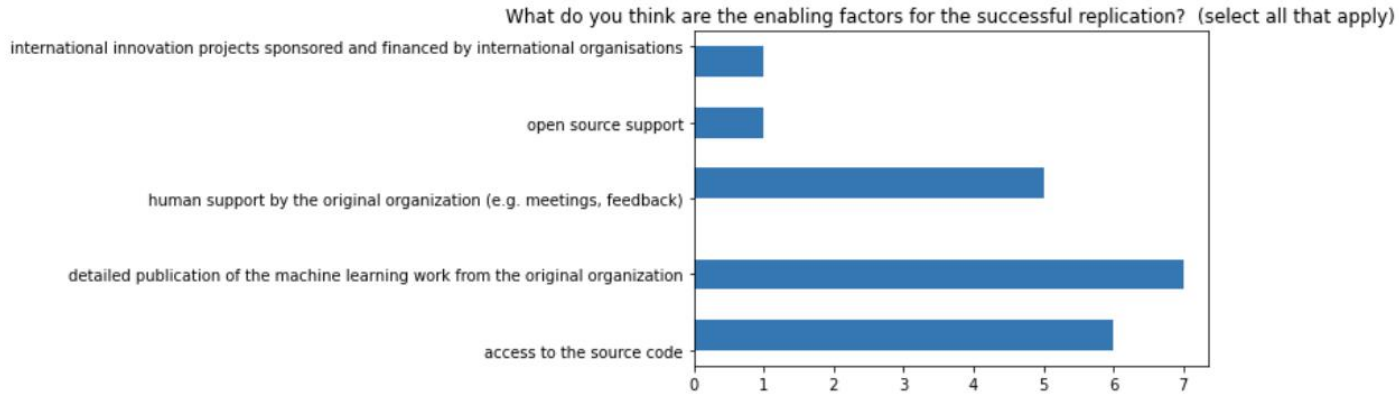

Figure 5: Enabling factors for successful replication

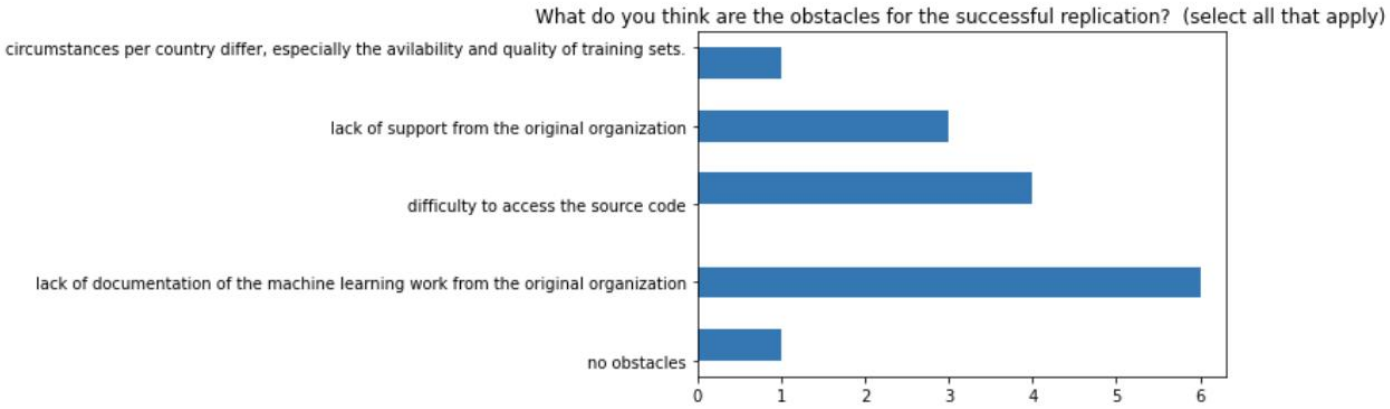

Figure 6: Obstacles for successful replication

Figures 5 and 6 shows the respondent's thoughts on the enabling and limiting factors for successful replication. By far the most enabling factor seems the be the existence of a detailed publication, followed closely by access to the original source code and human support by the original organisation. The biggest perceived obstacles to replication are opposite to the biggest enablers: the lack of documentation and source code of the original work and lack of support from the original organisation.

Finding successful applications of ML

This survey also aims to find out if there is a need for a structured repository of (un)successful ML projects. Such a repository could serve as a source of inspiration for statistical offices looking to see where ML can bring an added value to their work.

