The report of project W3: Kaggle-Predicting transparent conductors

Business understanding

Identification of business goals

Background

Transparent, conductive materials are a unique class of compounds, which have both: electrical conductivity and low absorbance in the visible range. These materials can be used in a large number of applications such as energy-conserving windows and solar cells.

Business goals

In spite of appealing properties relatively low number of compounds possess such properties. As the number of possible materials is large, to select compounds for experimental verification, computational screening of candidates is necessary. Lowering the cost of screening and increasing its accuracy, would benefit the discovery of these materials by reducing the development costs.

Business success criteria

As a success, we can count faster and more effective development of transparent conductors. This criterion can be verified by the number of publications presenting novel materials, which possess necessary properties.

Situation assessment

Inventory of resources

- Two university students
- Supervisors, who can provide additional guidance if necessary
- Python Data Science Platform Anaconda including SKlearn, Pandas and Numpy packages
- Dataset provided by Kaggle competition

Requirements, assumptions, and constraints

The requirements:

- At least 30 hours of work per student
- The project idea must be presented in a practice session during Nov 11-13
- The completed project must be presented in a poster session on Dec 19
- The project code repository must be accessible to instructors.
- As the project's topic is from Kaggle, all available kernels that have been used in the project must be declared

The assumptions:

- The objectives of the project are clear
- Relevant data science methods are used and described in the project

The constraints:

- Neither of the students is expert in the field

Risks and contingencies

The risks with contingencies:

- The data and approaches used for training fail to produce an accurate model
  - Study literature to find additional parameters, which can be relevant for predicting the properties
  - Consult with the lecturers from Chemistry or Computer Science Institute, who can give ideas
- The project is not finished by the deadline due to unexpectedly large workload
  - Implement ideas and approaches taken from the kernels that are uploaded to Kaggle competition. The used kernels must be declared
  - This risk can be mitigated by planning the tasks of the project and setting milestones
- The project code and data is deleted by mistake
  - The code in Github is regularly updated with the latest changes
  - Each team member keeps a copy of the project in their local hard drive

Terminology

- **Transparent conductors** - optically transparent and electrically conductive materials
- **Alloy** - Combination of a metal and one or more other elements
- **Band gap** - Range of energy in solid where no electron states exist.
- **Formation energy** - The energy required or released during the formation of a compound from its constituent elements in their standard state.

- **Unit cell** - The smallest group of particles in the material that constitutes a repeating pattern
- **Space group** - The symmetry group of a configuration of particles in crystal.
- **Coordination number** - The number of closest neighbouring atoms for a defined atom
- **Partial charge** - Charge on a specific atom. Can be used to quantify the degree of ionicity in crystal
- **Electronegativity** - a chemical property that describes the tendency of an atom to attract electrons towards itself
- **Ionization potential** - the minimum amount of energy required to remove the valence electron from an atom
- **Electron affinity** - The amount of energy released when an electron is added to an atom
- **Packing fraction** - the fraction of unit cell volume in a crystal structure that is occupied by the particles

Costs and benefits

The direct and indirect costs are insignificant as the two students working on the project are doing this without pay and work at university of from their dorms. Although if there are sufficient interest, more elaborate methods for evaluating different attributes can be used. In order to do so obtaining computation time from HPC cluster would be necessary.

Benefits of the projects are mostly educational, as both students widen their knowledge and expertise both in solid-state physics/chemistry and in data science. If the model developed by the team is successful, analogical models could be developed to be used in scientific work in the Institute of Chemistry.

Definition of data-mining goals

Data-mining goals

The screening of compounds is currently done using computational methods, which require large computational resources. Finding a machine learning model, which can accurately predict compounds formation energy and band gap, would benefit the discovery and development of these materials.

Data-mining success criteria

With the current data-mining project, we count as a success if we have developed machine-learning model(s), which can predict with at least 90% accuracy the formation energies and band gaps. The compounds, selected by the developed model narrows the number of candidates, which can be then studied with a more expensive computational model.

