# STORY HUNT

*Workshop #3: Scraping & Cleaning*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# Today´s Program:

Introduction to Scraping

Web Scraping

*Demo: Chrome Scraper Plugin*

Introduction to Data Cleaning

Basic Data Cleaning Techniques:

*Demo: Open Refine*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# The Data Pipeline: Theory

## Finding & Getting

* Where to find Data?
* Data portals
* Freedom of Information Requests
* Webs-scraping

## Cleaning

* transform documents into a structured data

## Analyzing

* How do I spot:
* Patterns & Trends ?
* Outlier ?
* Connections between Columns, Datasets
* relation between Data & Context

## Visualizing

* How to visualize the essence of a finding:
* appealing
* informative

## Communicating

* What´s the right:
* Format?
* Medium?
* Target group?

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# The Data Pipeline: Tools

## Finding & Getting

* Google Sheets
* Webscraper
* Unbubble
* Tabula

## Cleaning

* Open Refine
* Trifacta
* Google Sheets / Excel
* Python, R

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# What is Scraping?

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# What is Scraping:

unstructured Documents

Structured tabular data

3 Steps:

Fetch the document (most likely from the Internet) → Follow futher linked documents (especially in web scraping) → Extract interesting parts and store them in a structured format

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

## When is Web-Scraping useful?

- *avoid tedious big copy-pasting jobs*

## What is the challenge?

- *one scraper only works for documents from the same source (e.g. Website, PDF Documents frame the same collection)*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# Chrome Scraping Plugin

*A web scraper with user interface*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# Chrome Scraping Plugin

*Download-Link: http://bit.ly/RmLaUv*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# Basics: Data Cleaning

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# Basics: Clean Data

*Structured, machine-readable Data*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# What makes Data Clean?

Consistent spelling of corresponding values (U.S. ≠ USA)

WatchOut: No trailing Spaces in front/ behind a value

Unique column labels

No hidden rows / columns

No empty rows / columns

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

# Open Refine

*Free Open Source Tool for working with messy Data*

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.

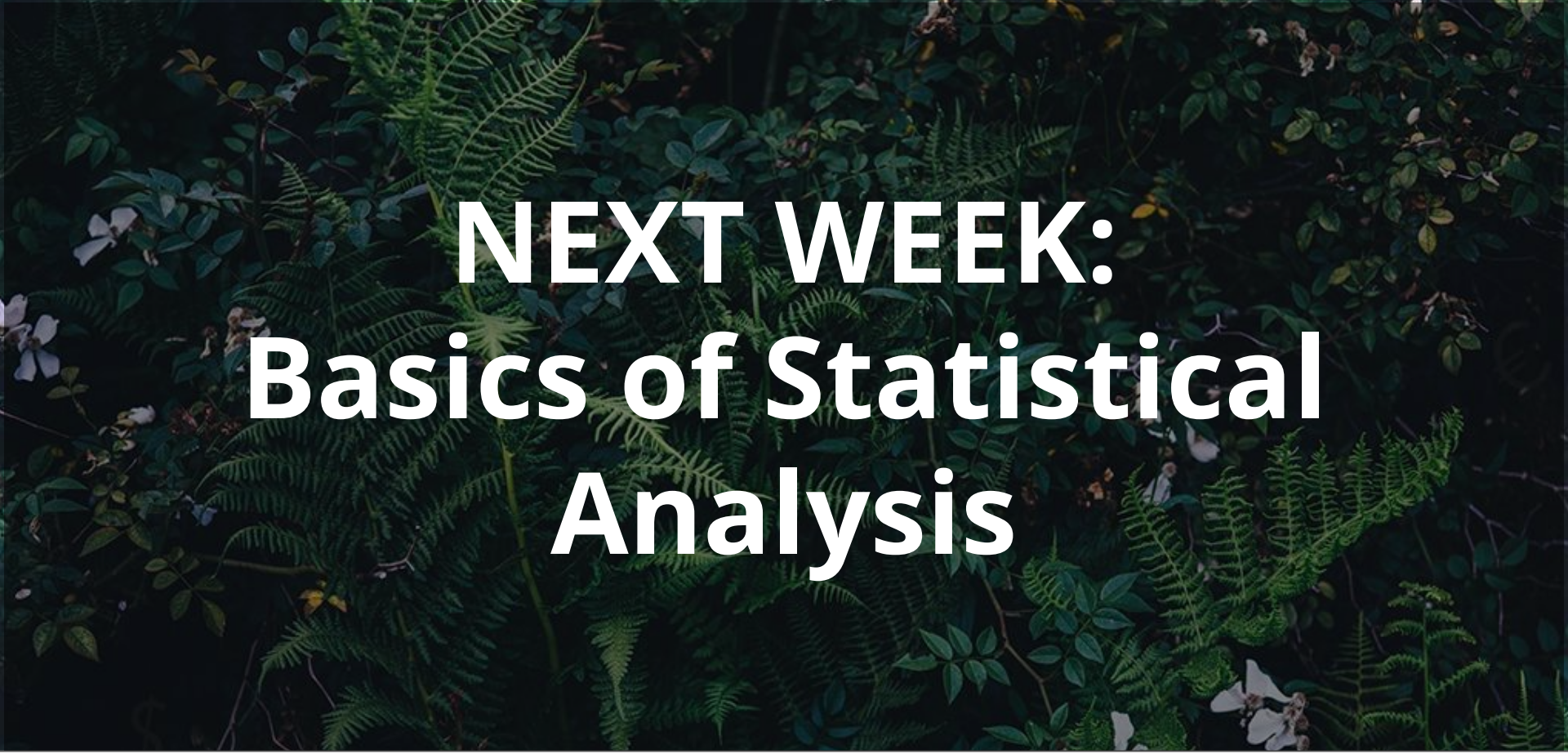
# Open Refine

http://openrefine.org/download.html

DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.




DATENSCHULE

OpenBudgets.eu

OpenBudgets.eu has received funding from the European Union's H2020 EU research and innovation programme under grant agreement No 645833.