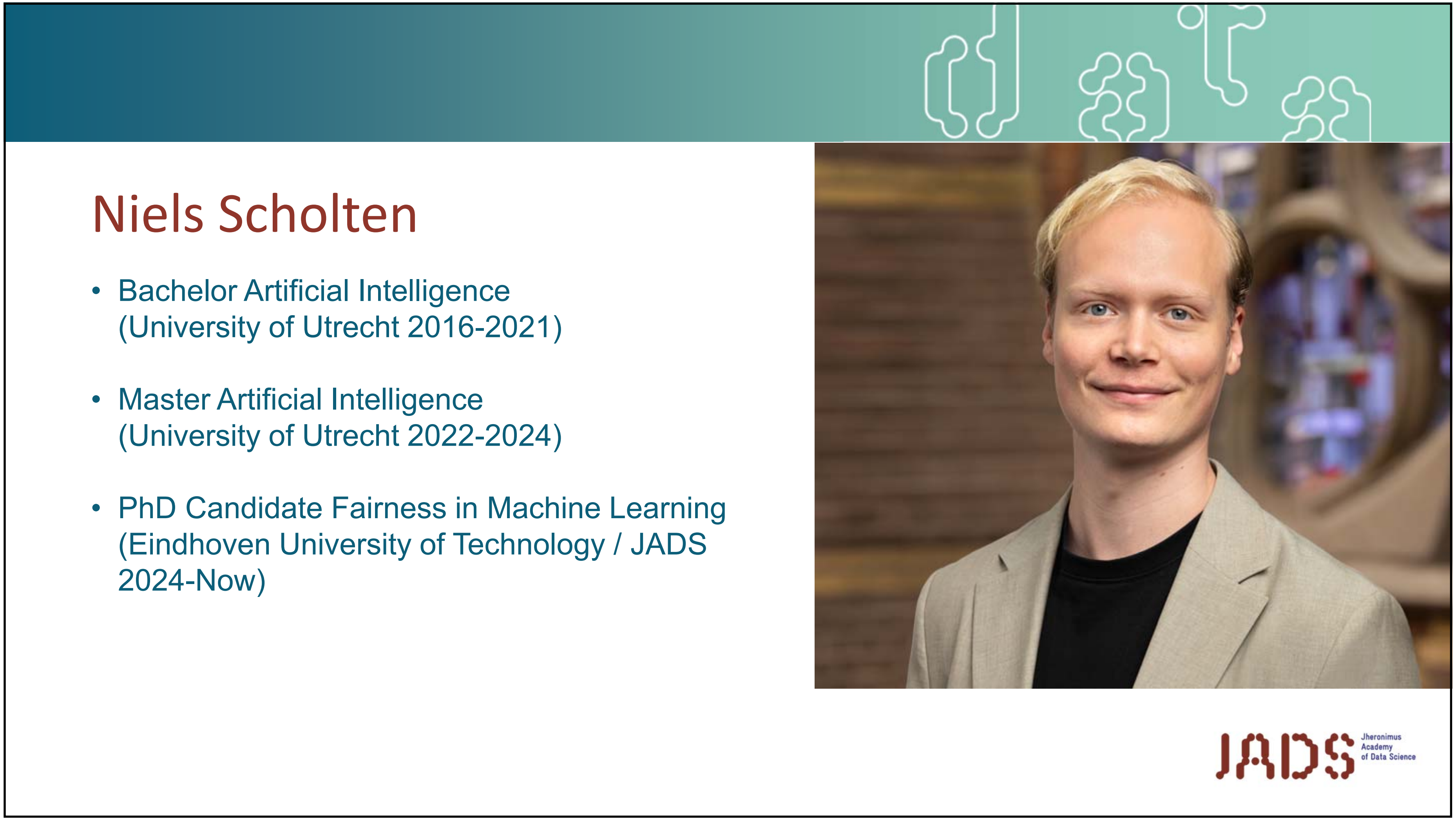
Instruction: Software Install

Niels Scholten | JM2050-M-6 Natural Language Processing

JADS is powered by



Niels Scholten

- Bachelor Artificial Intelligence (University of Utrecht 2016-2021)
- Master Artificial Intelligence (University of Utrecht 2022-2024)
- PhD Candidate Fairness in Machine Learning (Eindhoven University of Technology / JADS 2024-Now)

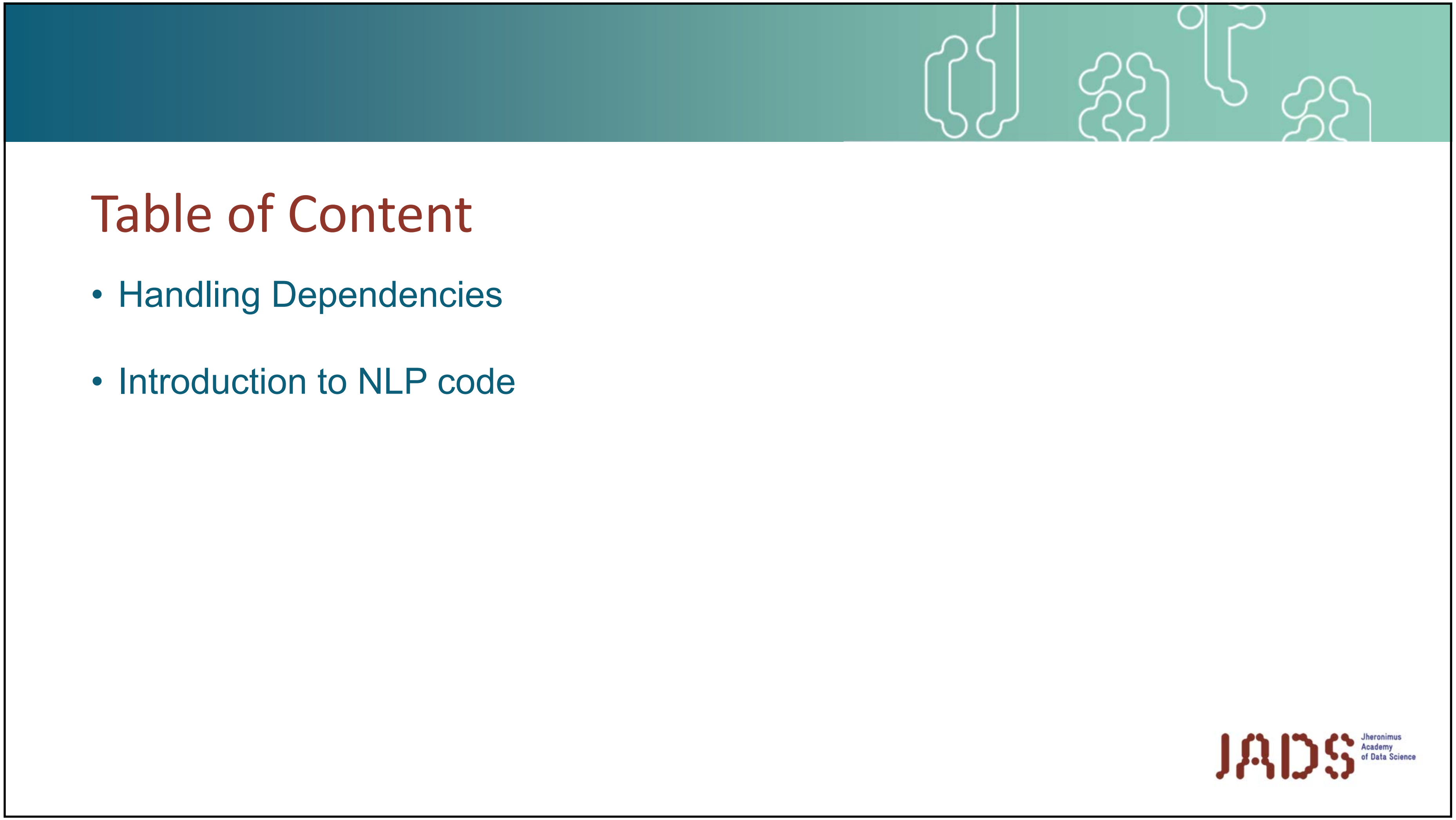
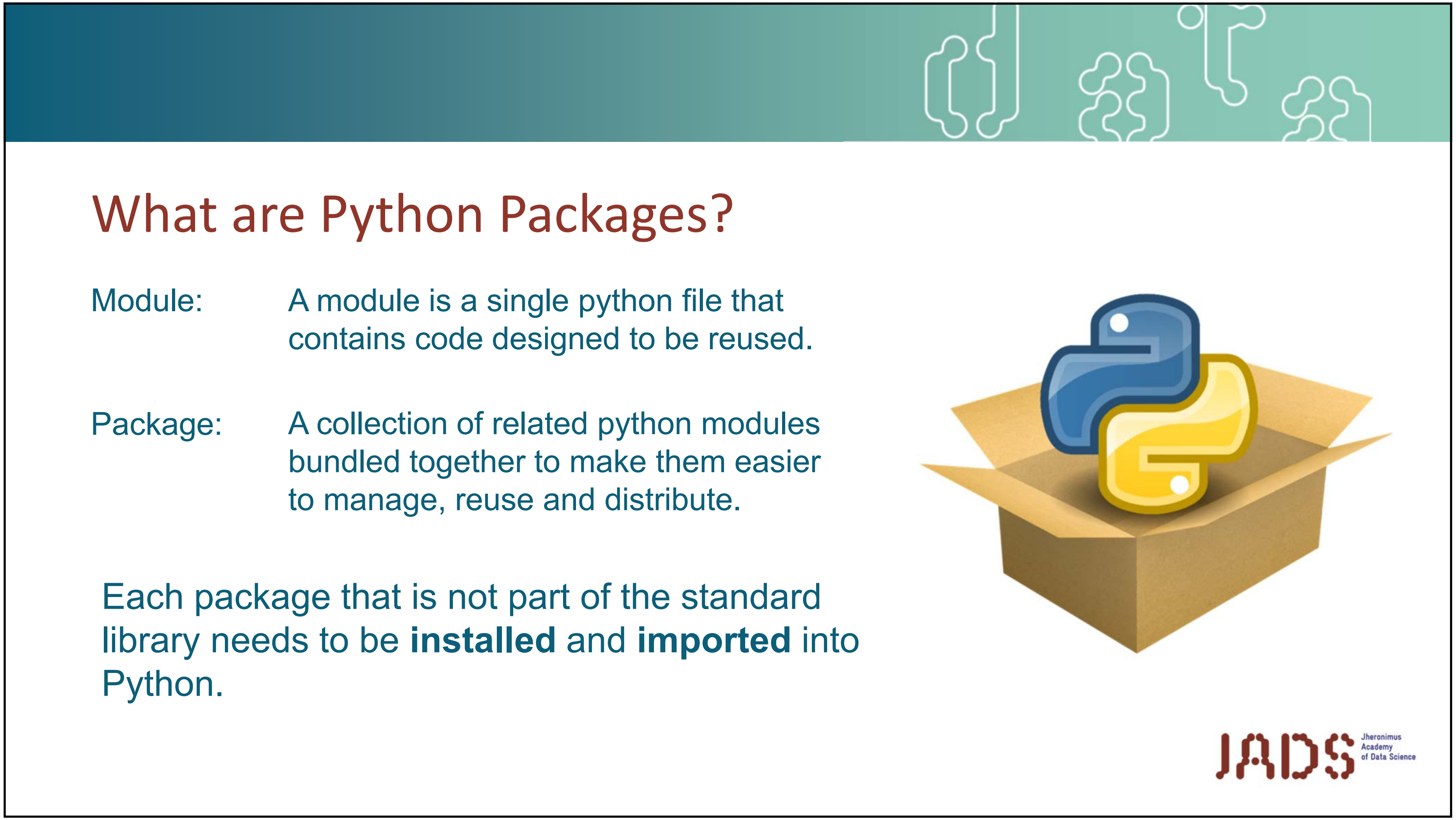