Data understanding

Gathering data

Outline of data requirements

In order to estimate the band gap and the stability of the material, we need to obtain data about its composition, structure and the interactions between the atoms in the crystal. In order to characterize these qualities we have chosen to use the following attributes:

- the symmetry group of the crystal,
- the number of atoms in a unit cell,
- the relative composition of the cell
- the distances between each pair of atom types (bond lengths)
- the coordination numbers for each pair of atom types (e.g. how many Ga atoms are in the equal nearest distance to Al atom
- the partial charges of atoms
- the packing parameter of the unit cell (how densely are the atoms in a unit cell situated)

Verification of data availability

The original data given to the participants of the Nomad 2018 competition is available at the competitions Kaggle page. As the attributes given by the Kaggle cannot comprehensively describe aforementioned qualities and additional data about so large number of compounds is difficult to collect, we have chosen to calculate other attributes by using the knowledge in theoretical chemistry. The methods, constants and additional quantities to be used are taken from:

- https://en.wikipedia.org/wiki/Ionization_energies_of_the_elements_(data_page)
- https://en.wikipedia.org/wiki/Electron_affinity_(data_page)
- https://wiki.fysik.dtu.dk/asap/Radial%20Distribution%20Functions#partial-rdfs-looking-at-specific-elements-or-subsets-of-atoms
- Pearson, Ralph G. "Absolute electronegativity and hardness: application to inorganic chemistry." Inorganic chemistry 27.4 (1988): 734-740.
- Bultinck, Patrick, et al. "The electronegativity equalization method I: Parametrization and validation for atomic charge calculations." The Journal of Physical Chemistry A 106.34 (2002): 7887-7894.

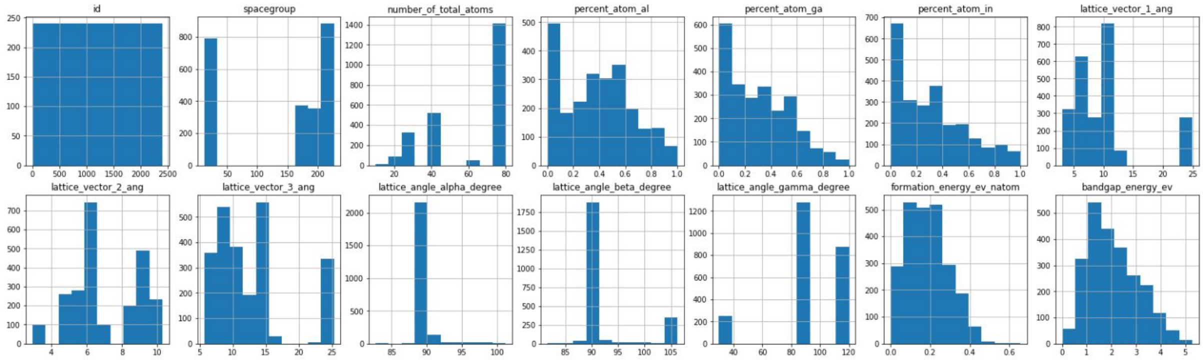
Data description

From the Kaggle competition, we have been given data about 3000 materials. The data contains each compound’s space group, total number of aluminium (Al), gallium (Ga), indium (In) and oxygen (O) atoms in the unit cell, relative composition of Al, Ga, and In, along with lattice vectors and angles, which are all stored in a .csv file. The data in the Kaggle .csv file can be used to form some idea how the atoms are located in the unit cell but in itself don’t describe how specific atoms are located relative to each other and what kind of chemical environment are they situated in.

The dataset also contains geometry files (.xyz) of unit cells for each compound, where are written its atoms’ element and position in xyz-coordinates, relative to each other. The geometries can be used to additionally calculate attributes using additional constants and quantities (e.g. ionization potential of atom), which we take from data tables because we did not find such information about all the compounds that have been given in Kaggle and decided not to search for the data that might be available in notebooks uploaded to other Kaggle by other users.

