[> Google Drive Folder](https://drive.google.com/drive/folders/129PC7vqXIrqhV777CdZbLs7VQW5lFMok)

Meeting Notes

# **5 Feb 2021**

## **Discussion**

1. How to organise related work?
   1. Klaus: Some related work I’ve already found
   2. Use a google doc to list all the papers, with link to the pdf version.
2. Use case:
   1. Find out the common issues among the use cases and use them as the targets/requirements for our work
      1. Grid search:
         1. Complexity:
         2. Time consuming for measure the performance of each parameter setting
         3. TensorFlow has built-in support for grid search: not sure before or after training
3. Development options:
   1. Provenance collection - provbook/provlab (focus on jupyter for now and consider more general support later)
      1. Jupyter is very common for ML tutorials
      2. Extension for provenance visualisation as a jupyter extension
      3. Is there any way to support users not using jupyter?
         1. Jupyter may be used during the learning of how to use ML methods
   2. Data format: Trrack? <https://vdl.sci.utah.edu/blog/2020/10/28/trrack/>.
      1. Doesn’t seem to provide much benefits for developers
      2. Still require a lot work from the developer, for example need to record every changes
      3. Storage model: snapshot difference. Not good for search? What about git model?
      4. The state capture is defined by each application: how to ensure that all the necessary information is captured?
      5. Sharing by URL: will it make the URL very long to store all the provenance information for a state (how about the source data?)?
   3. Front end: javascript/typescript for jupyter extension
      1. Klaus like to his extension:   <https://github.com/jku-vds-lab/jupyterlab_nbprovenance>

## **To do**

* Sheeba: talk to the student working on use case 1 to see if she can show us how she does grid search
* Kai/Klause/Michael: literature review on existing work on supporting grid search
* Klause: talk to AI/Machine Learning bachelor students about how they use grid search and whether they use notebooks
* Sheeba/Klaus/Kai: compare the way to save the provenance: jupyter native format, trrack,or both

# **19 Jan 2021**

Kai, Michael, Sheeba, Sneha, Marc, Klaus

## **Discussion**

### **Introduction**

Michael: just started PhD, direction using visual analytics to support machine learning

Sheeba: slides about provbook, reproduceme, provenance ontology, and other work from PhD

Klaus: jupyter lab provenance plugin

Marc: provectories

Kai: openml.org for potential dataset

### **Use case**

1. Hyperparameter tuning
   1. Need to be more specific
   2. Use case 1: machine learning/NLP for gentrification (Michael and Kai)
   3. Use case 2: computer vision (Sheeba)
   4. Use case 3: GAN for design transfer (Sneha)
2. Feature engineering: is even broader than hyperparameter tuning; maybe as the second stage

### **Next meeting: 5 Feb 2021 - 2pm CEST/ 1pm UK**

## **To do**

* Kai: create the shared google folder
* Kai/Klaus: find related papers and add them to the ‘related work’ folder
* Kai/Michael: create a description of use case 1: machine learning/NLP for gentrification
* Sheeba: create a description of use case 2: computer vision
* Sneha: create a description of use case 3: GAN for design transfer

### **Use case template**

* Title
* Contact person
* Analysis goal (what the user wants to achieve)
* Dataset, including data type and size
* Machine learning methods and workflow
* Current issues, particularly those related to hyperparameter tuning

**Use case 1 (finished the Master project, continue the work in his PhD work)**

* **Title**: Evaluating global patterns of leaf area using Machine Learning approaches
* **Contact person**: Vamsi Kommineni, Jena, Germany
* **Description**: Leaf traits are important and often used to understand the plant and functional diversity but the numbers of leaf trait values are still strongly limited in space and time. In contrast, leaf trait values are measured from Digital Herbarium Specimen images with the help of semi-automatic image recognition tool TraitEx. Different environmental data sets like WorldClim, TerraClim, CMIP5 etc are used for extracting corresponding environmental data for the specific Digital Herbarium Specimen. Finally we investigated environmental drivers of leaf area by taking various measures like features related to spatial variation in environment (soil data, current climate data), features related to climate change over temporal domains (change in climate data sets) and features representing the season (day of year) into the account using machine learning techniques.
* **Analysis goal (what the user wants to achieve):** Building machine learning models to investigate the intraspecific variability of leaf area for three species to environmental data, climate change data and seasonal data.
* **Dataset, including data type and size:** We used the custom data sets - shape of data set 1 - 1750\*69, shape of data set 2 - 2545\*69, shape of data set 3 - 1951\*69 -datatype is Pandas DataFrames contains int and float.
* **Machine learning methods and workflow:** Dependant variable: Leaf area, Independent variables: remaining 68 environmental variables

1. Visualize the data
2. Split train and test data
3. Creating data pipeline for preprocessing the data (scaling data if required, missing values, outliers etc)
4. Building basic models with random forests and PLS regression
5. Feature selection
6. Hyperparameter tuning (with random forest) - RadomizedSearchCV and GridSearchCV
7. Model Evaluation

* Current issues, particularly those related to hyperparameter tuning:

1. Computational Complexity for GridSearchCV is more compared to RandomizedSearchCV
2. Time-consuming
3. If you have no idea what value a hyperparameter should have, a simple approach is to try out the consecutive powers of 10.

