

Opening Lecture (week 1)

Online Data Collection & Management (2025/2026)

Hannes Datta

2025-10-20

Welcome to oDCM!

We're about to start with the first lecture of this class.

If you haven't done so, please check out the **Canvas page** for this course:

- explore the *course syllabus*
- complete the *software installation* (see “preparation before this class on Canvas)
- check out today's slides on the *modules* page on Canvas

Agenda

- Part 1 (10.45 to about 11.45)
 - Getting to know each other
 - Motivation for the course
 - Course framework and learning goals
 - Agenda and practical arrangements
- Break
- Part 2: Python Bootcamp on your laptops (about 12.00 – 13.45/14.00)

This course in a nutshell

- You will learn how to **write code that automatically downloads and structures information from the internet for the purpose of (scientific) analysis.**
- We call these programs “web scrapers” (for any internet pages) and “APIs” (for official data access)
- Web scraping are the foundation of Google Search (“web spiders”) and ChatGPT (e.g., for training); APIs are at the core of many business models (e.g., Twitter API – back in the days; OpenAI API)
- I also almost got sued doing scraping (more about it later...)

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects
- This is predominantly about web scraping and APIs — while I teach a bit of Python, becoming an expert requires years of practice

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects
- This is predominantly about web scraping and APIs — while I teach a bit of Python, becoming an expert requires years of practice
- About me

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects
- This is predominantly about web scraping and APIs — while I teach a bit of Python, becoming an expert requires years of practice
- About me
 - I record my lectures and post them on Canvas (remind me if I miss to post one—or forget to start the recording)

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects
- This is predominantly about web scraping and APIs — while I teach a bit of Python, becoming an expert requires years of practice
- About me
 - I record my lectures and post them on Canvas (remind me if I miss to post one—or forget to start the recording)
 - Consider me your coach

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects
- This is predominantly about web scraping and APIs — while I teach a bit of Python, becoming an expert requires years of practice
- About me
 - I record my lectures and post them on Canvas (remind me if I miss to post one—or forget to start the recording)
 - Consider me your coach
 - Slow me down when needed

Disclaimer

- Mix of lectures, teamwork and self-study — all are necessary
 - I do the tutorials, lectures and course coordination
 - Roshini Sudhaharan will coach you on the team projects
- This is predominantly about web scraping and APIs — while I teach a bit of Python, becoming an expert requires years of practice
- About me
 - I record my lectures and post them on Canvas (remind me if I miss to post one—or forget to start the recording)
 - Consider me your coach
 - Slow me down when needed
- You will have to invest a lot of time and energy (but the rewards are high!)

About myself

- scraping nerd — learned it in 2008 using Visual Basic in Excel
- started doing my own research with scraped and API-extracted data in 2012 (so, 10+ years experience)
- left Germany around your age, now 17+ years in NL
- Associate Professor at Tilburg University

Key areas of expertise

Substantive interests

- streaming business models (e.g., music, movies)
- marketing-mix modeling and optimization
- open science

Methodological interests

- online data collection via APIs and web scraping
- causal effects with observational data

Getting to know you

- How did dPrep go for you (e.g., programming)?
- Any experience in Python already?
- What are your passions & talents? (+ why I am asking you this...)

Motivation for the course

- started out as a PhD student without data
- was interested in music, and found website with data (<https://last.fm>)
- no best practices in scraping; learnt all by myself and made many mistakes
- scraping was undervalued in academic job market — **but** — key role in shaping relevance and rigor of your work
- now scraping and APIs are a large part of what defines my research

Selection of scraping projects

- scraped reviews at Amazon.com
- how music consumption changed with Spotify
- Spotify new releases monitor
- power imbalances in the music industry
- playlist ecosystem data
- 100k+ images from Amazon, Google Vision/NLP API
- video streaming wars (Netflix vs Disney+)
- methodological framework on scraping/APIs
- faced legal battles...

What is scraping, and what are APIs?

Web scraping: anything you can view in a web browser

- pricing data at [bol.com](#)
- reviews at [Amazon.com](#)
- movie data at [imdb.com](#)

APIs: official interfaces by firms for programmatic data access

- e.g., Instagram, Twitter/X, ChatGPT, AWS
- researchers use them to construct datasets from analytics firms


Introducing music-to-scrape.org

- Mock-up streaming service
- Developed and released as part of Guyt et al. (2024)
- “Safe” and controlled environment to learn scraping and APIs

We are a fictitious music streaming service without real data - use us to learn collecting data with web scraping and APIs.