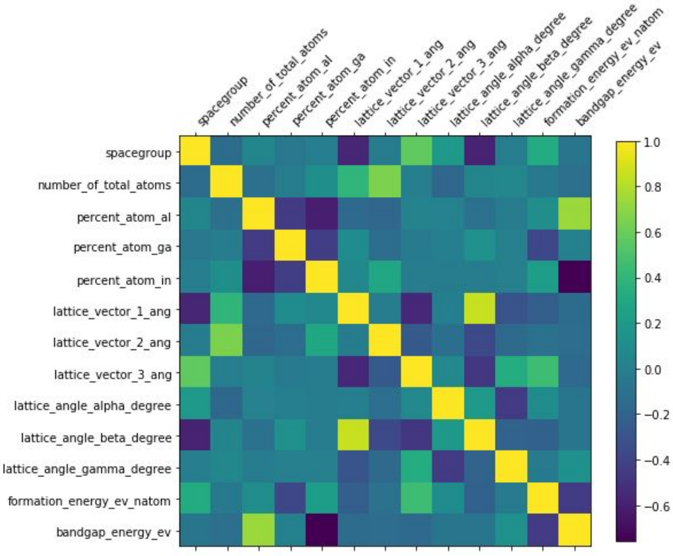
Data exploration

First of all, let’s look at the training data (2400 instances in total) from the Kaggle competition.



Histograms of features.

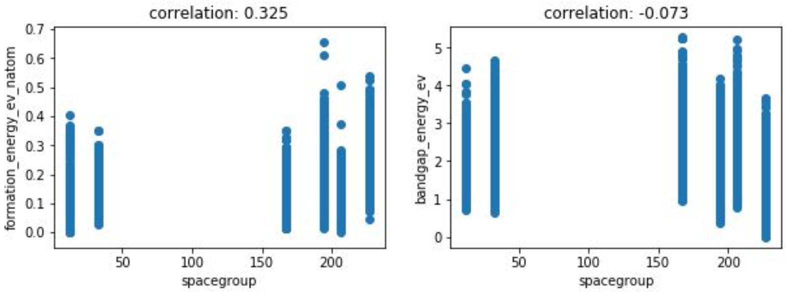
We can already remove the id column because it provides no useful information for our model. Something to note is that the features have a varying scale so it might be beneficial to standardize the features.
Luckily none of the features in the Kaggle dataset contained any missing values so we don’t have to worry about that.

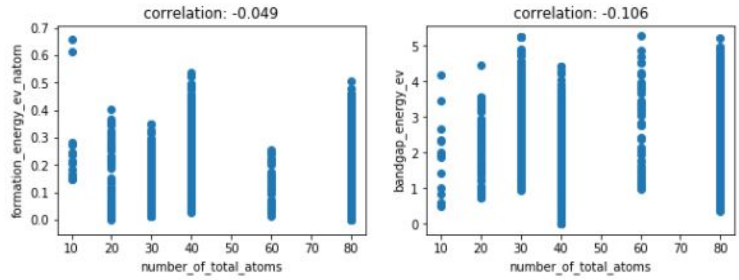


Correlation matrix.

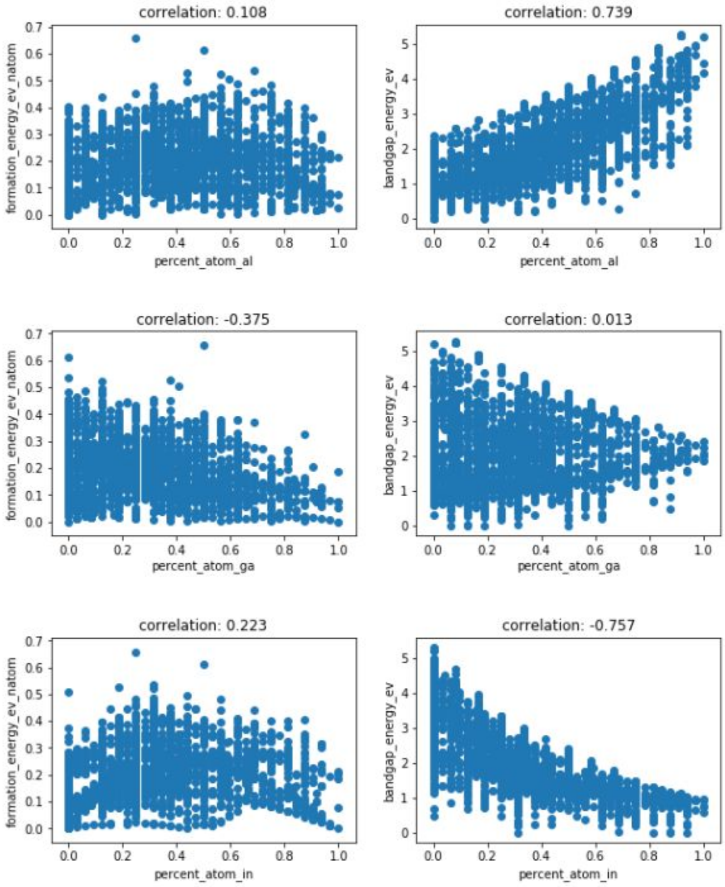
If we look at the correlation matrix we can see that there aren’t many features that are correlated with formation energy or band gap energy (the target values).