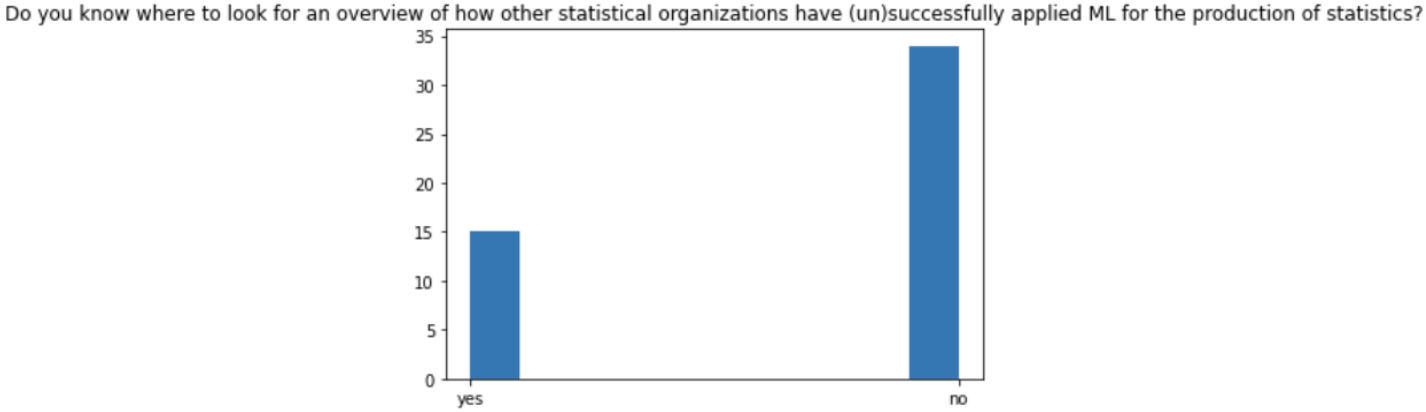

Figure 7: **Do respondents know where to look for successful ML applications?**

Figure 7 clearly shows that a lot of people do not know where to look to find successful ML applications. This shows that there is either a lack of knowledge around existing repositories, or that a suitable repository does not exist. Out of 49 respondents, only 15 report they know where to look. These 15 respondents go to the following resources when searching for successful ML applications in official statistics:

- UN HLG MOS/ML 2021 wiki
- Essnet
- Having key contacts in other organisations
- A survey done by researchers from Destatis: https://arxiv.org/abs/1812.10422 . Although the respondent mentioning this states it could use an update.
- International conferences (dach, dgins, eurosdr, ntts)
- Eurostat, for example the webpage on experimental statistics (https://ec.europa.eu/eurostat/web/experimental-statistics)
- Global network of data officers and statisticians (https://unstats.un.org/capacity-development/global-network-of-data-officers-and-statisticians)
- Websites of other national statistics offices

These responses show that there are various sources containing reports on ML applications, but that many people either are not aware of them, or that they do not suffice.

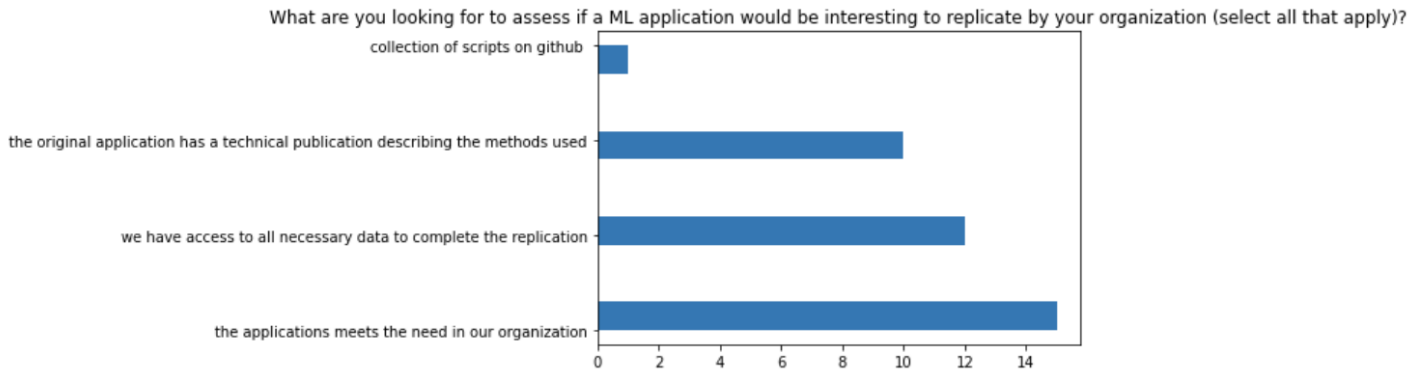

Figure 8: Choosing what to replicate

Figure 8 shows that when assessing the replication potential of other organisation's ML projects, the main consideration for choosing to replicate is that the ML application meets an organisational need. The second most important consideration is the existence of the required data in the target organisation. Third, a technical publication of the methods used by the original organisation is reported. A collection of scripts on Github, containing the actual programs used, seems less of a consideration.