MusicToScrape Home Learn how to scrape Learn how to use API More Resources ▾

Search for artists, songs, and users



We love being scraped.

We're a fictitious music streaming service with a real website and API. Built for educational purposes, you can use us to learn web scraping!

Recently Played

Quick web scraper in Python (I)

```
import requests  
url = 'https://music-to-scrape.org/'  
webrequest = requests.get(url)
```

Quick web scraper in Python (II)

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(webrequest.text)
weekly15 = soup.find('section', {'name': 'weekly_15'})
for song in weekly15.find_all('h5'):
    print(song.text)
```

Gabriel Yared

Danny Williams

Pascal Obispo

Vangelis

Solas

Joi

Stevie Ray Vaughan And Double Trouble

Magnatune Compilation

Mint Condition

Enslavement Of Beauty

Quick APIs in Python

```
import requests
api_request = requests.get('https://api.music-to-scrape.org')
api_request_json = api_request.json()
for song in api_request_json.get('chart'):
    print(song.get('name'))
```

El Ventilador

Decide

The Song Remains The Same (2007 Remastered Album Version)

Everyday

Boogaloo Wow

Opportunities with web data

For businesses

- connect to (AI) services (e.g., use ChatGPT in your own products)
- market research (e.g., retrieve pricing data regularly)
- build recommendation systems or search (e.g., for a web shop)

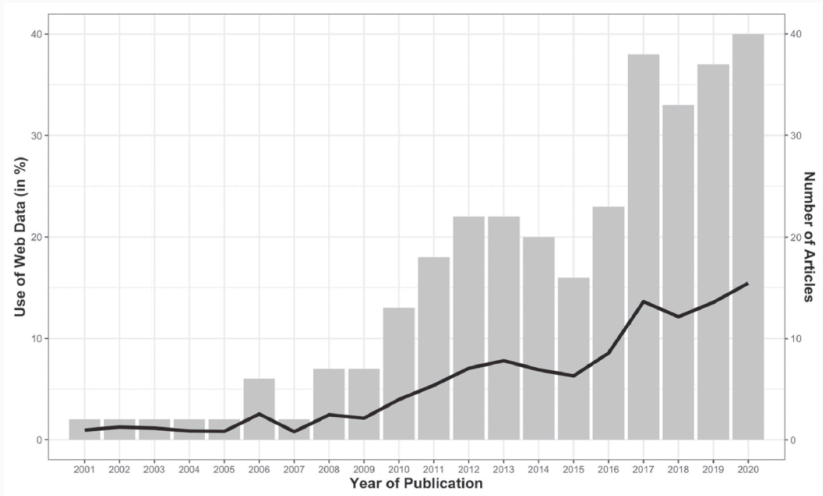
For research

- document novel phenomena without formal data access
- improve methods (text, image, video data, ...)
- achieve more accurate inferences by getting better control variables
- collect real-world metrics managers care about

Getting inspired

- What's the last app/website that made you say “wow”?
- Which three apps/websites would you definitely remove in the next month?
- Which three apps/websites would you definitely keep for the next year?
- What's a niche online community you're part of?
- What's the last thing you saw on TikTok that made you stop scrolling?
- Imagine an AI tool that makes you famous overnight — what would it do?

Why care (as a marketing researcher...)



Web data versus other marketing data (I)

Yes, but collecting *web data* is different!

- Finding the right data source (many exist)
- Different formats (website vs API vs CSV)
- Access to data that's not publicly available

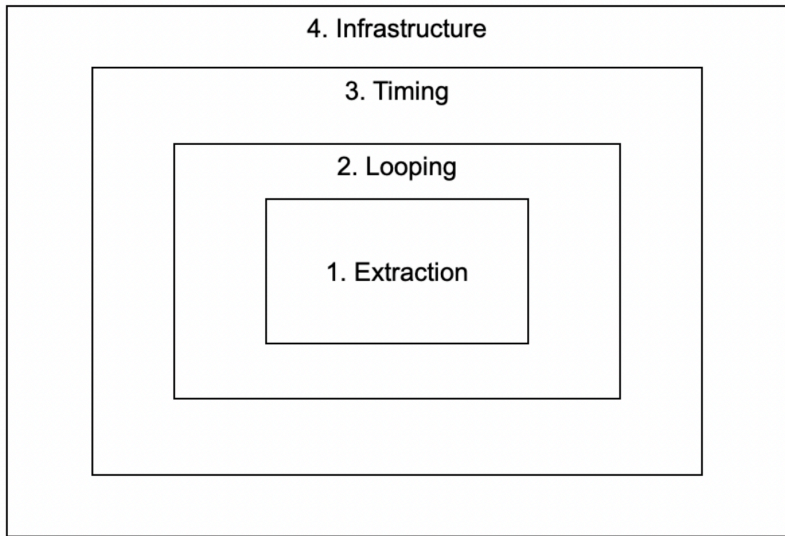
Web data versus other marketing data (II)

- Extraction design:
 - Which info to select
 - What variables exist
 - Legal and ethical considerations