Look at the individual features:

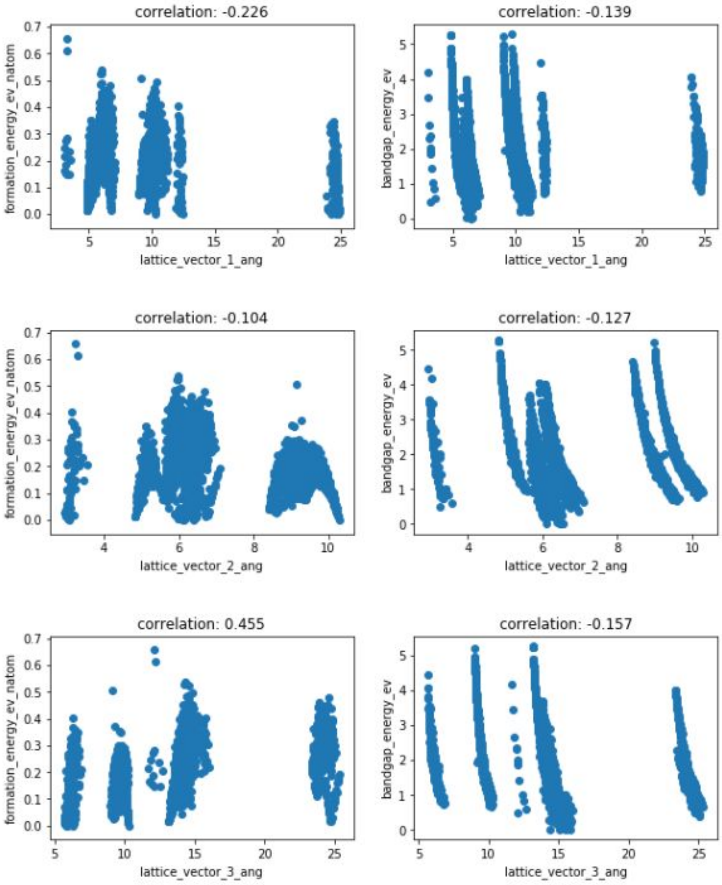
- **spacegroup**
    - We can see that the correlation values for it are quite low both with formation energy and band gap energy
    - Since spacegroup represents a complex concept, representing it by a numerical encoding is not very useful. It is worth trying grouping values with similar crystal systems and encoding them with one-hot encoding because spacegroup numbers are not in a particular order and it doesn't make sense to encode it numerically.
    - It is also worth generating some features from the spacegroup such as a binary feature of whether the crystal is symmetrical.
    - 

- **number_of_total_atoms**
    - This feature doesn't seem to be important right now. It has a low correlation with both of our targets.
    - 