Table of Content

- Handling Dependencies
- Introduction to NLP code

Handling Dependencies

JADS Jheronimus Academy of Data Science



What are Python Packages?

Module: A module is a single python file that contains code designed to be reused.

Package: A collection of related python modules bundled together to make them easier to manage, reuse and distribute.

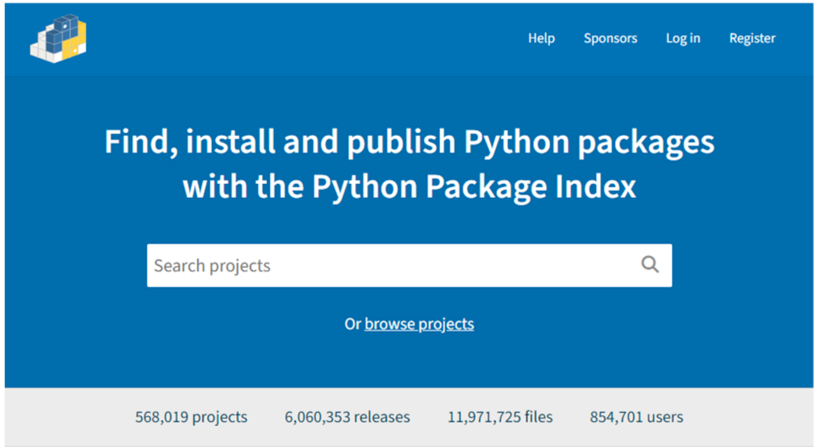
Each package that is not part of the standard library needs to be **installed** and **imported** into Python.

Python Package Index

- Python packages can be downloaded from PyPI (Python Package Index)

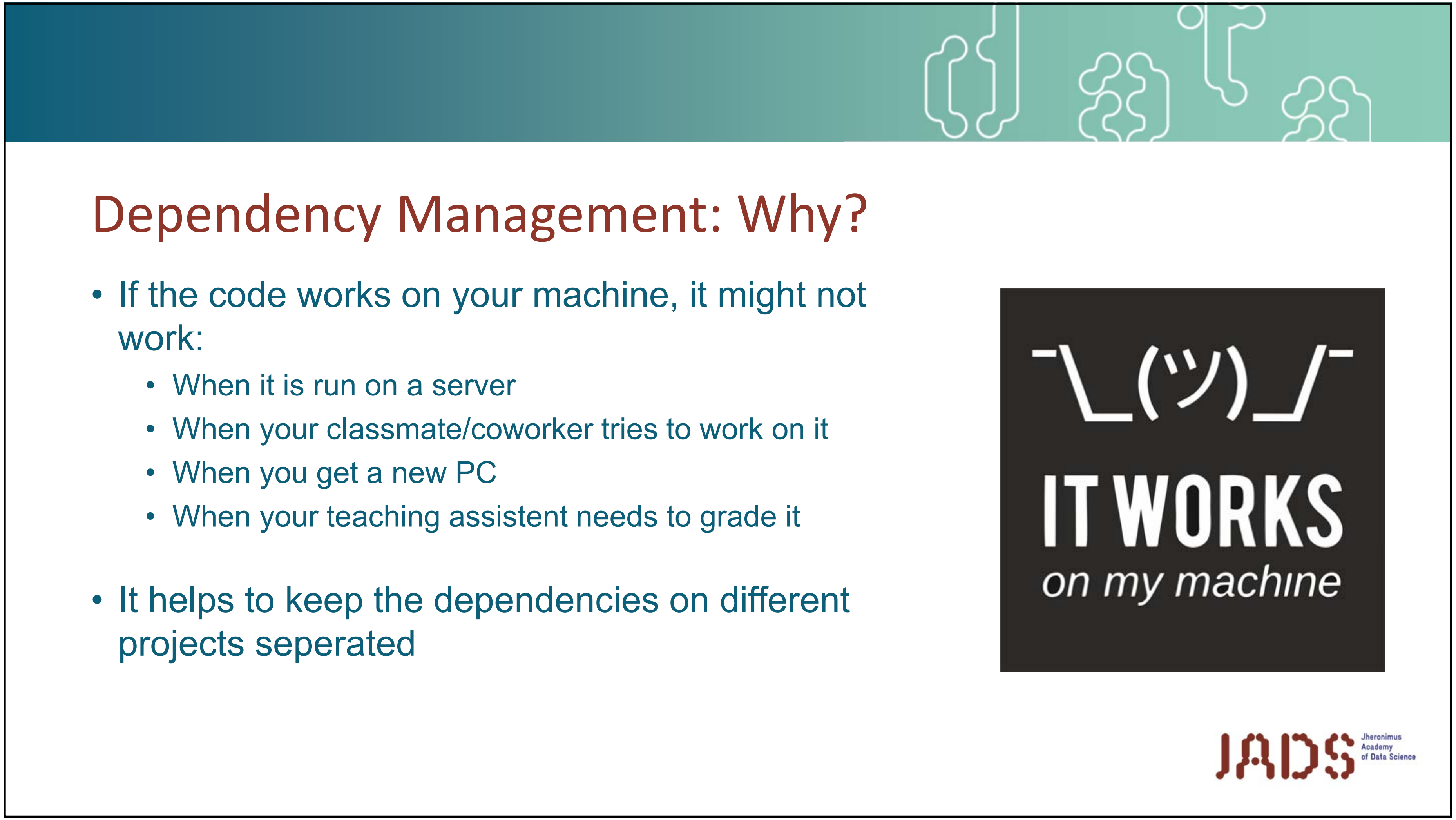
PyPI: A repository of software for the Python programming language

- You can find information and download packages from the PyPI website, but this **is not the best way** to manage your packages



Find, install and publish Python packages with the Python Package Index

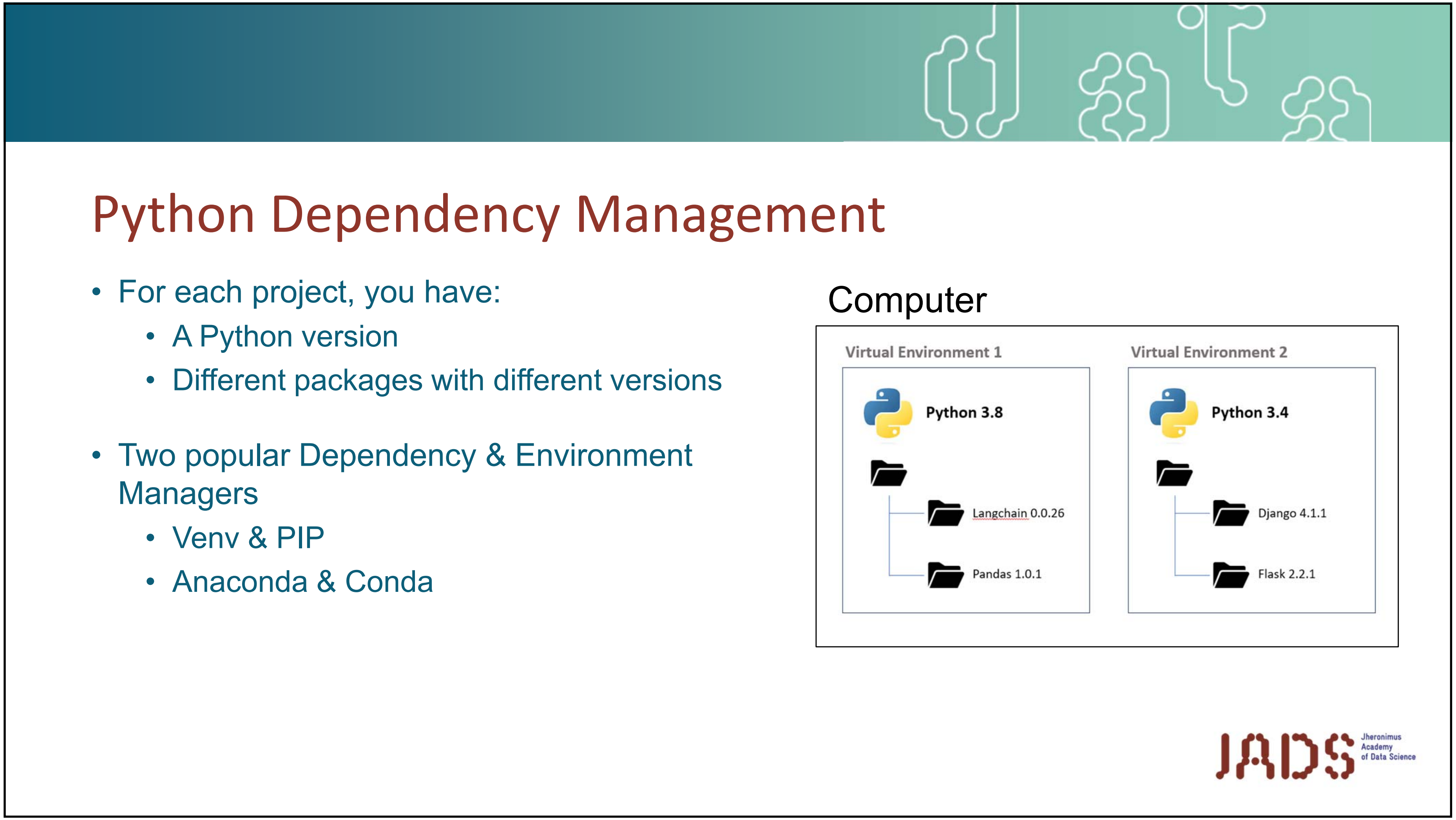
568,019 projects 6,060,353 releases 11,971,725 files 854,701 users



Dependency Management: Why?

