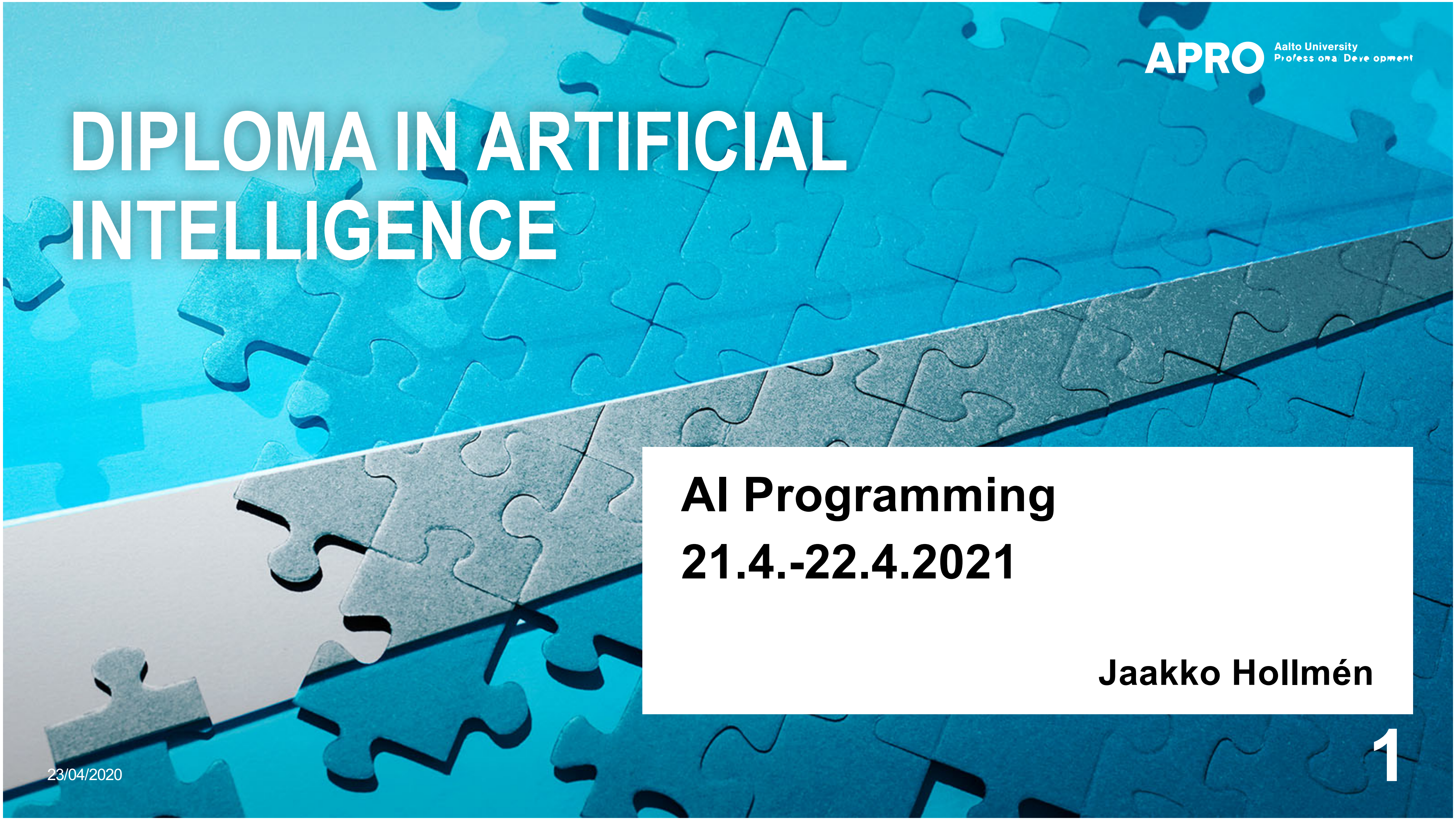

APRO Aalto University Professional Development

DIPLOMA IN ARTIFICIAL INTELLIGENCE

AI Programming

21.4.-22.4.2021

Jaakko Hollmén

AI PROGRAMMING

The two-day training is divided into thematic sessions, where problems are presented to the students and when students create the solutions to the problems during the session.

Each student should have a computer and preparedness to run Python programs in Jupyter notebooks.

Each session contains a brief introductory lecture to the topic, and description of the programming exercise. Then, student proceed by programming, either alone or in pairs. Towards the end of the session, solutions will be reviewed.

SESSION 1: DATA FORMATS, READING AND WRITING DATA IN PYTHON

- This session will be covered on the first day during 9.10-10.25.
- The learning objective is get familiar with different file formats for storing data sets, Python libraries and command for reading and writing data sets, and the data structures used to store the data in Python.

LECTURE CONTENTS

- File formats: text files, csv files, character encodings, Excel
- Databases: relational databases, JSON
- Description of the exercises

STORAGE AND FILE FORMATS

- Data sets are stored with persistent storage on hard disks or other media
- In-memory solutions are the fastest, but limited by practical considerations

TEXT FILES

- Text files are readable and easily transferrable form of data
- Typically, a relational database: each row has a record that has the fields in the schema
- Field separator may differ in the specific file: <field1><separator><field2><separator>…
- Separator (or delimiter) may be a space, or a tab, for instance
- Quite well portable from one platform to another

Python solutions:

- Open a file, read a line, do something for each line, close file
- Read the whole file to a variable

COMMA-SEPARATED VALUES

- Fields in the file may contain numbers, or sometimes strings
- Comma-separated values defines how to separate fields and how to quote strings
- Example: 002829, ”John Doe”,”plumber”,3400
- Separator is a comma, fields are within quotes when needed

Python solution

- Use the csv library (import csv)
- Use the pandas library: import pandas as pd, my_table = pd.read_csv(…)

EXCEL FILES

- Excel spreadsheet files are a common data file format
- Usually, records have the same amount of fields
- Directly readable into Python workspace

Python solution:

- Import pandas as pd, my_table = pd.read_excel(...)

RELATIONAL DATABASES

- Often, files are relational databases: there are records (as rows), fields on one row are the members of a schema
- SQL (structured query language) is used to interfacing a database
- Usually, you need a driver to connect to a particular database (SQLite, MySQL, PostgresSQL)

Not covered on this course

JSON – JAVASCRIPT OBJECT NOTATION

- JSON (Javascript Object Notation) is a data exchange format for data
- Standardized format for data-exchange: RFC 7159 and ECMA-404.
- Human-readable, allows for varying definition of records
- Example: { ”John": 5, ”Mary": 7 }
- Another example: [{”Hello”: ”moi”, ”John”: 5}, {”This”: ”that”, ”Mary”: 8}]

Python solution:

- Use the json library: import json
- Documentation: https://docs.python.org/3/library/json.html
- Pandas library read_json()

EXERCISES

- The exercises are presented in the Jupyter notebook Session-1-Data-formats.ipynb
- Work one exercise at the time
- Not all exercises need to be completed

REVIEW OF THE SOLUTIONS

- How do the solutions look like?