**Use case 2 (finished)**

* **Title**: Optimization methods analysis
* **Contact person:** Nora Abdelmageed, Jena, Germany
* **Analysis goal (what the user wants to achieve):** Create a neural network model and try different optimization methods then decide on which one of them is the best in terms of accuracy metric. The selected optimization methods are: stochastic gradient descent (SGD), Momentum, RMSProp and Adam.
* **Dataset, including data type and size**: [Moon dataset](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_moons.html) (toy dataset as a part of sklearn library) . The used configuration created a dataset with a shape of (300,2)
* **Machine learning methods and workflow**: Panda, which means, we create a single neural network at a time and associated with one optimization method, record the accuracy and repeat
* **Current issues, particularly those related to hyperparameter tuning**
  1. Time consumption

**Use case 3 (planning stage)**

* **Title:** Creating knowledge graph using BERT
* **Contact person:** Fusion group, Jena, Germany
* **Current issues, particularly those related to hyperparameter tuning**
  1. Planning to use the state of the art hyperparameters
     + Known to have good performance
     + Require a lot of efforts to search for new set of parameters

**Use case 4 (finished first stage, potential for further extension)**

* **Title:** Use of GAN and Cycle-GAN for style transfer
* **Contact person:** Fusion group , Jena , Germany
* **Current issues, particularly those related to hyperparameter tuning**
  1. Hyperparameters :
     + Using state of art parameters? Will add the names of the parameters later.
  2. Datasets : (Architecture and Portraits)

1. Train A : Bauhaus style images
2. Train B : Non-Bauhaus style images

      C. Test A : Another set of Bauhaus style images

      D. Test B: Another set of Non-Bauhaus style images

* 1. Changes to parameters :
     + Changes to loss function ( add and remove identity loss ) ,
     + Change the number of epochs for training
     + Change the learning rate , so that initially the learning rate is higher and once the network is trained, the learning rate automatically decreases and flattens.
     + Batch size
     + Training time is one of the factors that need to be considered.

Michael’s interview with a data scientist:

I have conducted a user interview from a data scientist/machine learning specialist to understand the problems user face. The conducted interview was structured and the interview questions were devised based on Eric Migicovsky's guide on "How to Talk to Users" suggested by Mr. Kai. While asking the questions and recording the responses, I have tried best to have the deep understanding of user needs.

User interview/response:Participant: 1 Interview length: 15 minutes

Response Summary:

Question: What is the occupation of the interviewee?

Interviewee: is a full-time working employee as a data scientist and a researcher. He is also a graduate student doing specialization in Imaging and Machine Learning. He also has various research papers published in reputable conferences and Journals.

Question: What kind of data have you worked on?

Interviewee:I have worked on Imaging Datasets including Medical, X-RAY, CT scan and various Tabular Datasets including Heart Disease datasets and Cancer Detection datasets.

Question: What kind of analysis have you done in your career uptil now?

Interviewee: I have worked on Semantic Image Segmentation to segment lungs imagery from Chest X-RAY, Neural Network Image Classification, parameter value selection (hyperparameter selection) and various Clustering methods.

Question: What tools and softwares have you used?

Interviewee:For data visualization and business intelligence, I have used Tableau and Power BI. For hyperparameter optimization and data pre-processing, I have used various Python machine learning libraries including scikit-learn, Tensorflow, NumPy, Pandas, Keras, SciPy and PyTorch.

Question: What basic steps you take when you have a machine learning problem to solve?

Interviewee:The first step is to analyze the problem in a sense to figure out the input and the output. Once it is done, I see which model is best suited for the problem and make a list of them from model of least complexity to model of highest complexity. Then I check which type of data should I use, on basis of my inputs and outputs I search if I have data available online or should I need to collect it myself. Then I preprocess data to be fed to model and train the least complex model to check the performance and move up to the ladder and stop where I get the desired result without exhausting my resources.

Question: What is the most painful process you think in training a model?

Interviewee:  According to my experience, the most painful process would be to prepare and pre-process the data. Sometimes I collect the data and it is impossible to pre-process it according to my needs. Or if I pre-process it and extract features, then these features are not correct for getting desired output and the accuracy turns out to be bad.

Question: Are you aware of hyper-parameter tuning?

Interviewee:Sure, I am. It is one of the important things in ML to get optimal solution by controlling the learning process

Question: What is the most painful thing about hyper-parameter tuning?

Interviewee:Whenever I build a model, I want the solution to be optimal, but there are a lot of parameters that needs to be tuned to get that optimal solution. So, I have to try all the possible values to get a better fit for my model and dataset. Even if I tune the model, whenever I change the data, accuracy decreases and I have to tune the parameter values again.

Question: What do you think is the most troublesome step or process for you as a machine learning specialist?

Interviewee: Personally, I find the data pre-processing most troublesome in the process of training a model.

Question: What tools do you use for pre-processing and hyper-parameter tuning?

Interviewee: I have not used any built-in software or tool to do that yet. For hyper-parameter tuning I have used sklearn libraries, hit a trial and sometimes my own experience in the field helps me for simple models.

Question: Why do you think pre-processing is more painful than hyper-parameter tuning?

Interviewee:  There are many models which apply brute force to get optimal hyper-parameter values for optimal accuracy. Moreover, I know that there is some optimal value but in case of pre-processing I’m never sure, the data can be wrong or the pre-processing technique I’m using can be wrong. In my career so far, I have come across problem of hyper-parameter tuning but I’ve solved it easily, but pre-processing is always time consuming.