- If the code works on your machine, it might not work:
  - When it is run on a server
  - When your classmate/coworker tries to work on it
  - When you get a new PC
  - When your teaching assistent needs to grade it
- It helps to keep the dependencies on different projects seperated

¯_(ツ)_/¯
IT WORKS
on my machine



Python Dependency Management

- For each project, you have:
  - A Python version
  - Different packages with different versions
- Two popular Dependency & Environment Managers
  - Venv & PIP
  - Anaconda & Conda

Computer

Virtual Environment 1

Python 3.8

Langchain 0.0.26

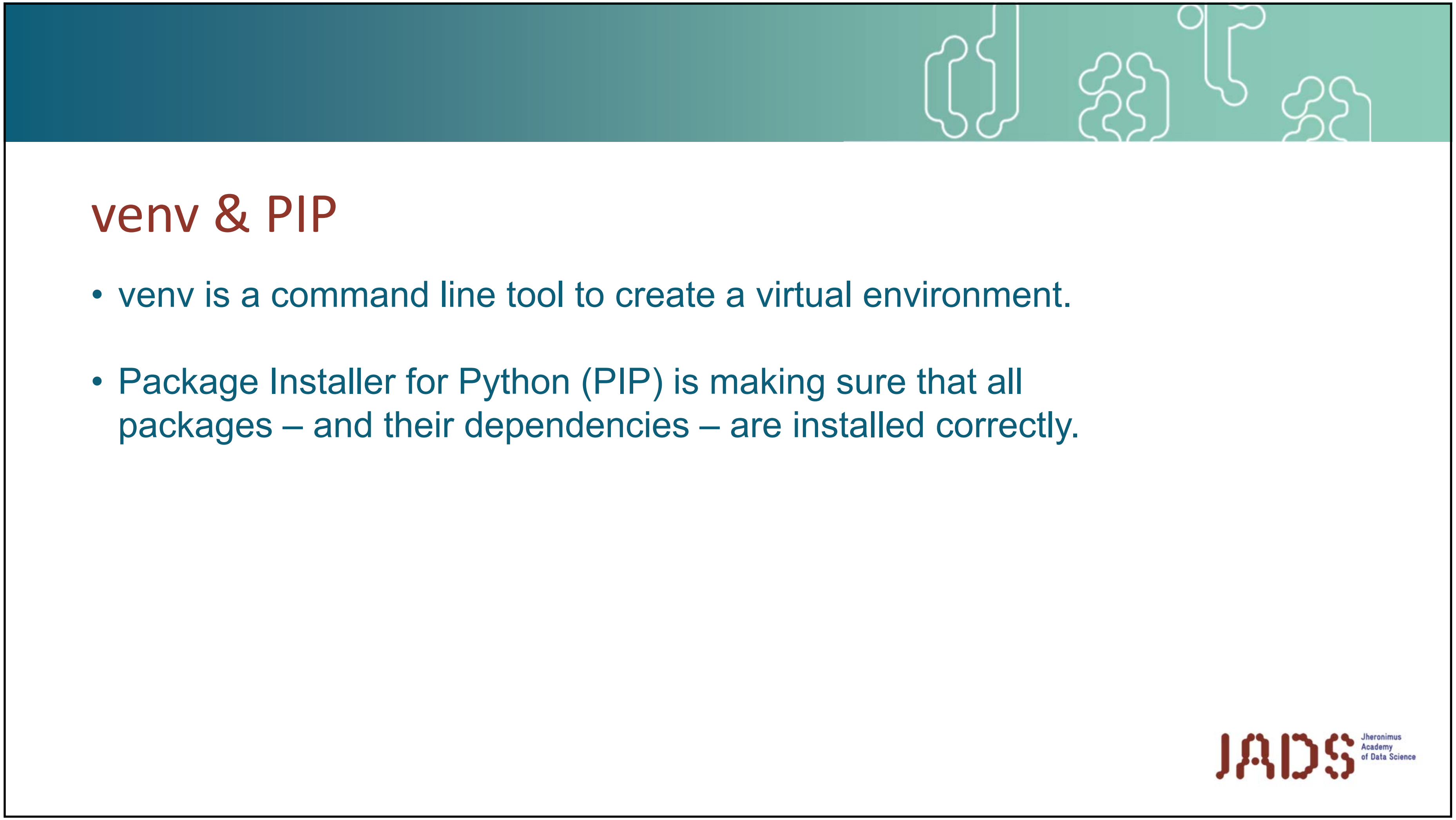
Pandas 1.0.1

Virtual Environment 2

Python 3.4

Django 4.1.1

Flask 2.2.1



venv & PIP

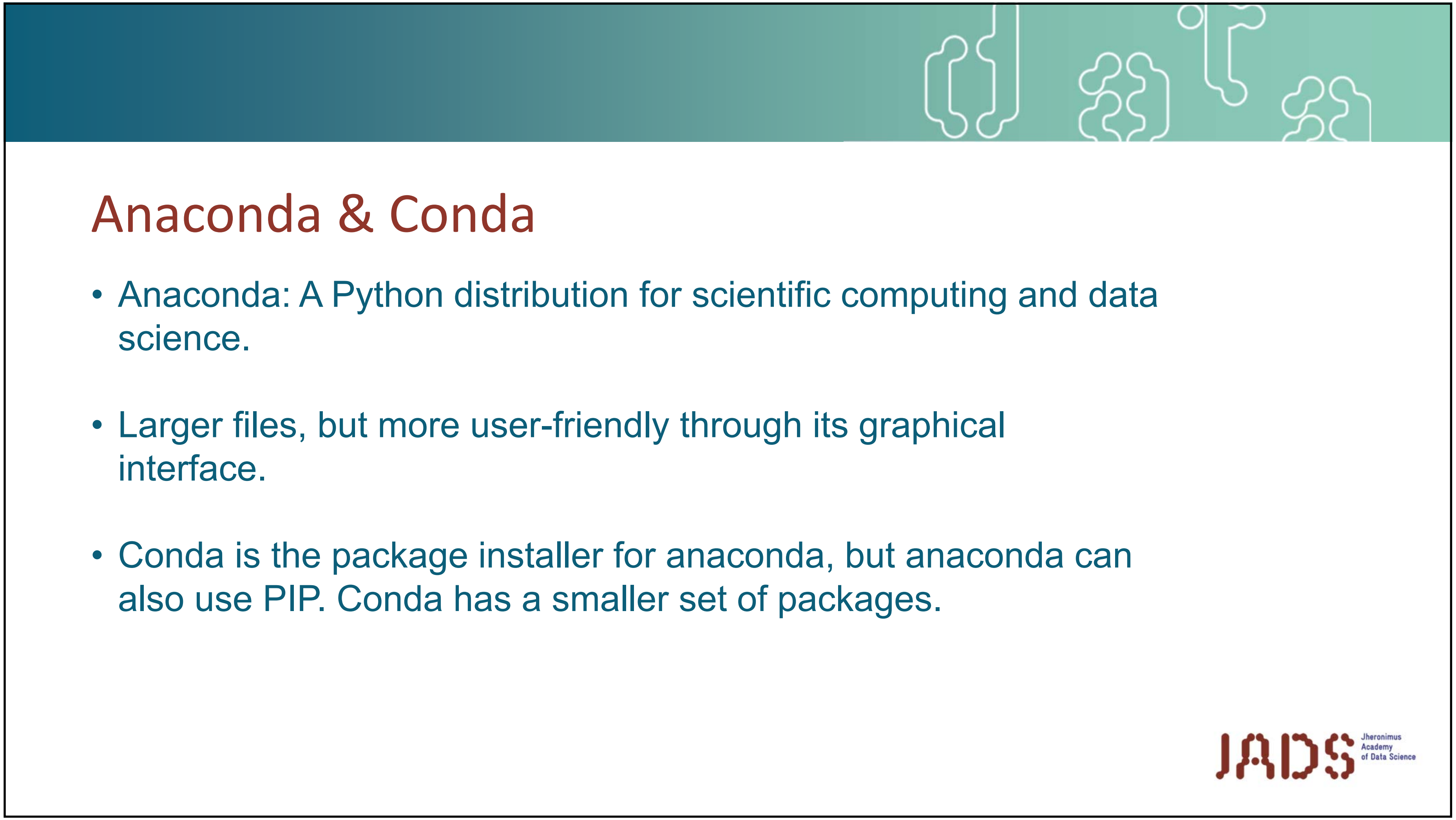
- venv is a command line tool to create a virtual environment.
- Package Installer for Python (PIP) is making sure that all packages – and their dependencies – are installed correctly.