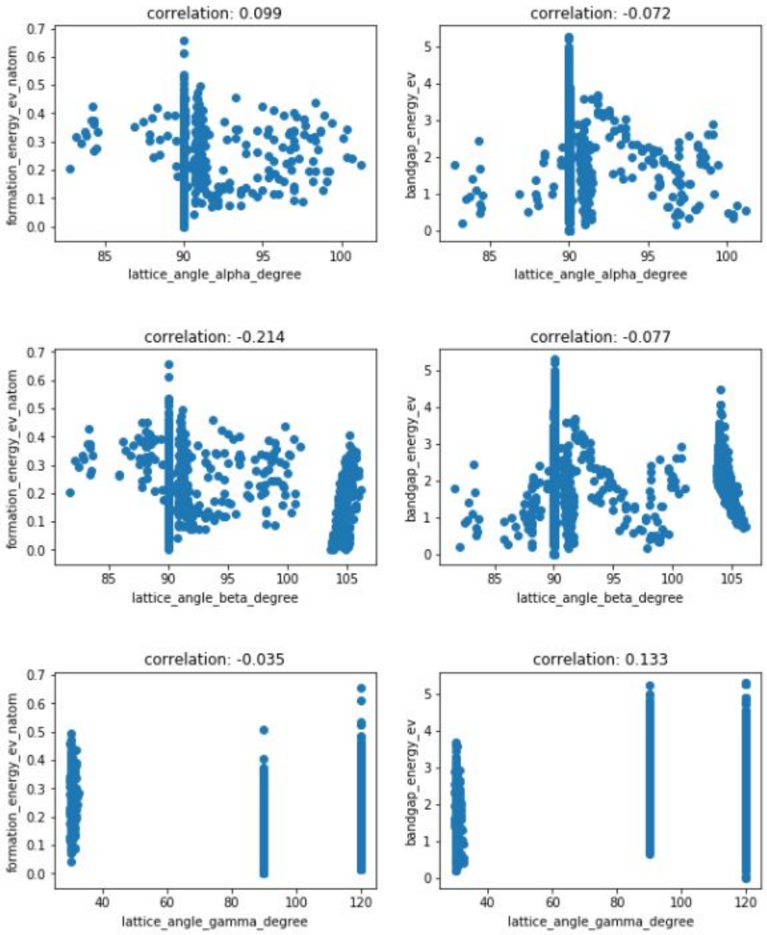
- **percent_of_atom_al, percent_of_atom_ga, percent_of_atom_in**
    - percent_of_atom_al and percent_of_atom_in seem to be somewhat correlated with band gap energy, while the correlation of percent_of_atom_ga doesn't seem to have any correlation with band gap energy.
    - percent_of_atom_ga is somewhat correlated with formation energy, while the other percentages don't show any significant correlation with this target



    -

- **lattice_vector_1_ang, lattice_vector_2_ang, lattice_vector_3_ang**
    - Lattice vectors could also be somewhat useful because it seems that they are somewhat correlated with the target values.
    - 

- **lattice_angle_alpha_degree, lattice_angle_beta_degree, lattice_angle_gamma_degree**
    - There doesn't seem to be any clear relationship between these features and the target values. These might not be useful by themselves and might need some transformation.

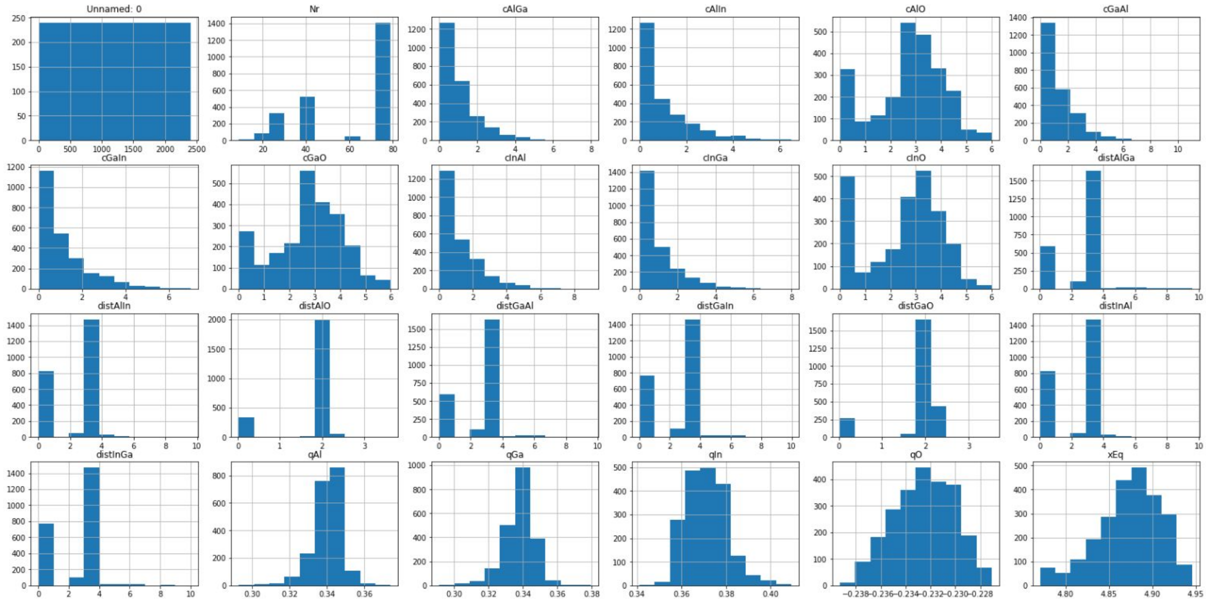


○

Targets:

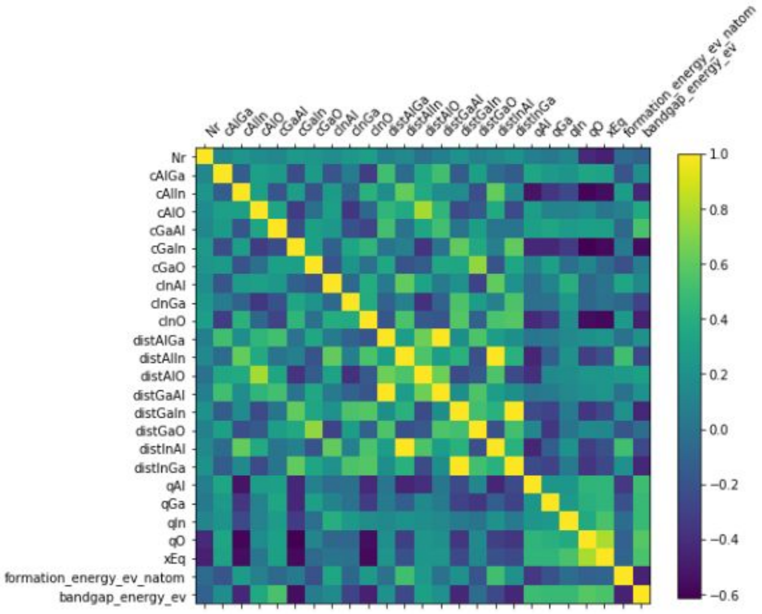
- **formation_energy_ev_natom**
- **bandgap_energy_ev**

We have also already engineered 23 new features:



Histograms of engineered features. 'Unnamed: 0' is just an id column and can be dropped.

Some of the generated featured need further engineering because they contain missing values. Here we also need to consider standardizing the values to transform them to a similar scale. The quality of the calculated properties for each material can be disputable, as we use methods that are modest in computational cost but due to the use of empirical parameters, their suitability is not verified for given compounds and therefore can give mixed results.



Correlation Matrix of engineered features and target values (last 2 columns).

We can see that the engineered features contain more features that have a significant correlation than the data from Kaggle.

Data quality verification

The data from Kaggle competition has good quality. As noted previously none of the features was missing (had null/nan value). Some features need standardization and different encodings to make them useful. Also, we might need to generate new features from some of the more complex features to make them more useful for training a model.

We have already generated 23 new features from the data and we are planning to engineer more features. The task we have is quite complex, but we have enough good quality data to solve it and it’s not too complicated to engineer new features from the data we have when we need them

Project plan

Accomplished Tasks:

1. **Familiarizing with the literature.** Reading the literature has given us the ideas to calculate attributes such as coordination numbers and partial charges.

   Taido: 2 h

   Heigo: 3 h

2. **Programming the methods to calculate physical attributes from scratch or searching for available implementations and adapting them.** This task involved implementing reading in geometries, calculating partial charges and radial distribution function.

   Heigo: 5 h

3. **Preliminary data analysis.** Verify that the attributes we have chosen to engineer are relevant or different attributes need to be looked for. The results of the analysis are shown in Task 3.

   Taido: 3 h

Future tasks:

1. **Data transformation/encoding.** Some features need further encoding to make them useful for our dataset (such as spacegroup). Some of our calculated features have missing values which need to be replaced with something that makes sense in the context.
   Taido: 2 h
   Heigo: 2 h
2. **Obtaining baseline results.** Try obtaining results for comparison. Try out various regression algorithms.
   Taido: 3 h
3. **Generating additional features.** For example, generating new features from the spacegroup feature or using geometry files. In addition to encoding it, there could be useful information extracted from it. One idea is to categorize the spacegroups into symmetry classes.
   Heigo: 4 h
   Taido: 2 h
4. **Training models and parameter optimization.** We need to train two models: one for predicting generating formation energy and one for band gap energy. The parameters of both models need to be optimized. In addition to that, we need to test various encodings and possibly leave out some features that aren't useful.
   Taido: 12 h
   Heigo: 10 h
5. **Writing the final report of the results. Making a poster.** Concluding our results and designing a poster.
   Taido: 8 h
   Heigo: 8 h