Opportunities for replication

This section discusses the number of organisations that have already applied ML for official statistics, and their thoughts on being replicated.

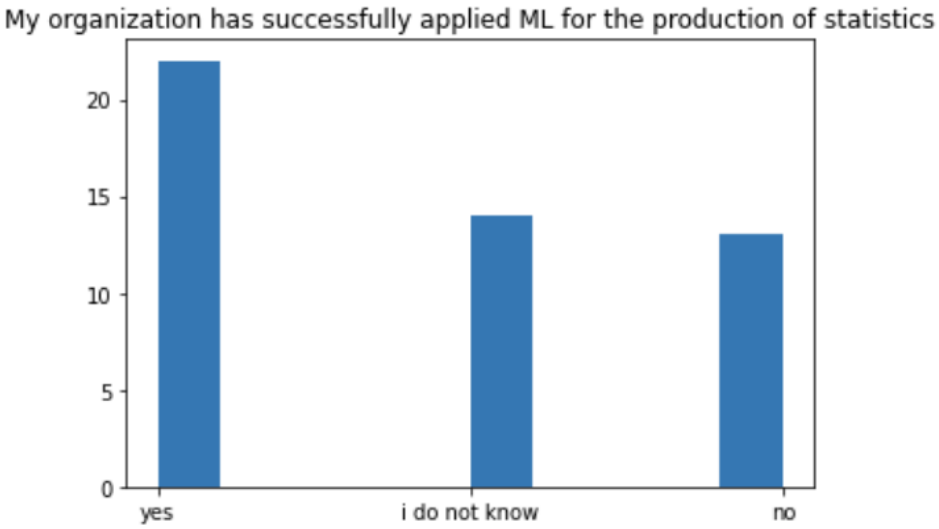

Figure 9: ML experience

Figure 9 shows that 22 respondents work for an organisation that already successfully applied ML to produce statistics. Figure 10 shows that out of these 22 respondents, 14 report that their ML application could be replicated by other organisations.

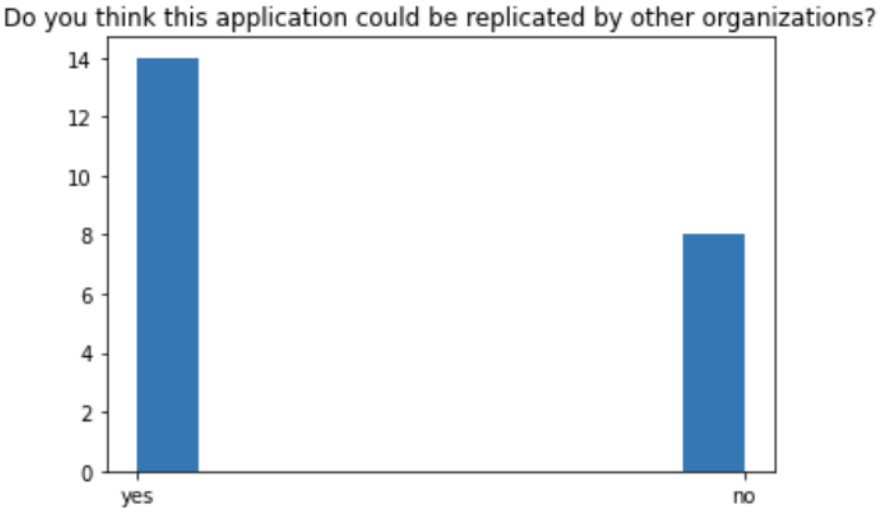

Figure 10: Organisations having replicable ML projects

Respondents who report that their application could not be replicated provide the following reason:

- There is no technical publication of the solution yet
- The solution depends on confidential data, or data that is behind a paywall
- The ML solution only makes sense as part of a wider application
- The solution needs to be improved before another organisation's should replicate it

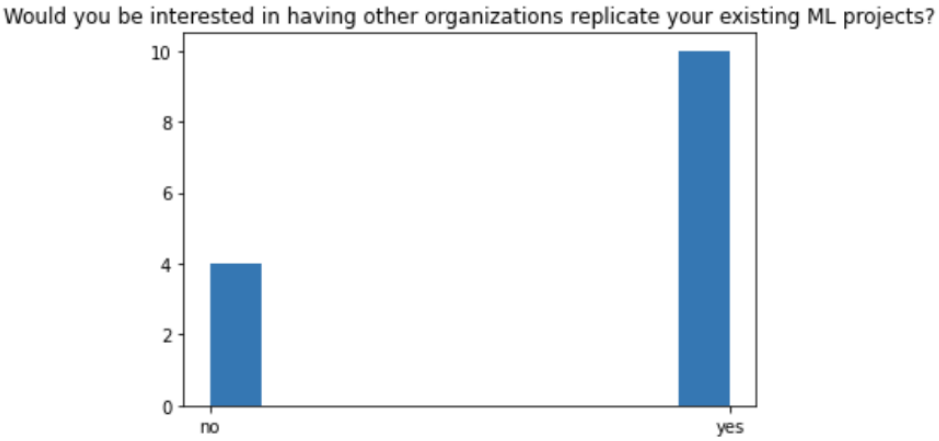

Figure 11: Interest in being replicated

Out of the 14 respondents who think their ML solution could be replicated by another organisation, 10 report to be interested in having their work replicated. Appendix 2. shows a description of those projects that can be replicated, and of which the original organisation is interested in having it replicated.