JADS Jheronimus Academy of Data Science

Demo time

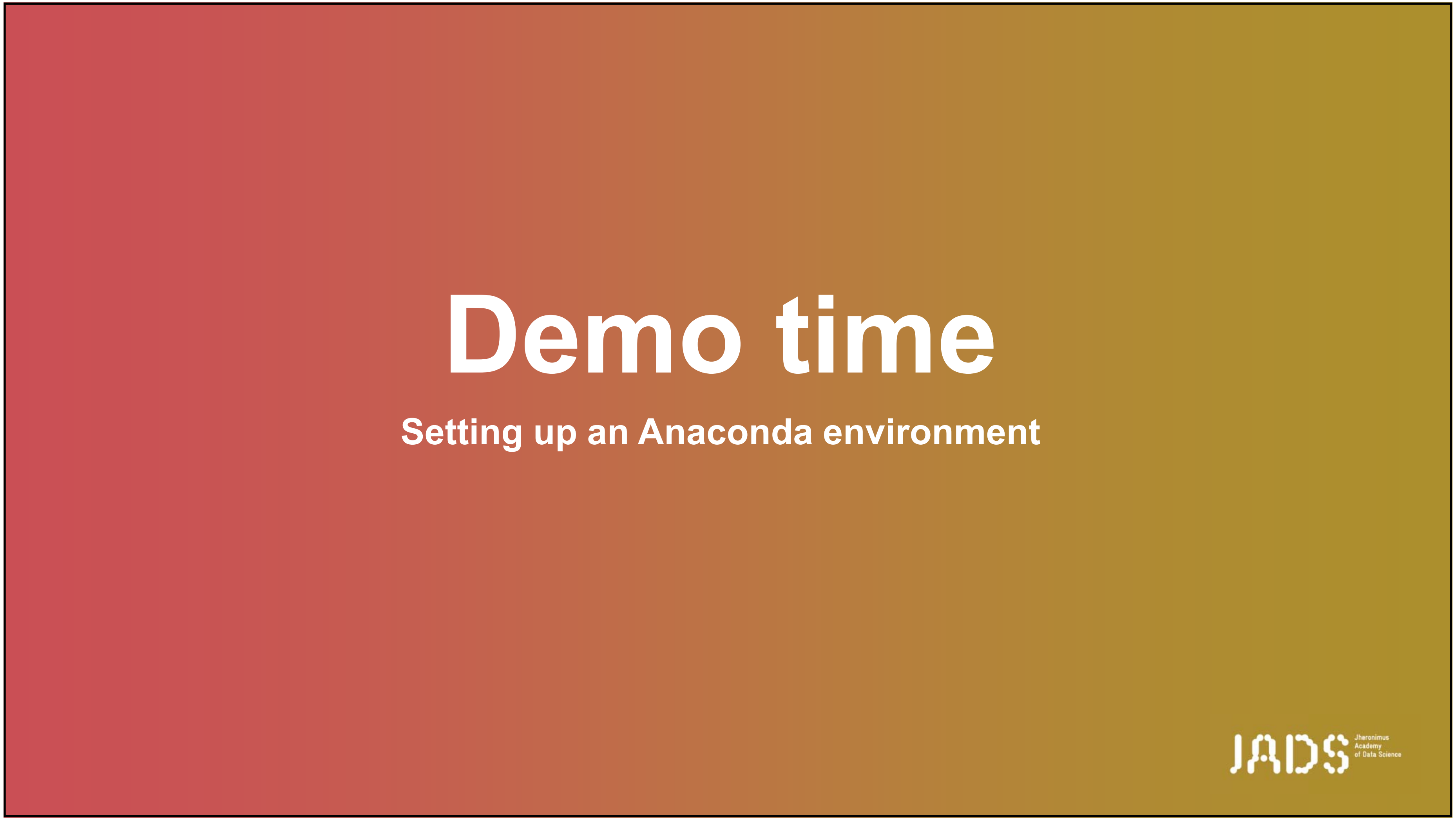
Virtual Environments with venv and PIP

JADS Jheronimus Academy of Data Science



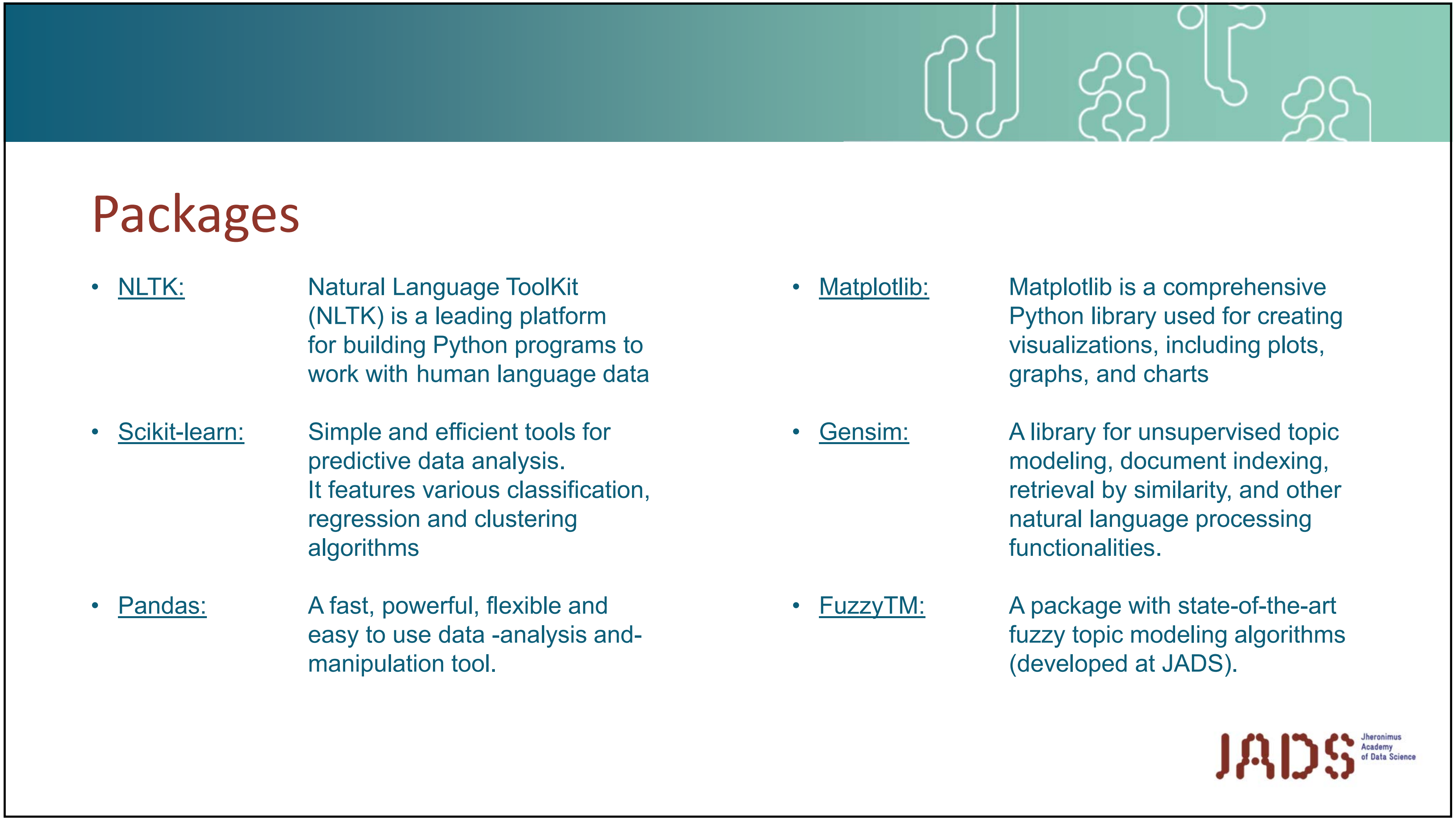
Anaconda & Conda

- Anaconda: A Python distribution for scientific computing and data science.
- Larger files, but more user-friendly through its graphical interface.
- Conda is the package installer for anaconda, but anaconda can also use PIP. Conda has a smaller set of packages.



