# University of Helsinki
# Data Science for the Internet of Things (DS4IoT) - Spring 2024

*Home Exam*, due on 20th May 2024 by 23:59, Helsinki time

NGOC THI NGUYEN, AGUSTIN ZUNIGA, PETTERI NURMI, University of Helsinki

**ACM Reference Format:**
Ngoc Thi Nguyen, Agustin Zuniga, Petteri Nurmi. 2024. University of Helsinki. Data Science for the Internet of Things (DS4IoT) - Spring 2024: *Home Exam*, due on 20th May 2024 by 23:59, Helsinki time. 1, 1 (May 2024), 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

**Instructions:**

- All course participants are requested to submit their solutions (in English) on Moodle. If you encounter any problems with Moodle, please send an email to the course instructors Ngoc Thi Nguyen (ngoc.t.nguyen@helsinki.fi) and Petteri Nurmi (petteri.nurmi@helsinki.fi) by the due date and time.
- Please take into account the following guidelines about including AI-generated content: https://studies.helsinki.fi/instructions/article/using-ai-support-learning

**Formatting:**

- Submit the report of non-programming tasks in PDF file format. Submissions should be formatted according to the ACM master template available from https://www.acm.org/publications/proceedings-template. Please use the single column formation option for the document class. Your resulting file should be formatted similarly to how the exam questions (this document) is formatted.
- Submit the solutions and the source code of the programming tasks in Python using Jupyter Notebook format (*.ipynb*).

**Assessment:** Participants are encouraged to review course material to answer the problems and in some cases write computer programs to derive solutions. In all the tasks, do not just give the answers, but also justify, derive and contextualise the answers in Data Science for the IoT contexts. The assessment parameters for each of the tasks are: solution (25%), justification and derivation (50%), contextualisation (25%).

**Tasks 1 and 2 are compulsory for everyone.**

**Task 1: Concepts (8 pts.)**
Describe **what** the following concepts mean and **why** they are relevant to the Data Science for the Internet of Things (DS4IoT). For full points give at least one example of the concept in a DS4IoT context or otherwise illustrate their importance. Do not copy the definitions from Wikipedia or other source, but explain the concepts in your own words and relate them with DS4IoT. Use about 150–200 words to describe each of the concepts.

(1) Inter-rater reliability
(2) Spatial sampling
(3) Missing at random (MAR)

Author's address: Ngoc Thi Nguyen, Agustin Zuniga, Petteri Nurmi. University of Helsinki.

2024. Manuscript submitted to ACM

(4) Spectral feature
(5) Inter/Cross-x variability
(6) Stratified cross-validation
(7) Sensitivity (in differential privacy)
(8) Convolutional layer (in deep learning)

**Task 2: DS4IoT Sensing Pipeline (8 pts.)**

Read the following description of an DS4IoT application https://www.starship.xyz/press_releases/starship-launches-robot-grocery-delivery-services-in-finland-partners-first-with-a-leading-retail-operator-hok-elanto-group/. Answer the questions below using the described application:

(1) What/Where are the science, data science, and pervasive computing parts? How would the selected application benefit from Data Science for the Internet of Things?
(2) Design a concept map to describe the different stages of the sensing pipeline. In each stage, describe and justify the methods, techniques, models, evaluation metrics, etc. that you think they are used in the application.
(3) The following article describes an issue experienced during the first week of operation of the autonomous vehicles: https://www.bbc.com/news/uk-england-cambridgeshire-63821535. How would you improve the sensing pipeline described in (2) to prevent this issue from happening in the future? In particular, discuss how would the system be evaluated?
(4) Describe a security or privacy attack that can impact the correct performance of the application. Describe how the attack works in practice. How can the system be secured against such attacks?