Those that are not interested in having their project replicated report the following reasons:

- The solution is very simple, so very little would be gained by using the original code.
- There exist better solutions than the one we have implemented
- The respondent did not develop the solution themself
- The respondent is not in a position to decide if their organisation is able to work with other organisations on replication of their solution.

Conclusions and lessons learned

In summary, the results presented above result in the following insights:

- Most organisations do not replicate other organisations ML work and are not having their ML work replicated by others.
- Replication of ML projects has clear benefits to both original and target organisations
- The results above shed light on what enables and what prevents successful replication. By far the most enabling factor seems the be the existence of a detailed publication, followed closely by access to the original source code and human support by the original organisation.
- Although some resources exist, most respondents do not know where to look to find (un)successful ML applications.

In the future (for example during the continuation of ML2021) the following actions could be taken based on these insights:

- Motivate the replication of existing solutions. Appendix 2 can serve as inspiration for potential projects to replicate. We expect replication to result in faster adoption of ML in statistical offices. Furthermore, by having different organisations the same ML methodology, work can be done on analysing the comparativeness of statistics based on ML.
- Create a repository of (un)successful ML projects, that can serve as a basis for replication.
- Provide some guidance on solutions to data availability (Privacy Preserving Techniques, third party contracts, ...)
- ...

Appendix

Appendix 1: description of replicated projects

- Classification of web data into COICOP classes
- 'the replication was not exact. comparison of outcomes from models were used to reduce the search criteria necessary to select the optimal classifiers. complex processes/pipelines from other nsos were also used as examples and directly replicated in the first phase of research towards developing new methods. discussions were held with other nsos during development to maximize the benefit from the work done to replicate their research. '
- 'in one of our covid-related projects, we used the hierarchical bayesian modeling to reproduce the effects of non-pharmaceutical interventions on province-wide spread of the disease (estimating the time-varying reproduction number at provincial level) and then generated death-count forecast using the fitted model. in that project, we replicated the work of flaxman et al. (imperial college), which is available on github under mit license (open source). with that said, for majority of our projects we use open source tools ranging from python/r packages and libraries related to various ml models and neural network architectures to different visualization tools, web applications and mlops tools (e.g. continuous development and integration libraries, docker, kubernetes, etc). we are also releasing and open sourcing some of our in-house developed code for various projects.'
- experimental twitter sentiment statistic
- web scraping company's websites + ml classification to get a more complete view of the number of innovative companies in the region
- text mining algorithms and other (image) feature detection software was used and are being used. the originating software is from public repositories and the authors are from a large variety of international academic institutions or research institutes.
- actually the replication was on both sides, more like a cooperation, but your questions do not allow me to fill in that. So it started with an idea to derive enterprise characteristics from web data based on a training set, by istat. they did some early experiments. then we replicated, improved and extended the technique, merging knowledge back into the international project. various known ml methods were used. training sets are the bottleneck.'
- scrapping products data from pharmacies

Appendix 2: description of projects that the original organisation would be interested in having others replicate

- Web scraping for prices
- automated coding improvement of our search tool
- automated detection of urban and rural areas
- automated identification of graveyards
- gender identification in occupations
- text classification for economic activity data
- automated coding of occupation classifier with neural networks in household surveys
- classification of industry, products, ... (using fasttext, xgboost, ...)
- classification of survey comments, narrative text (using lstm, lda, ...)
- classify or identify information on satellite images (neural network)
- model covid-19 related information such as occupancy forecast (hierarchical bayesian modeling), learning intervention strategy (re-inforcement learning using markov decision process), informing personal protective equipment (epidemiology modeling), identifying pandemic hubs (lstm)
- model crop yield during in-season (xg-boost)
- read pdf documents, newspaper articles and extract key information (many techniques are used)
- Creation twitter sentiment statistics
- Estimating response probabilities and feature selection for a survey by using random forests and cross validation. the goal is mainly to reduce overfitting of a more standard greg approach. can also be used to improve future sampling strategies.
- Generic software for finding websites of enterprises using a search engine and machine learning. ('see https://github.com/snstatcomp/urlfinding')