Demo time

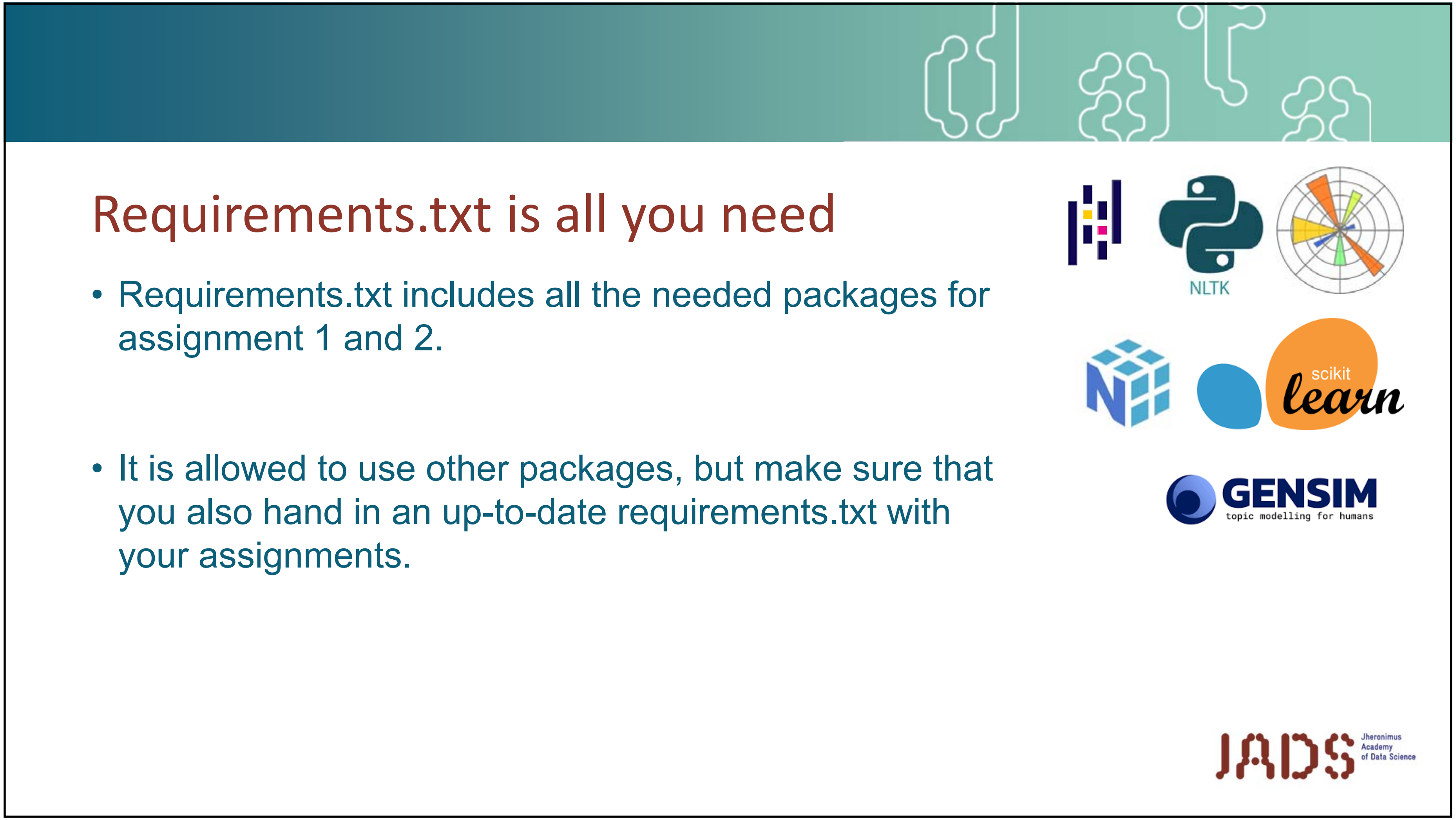
Setting up an Anaconda environment



Packages

- NLTK: Natural Language ToolKit (NLTK) is a leading platform for building Python programs to work with human language data
- Scikit-learn: Simple and efficient tools for predictive data analysis.
  It features various classification, regression and clustering algorithms
- Pandas: A fast, powerful, flexible and easy to use data -analysis and-manipulation tool.
- Matplotlib: Matplotlib is a comprehensive Python library used for creating visualizations, including plots, graphs, and charts
- Gensim: A library for unsupervised topic modeling, document indexing, retrieval by similarity, and other natural language processing functionalities.
- FuzzyTM: A package with state-of-the-art fuzzy topic modeling algorithms (developed at JADS).

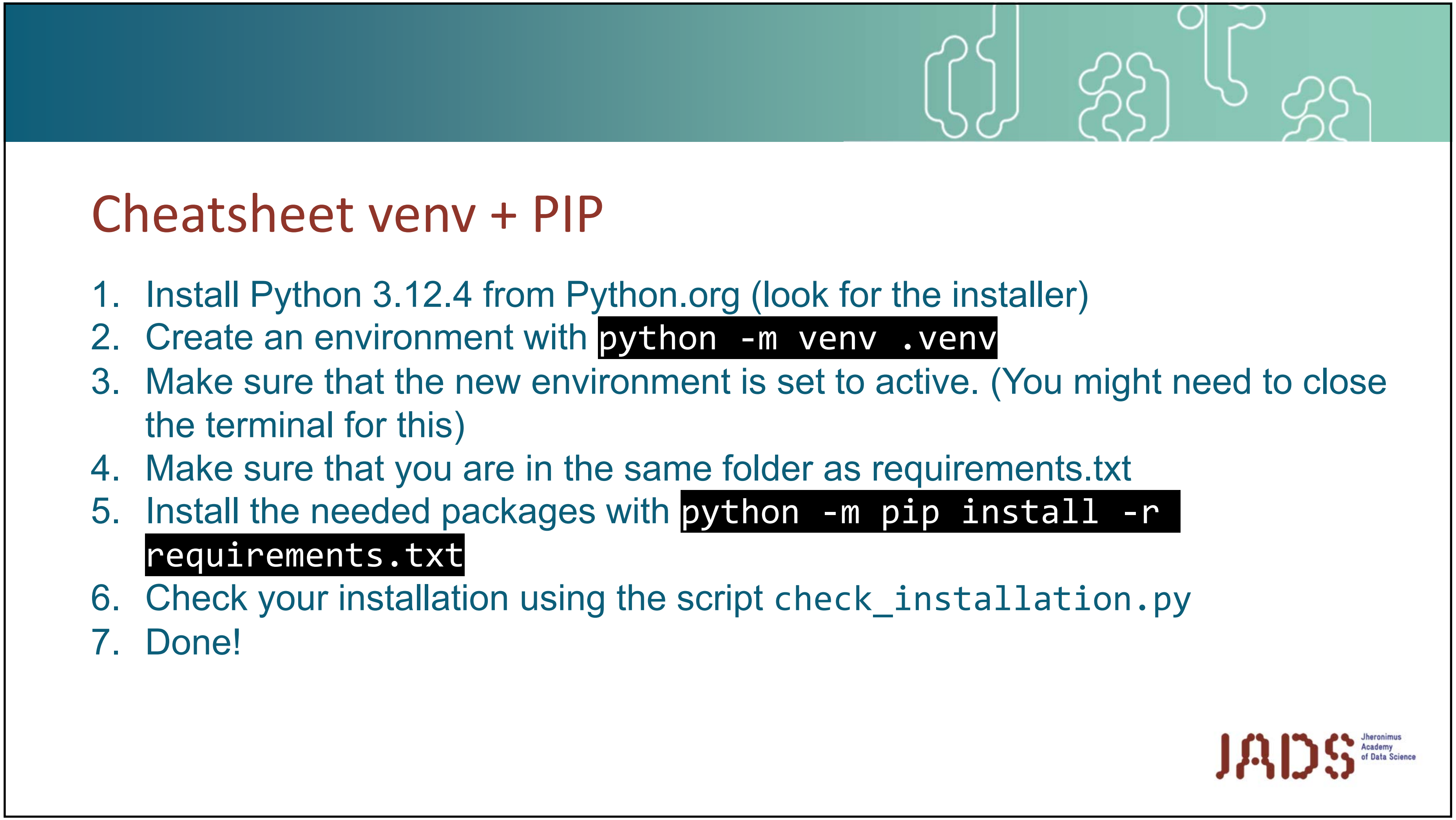
JADS Jheronimus Academy of Data Science



Requirements.txt is all you need

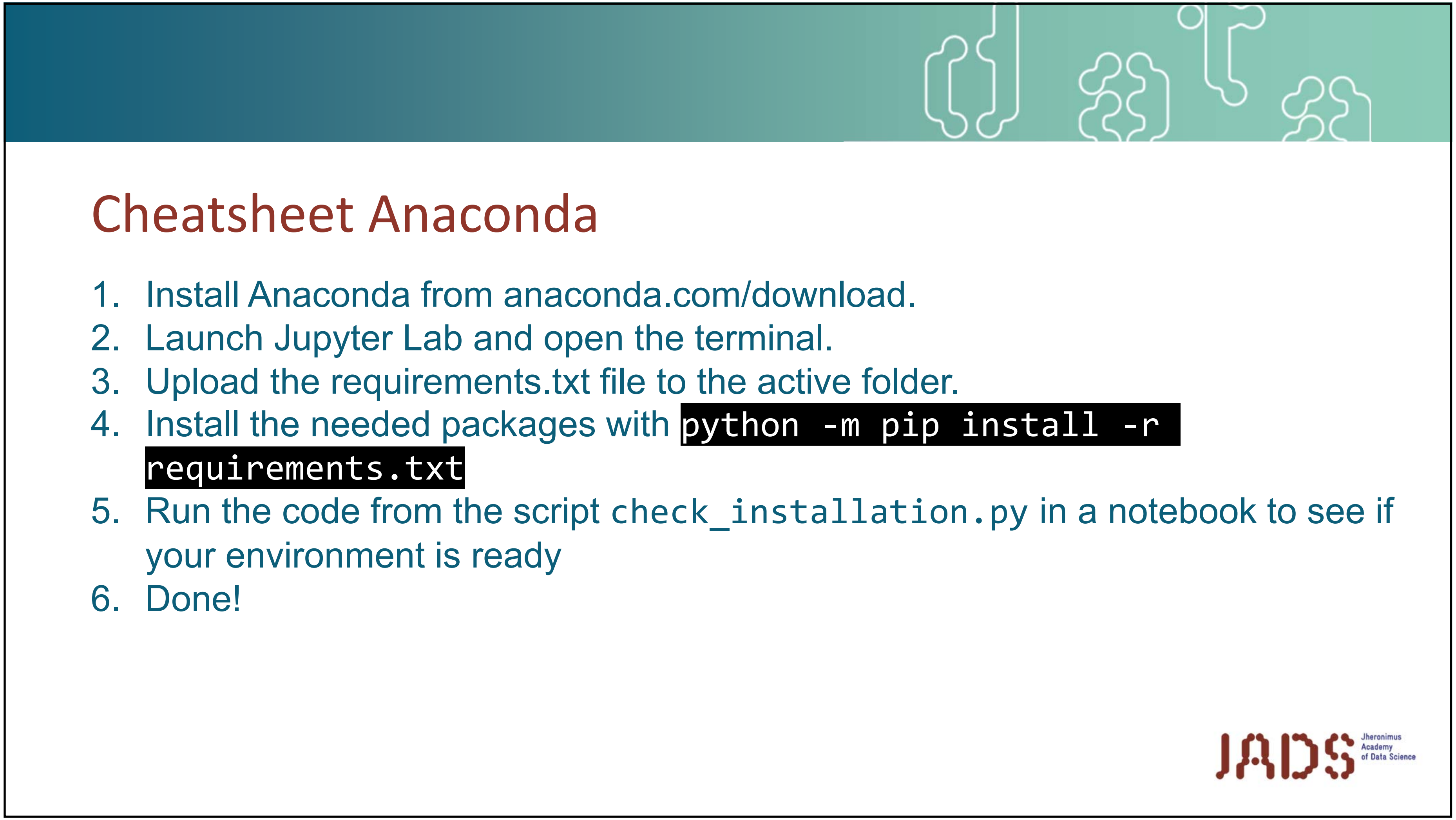
- Requirements.txt includes all the needed packages for assignment 1 and 2.
- It is allowed to use other packages, but make sure that you also hand in an up-to-date requirements.txt with your assignments.