**Please select ONE of the two following programming tasks.**

**Task 3: Preprocessing and Feature engineering - Programming (8 pts.)**

The tinamous are ground birds native from Central and South America. There are 46 different species of tinamous that can be found in diverse habitats up to 5000 meters. Tinamous are characterised by their particular melodic call, which can be very similar to a flute. An audio recording of a tinamou's call collected at the tropical forest can be found on Moodle (i.e.,*Tinamou.wav*). The great tinamou and the white-throated tinamou are two of the 46 species. The former inhabits Central and South American forests at altitudes between 300 and 1500 meters, and the latter inhabits in sub-tropical and tropical areas of South America at altitudes around 500 meters. The white-throated tinamou is considered as a near-threatened specie. The audio recordings of the great tinamou and the white-throated tinamou can be found on Moodle (i.e., *GreatTinamou.wav* and *WhiteThroatedTinamou.wav*). Your task is to identify which of these two tinamou species corresponds to the collected recording.

(1) Preprocess the three audio signals *(Tinamou.wav, GreatTinamou.wav, WhiteThroatedTinamou.wav)* so that calls are isolated from other environmental sounds. Plot the audio signals before and after the preprocessing. Describe and justify the method used to remove the noise of the signal.
(2) Plot spectrograms of the three audio signals after the preprocessing and answer the following questions:
   (a) Which are the frequency ranges and periods for the three tinamous calls?
   (b) Which of the two calls (great tinamou and the white-throated tinamou) is in the same frequency with the call recorded at the tropical forest?
(3) Calculate the similarity between the signals and describe the results.
   (a) Which of the two calls (great tinamou and the white-throated tinamou) has the lowest distance to the call recorded at the tropical forest?

(b) How does this result variate compared to when using spectrograms? Why?

(4) Use differential privacy with Laplace mechanism to share the white-throated tinamou audio signal. Plot the spectrograms of the audio signals before and after applying differential privacy. Justify the way you calculated the sensitivity and the selection of epsilon.

**Task 4: DS4IoT Modeling and Evaluation - Programming (8 pts.)**

The file *air_measurements.csv* contains air quality measurements collected by the Finnish Meteorological Institute (FMI) at Helsinki between the 1st April 2021 to the 30th April 2022. Your company wants to implement a model for predicting air quality using this dataset.

(1) Calculate data statistics of air quality index, $NO_2$, and $PM_{2.5}$, for each month. Then aggregate the data using different time windows (i.e., 1-hour, 12-hour, 24-hour) and describe the effect of different time windows on the monthly reports. Use data statistics and statistical plots to justify your answers. HINT: impute missing measurements using the median before calculating data statistics and aggregating the data.

(2) Implement a sensing pipeline (preprocessing, feature engineering, data modeling and evaluation) for predicting the air quality index using $PM_{2.5}$ values. Describe the sensing pipeline and justify the selection of methods used in the sensing pipeline (e.g., window size, feature selection, model selection, evaluation metrics, etc.). Report the accuracy of your model.

(3) Assume that the request changes to use $NO_2$ samples to predict the air quality index instead of $PM_{2.5}$. How does the accuracy of your model change compared to that of the model in (2)? Why? HINT: explore the the correlation between variables.

(4) Use differential privacy with Laplace mechanism to report the daily mean and max values of $NO_2$. Plot the results before and after applying differential privacy. Justify the way you calculated the sensitivity and the selection of epsilon. HINT: use 1-hour time window.

## OPTIONAL: The following question is a BONUS task that can be used to increase your points tally. This task is OPTIONAL and can result in an up to HALF A GRADE increase.

This task aims to reflect on the use of generative models to answer evaluation queries. Use a generative model-based application (e.g., ChatGPT, Copilot, Bing, etc.) to solve the following tasks of the home exam.

(1) Task 1 (all the parts)

(2) Task 2 parts: (1), (2), (4).

(3) Task 3 (4), or Task 4 (4) - depending on whether you choose Task 3 or Task 4 above

Complete the task by describing the following:

(1) Write the name of the application that you used and a brief description (max 300 words, using your own words) of the model architecture and evaluation parameters that the application uses to generate their answers.

(2) Copy the prompts you provided to the application for generating the answer to each task.

(3) Copy the answer you obtained from the application for each task.

(4) Write a reflection (450-500 words, using your own words) on the answers that you obtained from the application and the answers that you provided in the exam. The reflection should cover the following: similarities and limitations between the answers provided by the application and those provided by you in the exam, privacy concerns, and how data science for the IoT would benefit from this application.