NLTK
scikit
learn
GENSIM
topic modelling for humans
JADS
Jheronimus Academy of Data Science



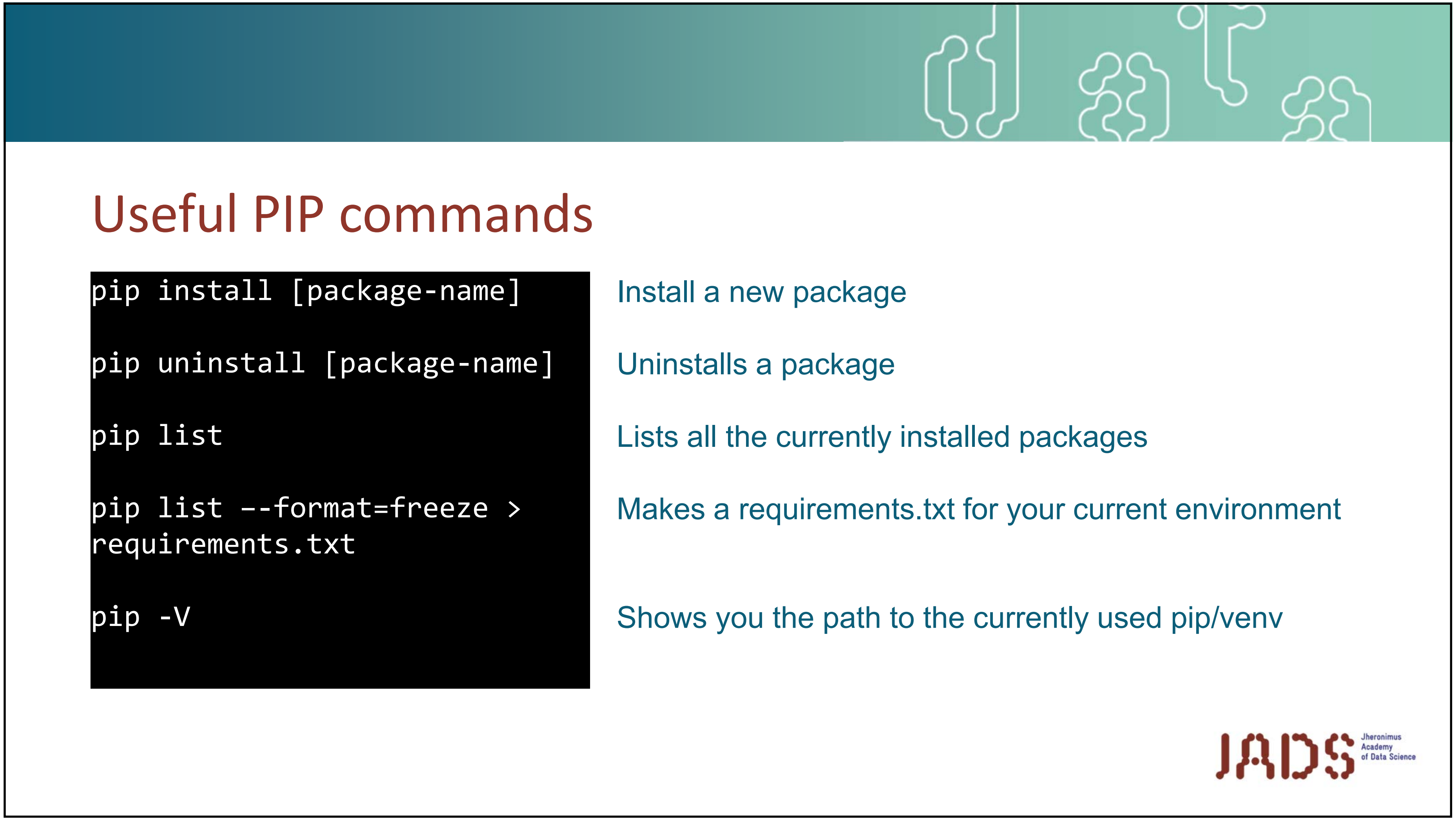
Cheatsheet venv + PIP

1. Install Python 3.12.4 from Python.org (look for the installer)
2. Create an environment with `python -m venv .venv`
3. Make sure that the new environment is set to active. (You might need to close the terminal for this)
4. Make sure that you are in the same folder as requirements.txt
5. Install the needed packages with `python -m pip install -r requirements.txt`
6. Check your installation using the script `check_installation.py`
7. Done!



Cheatsheet Anaconda

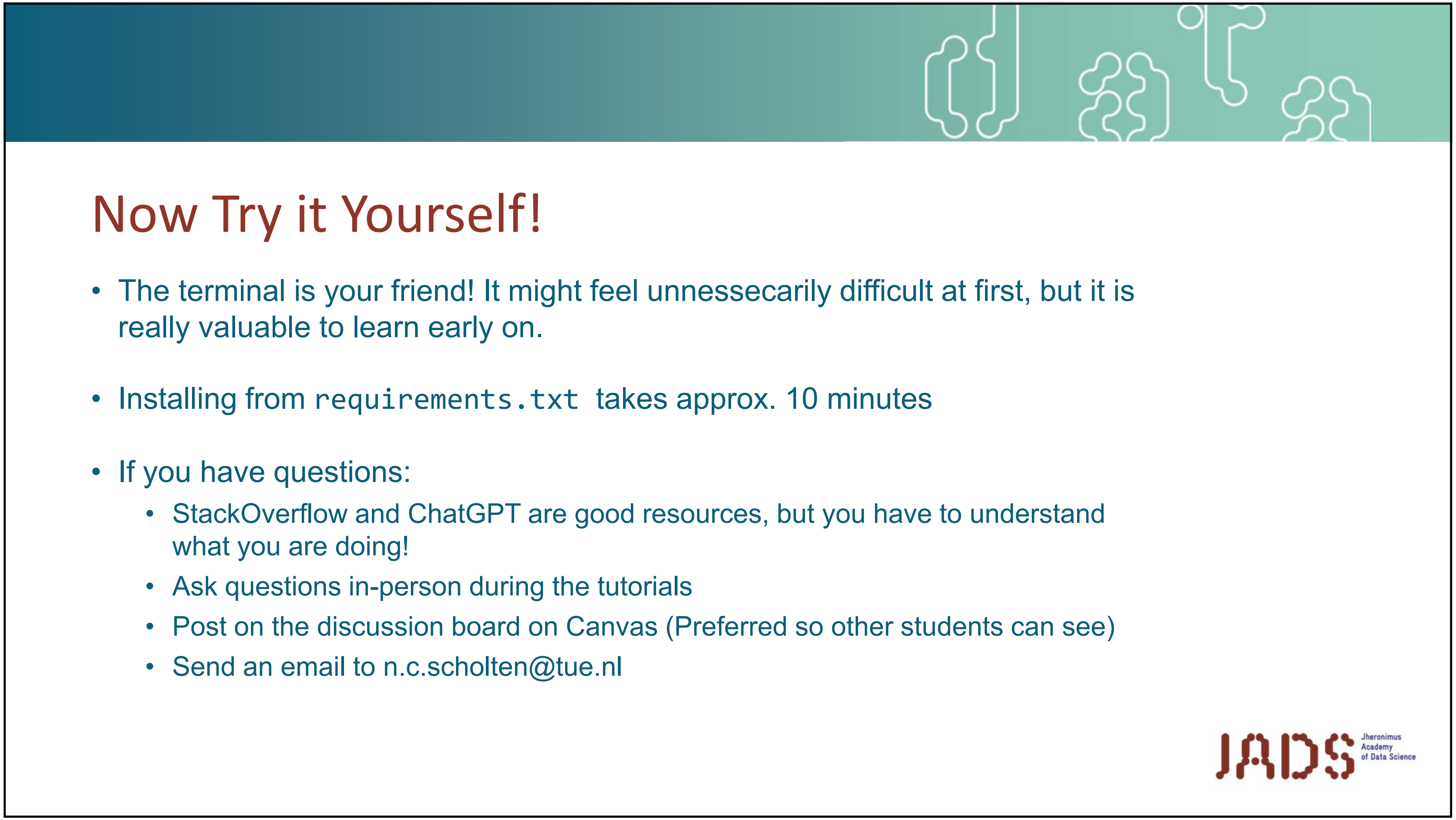
1. Install Anaconda from anaconda.com/download.
2. Launch Jupyter Lab and open the terminal.
3. Upload the requirements.txt file to the active folder.
4. Install the needed packages with `python -m pip install -r requirements.txt`
5. Run the code from the script `check_installation.py` in a notebook to see if your environment is ready
6. Done!



Useful PIP commands

`pip install [package-name]`	Install a new package
`pip uninstall [package-name]`	Uninstalls a package
`pip list`	Lists all the currently installed packages
`pip list --format=freeze > requirements.txt`	Makes a requirements.txt for your current environment
`pip -V`	Shows you the path to the currently used pip/venv

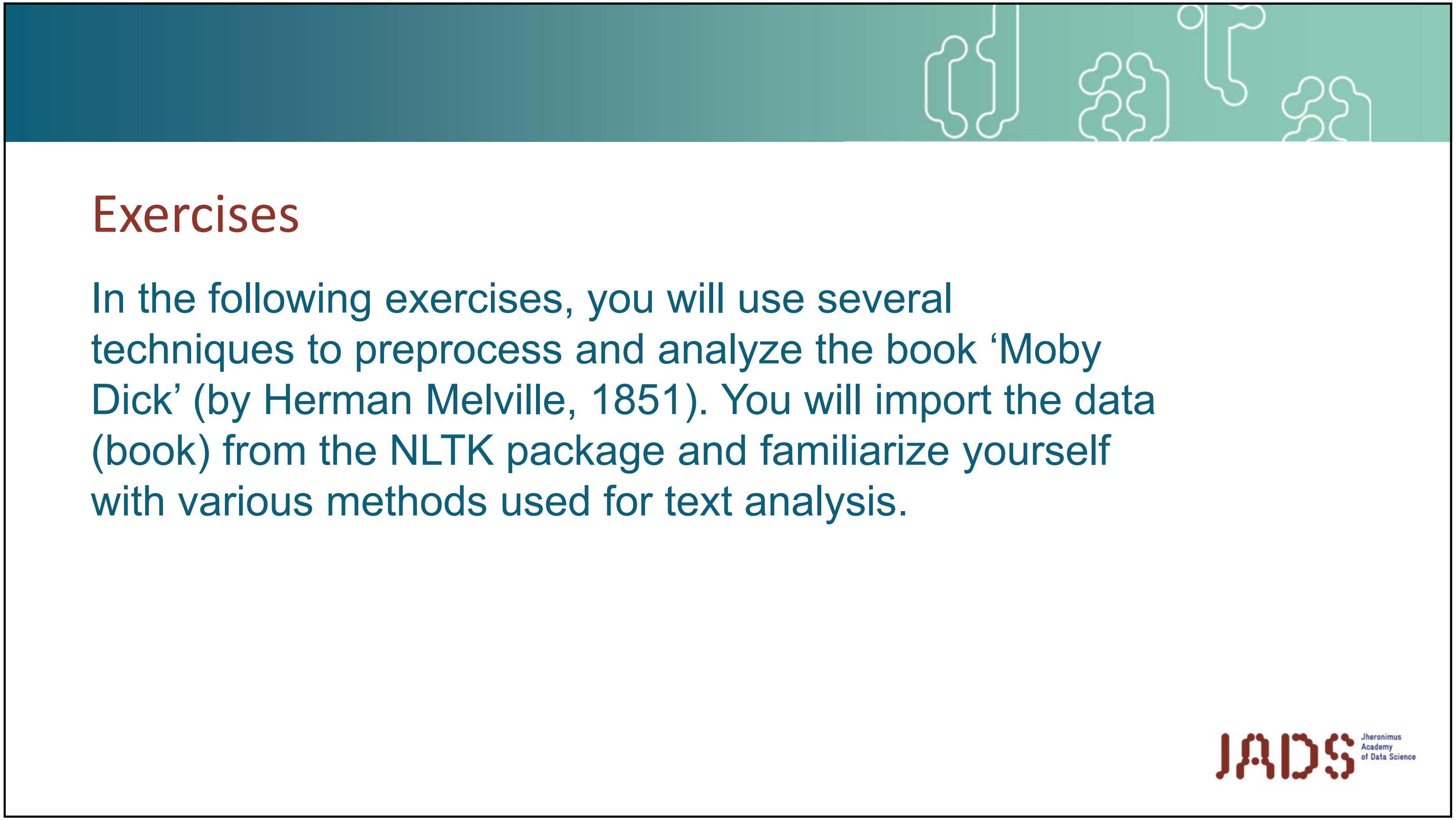
JADS
Jheronimus Academy of Data Science



Now Try it Yourself!

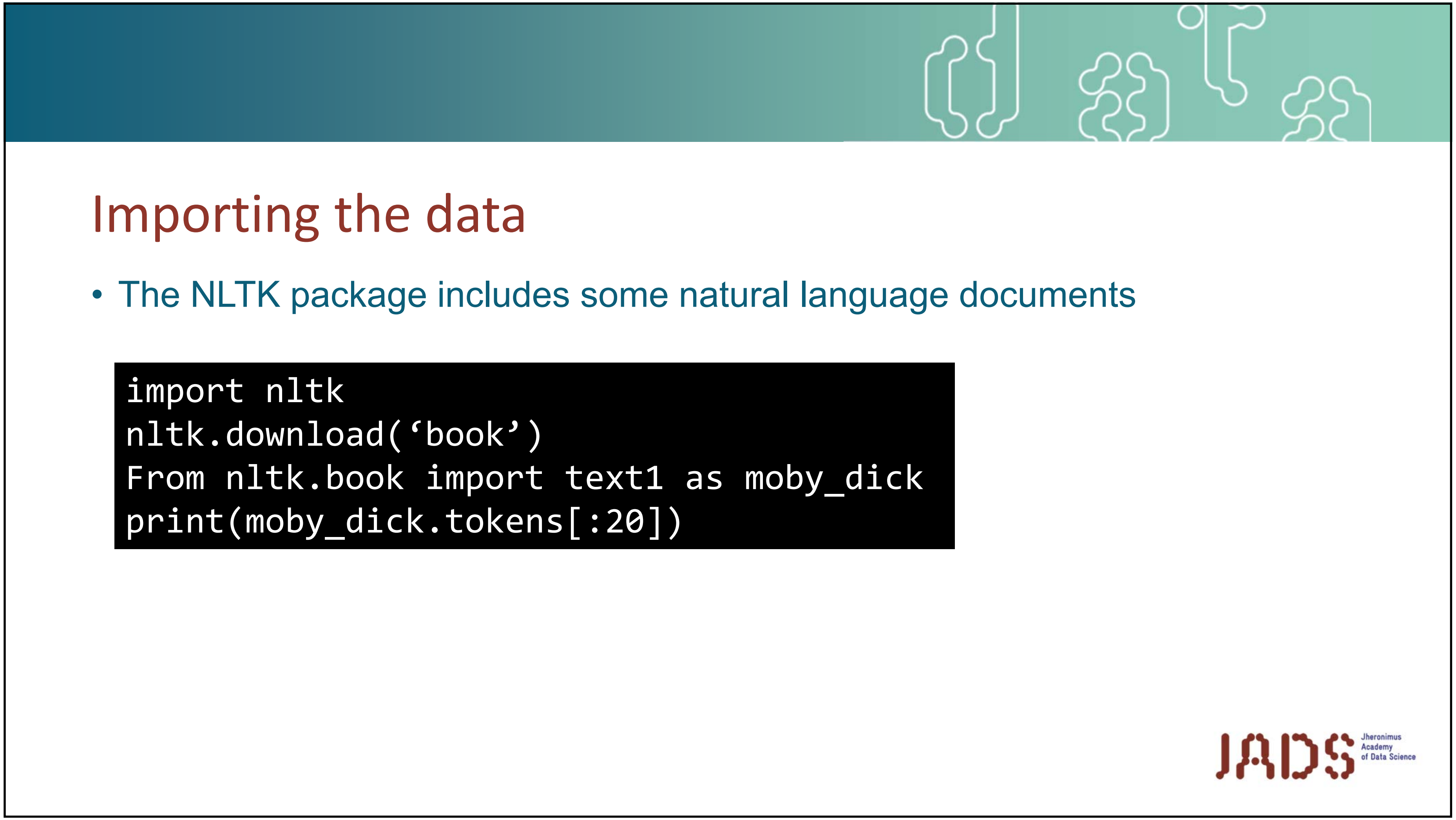
- The terminal is your friend! It might feel unnessecarily difficult at first, but it is really valuable to learn early on.
- Installing from `requirements.txt` takes approx. 10 minutes
- If you have questions:
  - StackOverflow and ChatGPT are good resources, but you have to understand what you are doing!
  - Ask questions in-person during the tutorials
  - Post on the discussion board on Canvas (Preferred so other students can see)
  - Send an email to n.c.scholten@tue.nl

Introduction NLP Code



Exercises

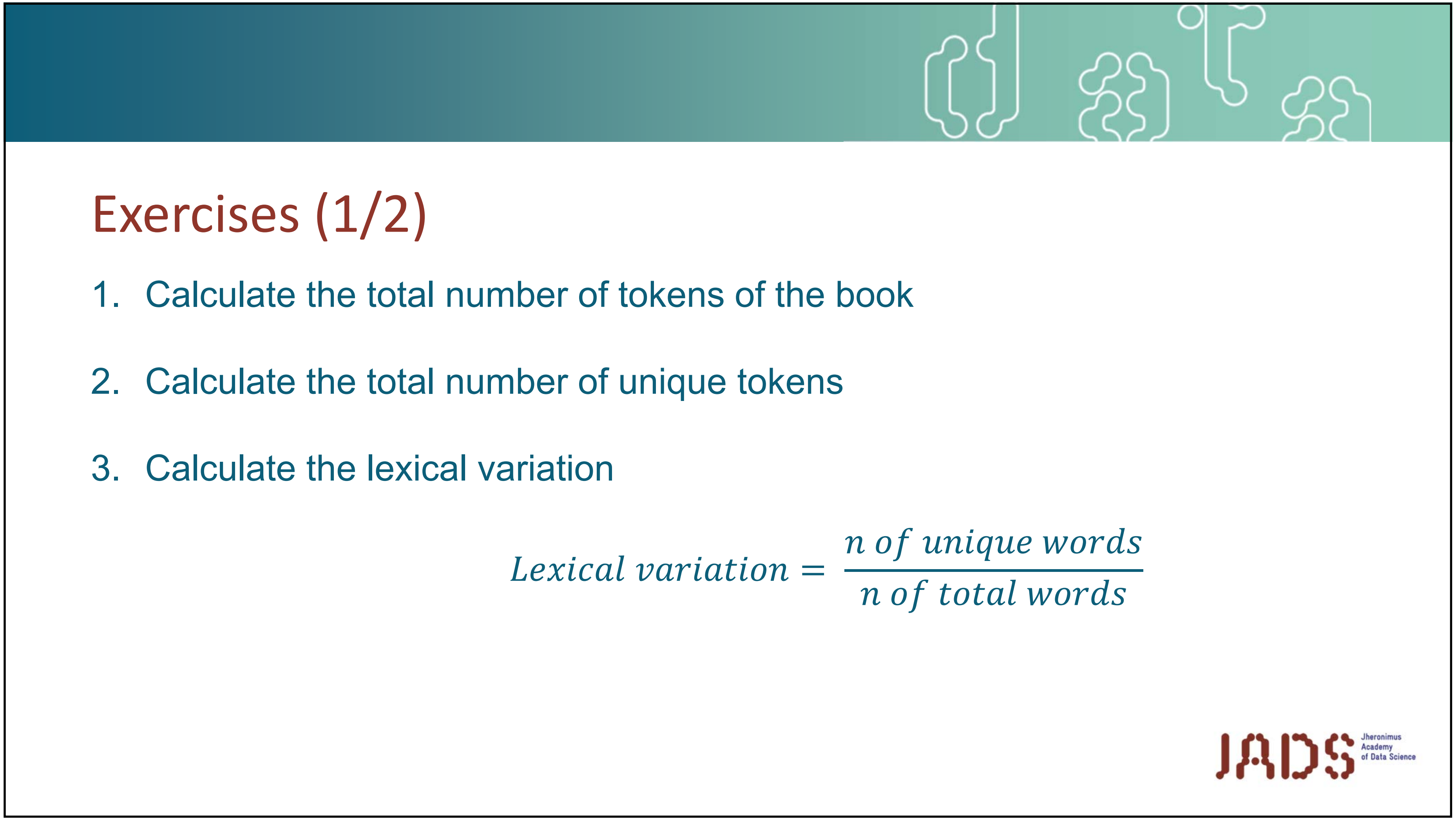
In the following exercises, you will use several techniques to preprocess and analyze the book ‘Moby Dick’ (by Herman Melville, 1851). You will import the data (book) from the NLTK package and familiarize yourself with various methods used for text analysis.



Importing the data

- The NLTK package includes some natural language documents

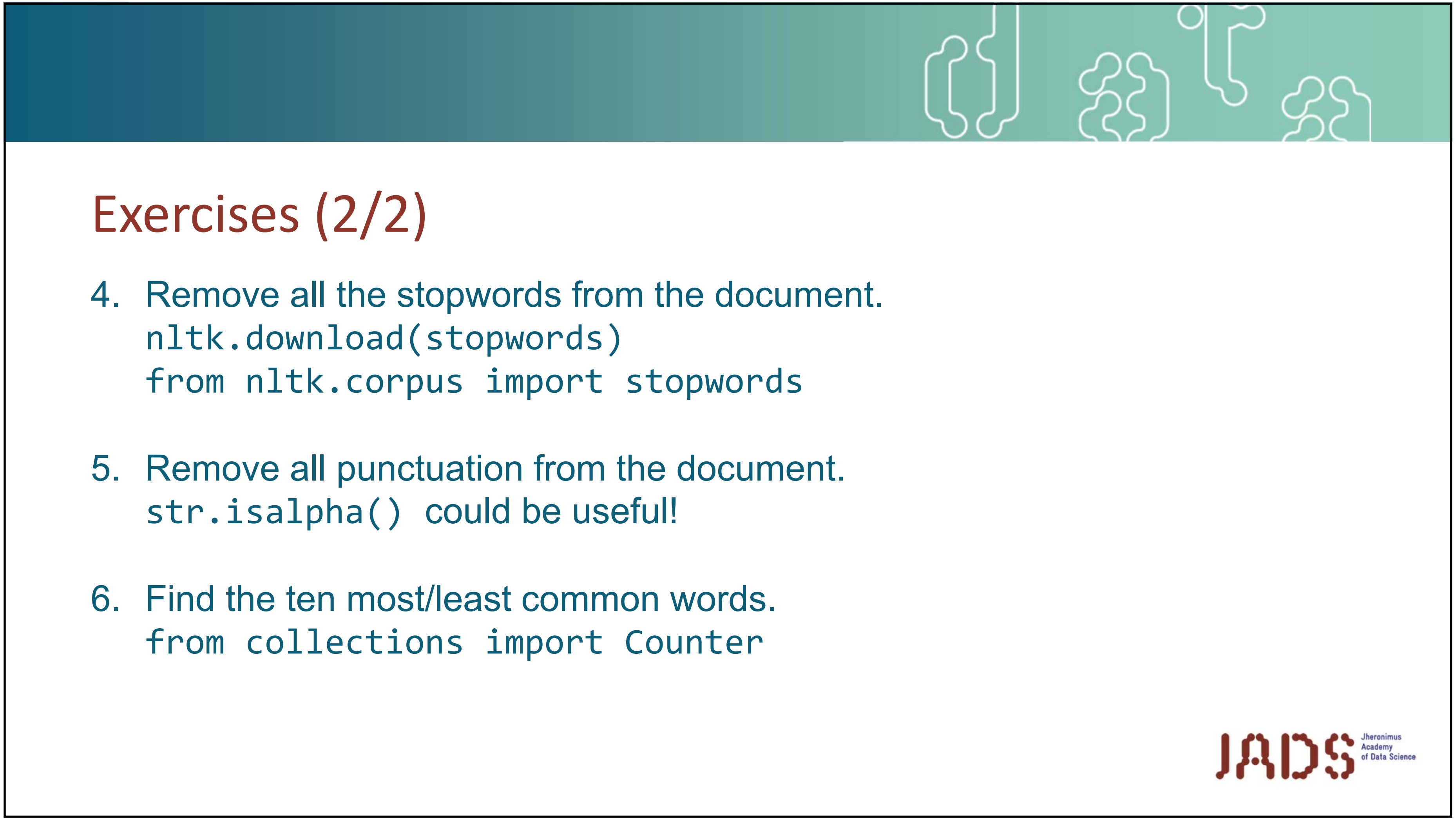
```
import nltk
nltk.download(‘book’)
From nltk.book import text1 as moby_dick
print(moby_dick.tokens[:20])
```



Exercises (1/2)

1. Calculate the total number of tokens of the book
2. Calculate the total number of unique tokens
3. Calculate the lexical variation

$$Lexical\ variation = \frac{n\ of\ unique\ words}{n\ of\ total\ words}$$



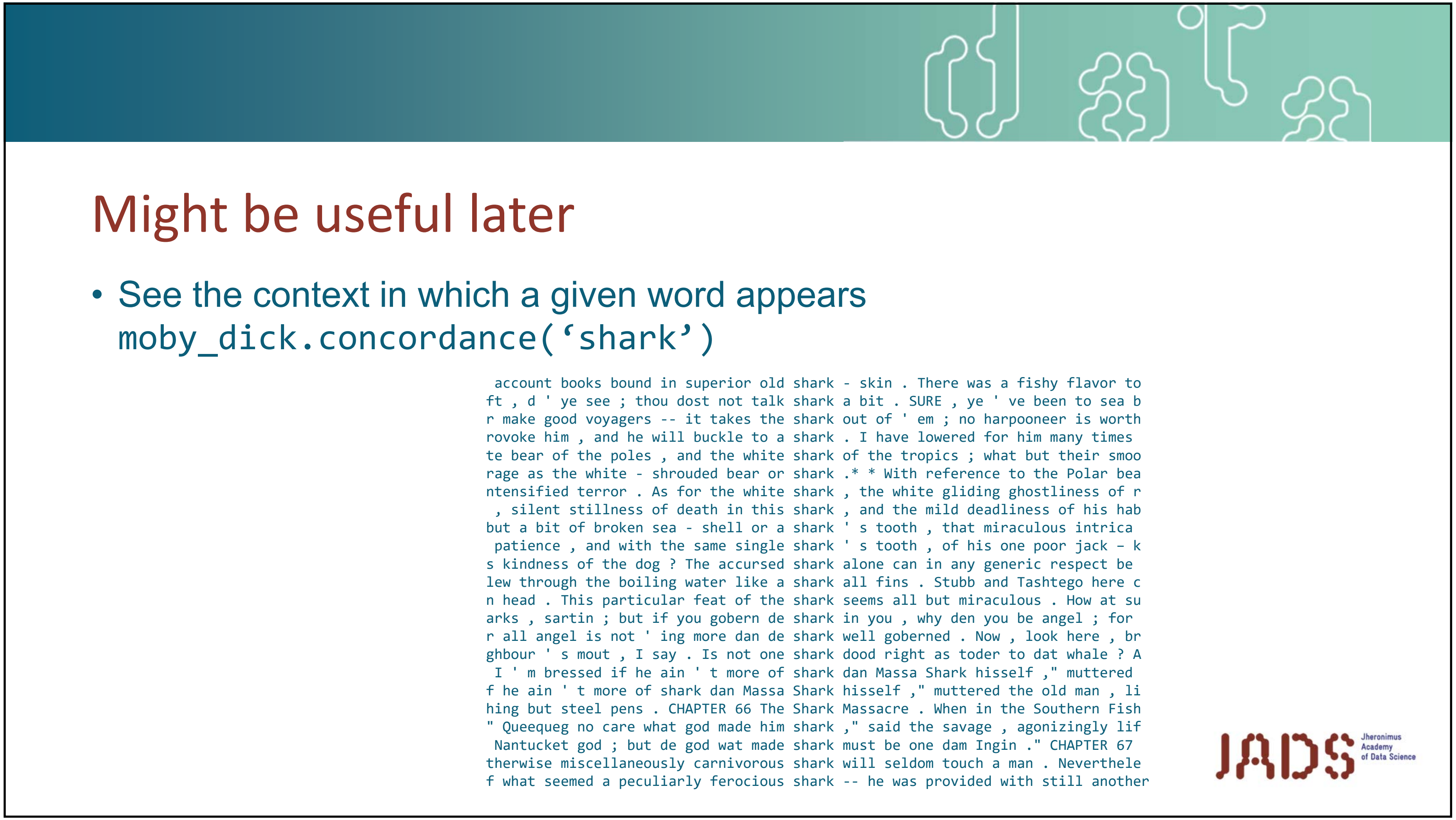
Exercises (2/2)

4. Remove all the stopwords from the document.
   ```
   nltk.download(stopwords)
   from nltk.corpus import stopwords
   ```

5. Remove all punctuation from the document.
   `str.isalpha()` could be useful!

6. Find the ten most/least common words.
   ```
   from collections import Counter
   ```

JADS Jheronimus Academy of Data Science



Might be useful later

- See the context in which a given word appears

```
moby_dick.concordance('shark')
```

```
 account books bound in superior old shark - skin . There was a fishy flavor to
ft , d ' ye see ; thou dost not talk shark a bit . SURE , ye ' ve been to sea b
r make good voyagers -- it takes the shark out of ' em ; no harpooneer is worth
rovoke him , and he will buckle to a shark . I have lowered for him many times
te bear of the poles , and the white shark of the tropics ; what but their smoo
rage as the white - shrouded bear or shark .* * With reference to the Polar bea
ntensified terror . As for the white shark , the white gliding ghostliness of r
 , silent stillness of death in this shark , and the mild deadliness of his hab
but a bit of broken sea - shell or a shark ' s tooth , that miraculous intrica
 patience , and with the same single shark ' s tooth , of his one poor jack – k
s kindness of the dog ? The accursed shark alone can in any generic respect be
lew through the boiling water like a shark all fins . Stubb and Tashtego here c
n head . This particular feat of the shark seems all but miraculous . How at su
arks , sartin ; but if you gobern de shark in you , why den you be angel ; for
r all angel is not ' ing more dan de shark well goberned . Now , look here , br
ghbour ' s mout , I say . Is not one shark dood right as toder to dat whale ? A
 I ' m bressed if he ain ' t more of shark dan Massa Shark hisself ," muttered
f he ain ' t more of shark dan Massa Shark hisself ," muttered the old man , li
hing but steel pens . CHAPTER 66 The Shark Massacre . When in the Southern Fish
" Queequeg no care what god made him shark ," said the savage , agonizingly lif
 Nantucket god ; but de god wat made shark must be one dam Ingin ." CHAPTER 67
therwise miscellaneously carnivorous shark will seldom touch a man . Neverthele
f what seemed a peculiarly ferocious shark -- he was provided with still another
```

Contact

Niels Scholten

n.c.scholten@tue.nl