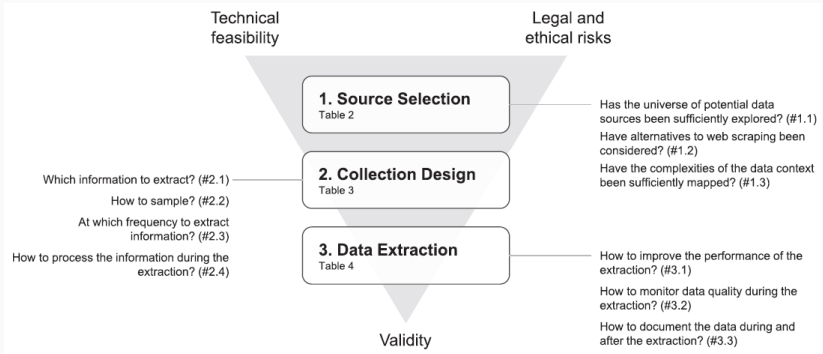
Web data versus other marketing data (III)

- Collecting at scale
 - Fully automatic but error-prone
 - Monitoring required
 - Poor documentation, unclear generalizability

Structured approach to data collections



Detailed guidance (Boegershausen et al. 2022)



Learning goals

- Explain how to use web data for **creating marketing insight**
- **Select data sources** and **evaluate** their value
- **Design web data collections** balancing **validity, feasibility, ethics**
- **Collect data** via scraping and APIs
- **Document and archive** for public reuse

Positioning in the study program

Research workflow



Data Collection

- **Online Data Collection & Management**
- Experimental Research

Substantive courses

- Brand Management
- Marketing Channels
- Marketing Communication
- Strategic Marketing Management

- Market Assessment
- Conjoint Analysis
- Survey Research
- Customer Analytics
- Pricing and Revenue Analytics
- Social Media and Web Analytics
- Pricing and Monetization
- Marketing models

Course structure I: Overview

- Weeks 1-3: Tutorials (Hannes)
- Weeks 4-7: Team Project (coach Roshini) + two lectures (Hannes, weeks 4-5)

Course structure II: Tutorials

- Five tutorials:
 - **This week:** Python Bootcamp
 - Week 2: Web scraping 101 (static) + Web scraping advanced (dynamic)
 - Week 3: API 101 (data extraction) + API Advanced (OpenAI's API)
- Structure of each tutorial
 - interactive walk-through (“semi-lecture”) with a Jupyter Notebook posted on Canvas (no slides)
 - discussion how to solve and accompanying you to learn coding to solve questions
 - solutions available

Key responsibility: **Hannes Datta**

Course structure III: Lectures

- provide **theory and context**...
- in the form of
 - lectures, and
 - recorded webclips

key responsibility: **Hannes Datta**

Course structure IV: Team project

- Collect data (scraping/APIs)
 - Apply frameworks (Boegershausen et al. 2022, Guyt et al. 2024)
 - Collect data for a research project
 - Online coaching sessions (10-15 minutes per team)
 - 5 group members (4 = exception)

Key responsibility: **Roshini Sudhaharan**

Course structure V: Team project (AI Use)

Level of AI allowed for this assignment: AI-assisted idea generation and structuring (Level 3 on AI Index Tilburg University)

- You are allowed to use generative AI tools to develop or refine initial ideas, materials, paraphrasing, structures, or outlines.

This includes generating code, e.g., for Python.

- Failing to declare AI use, or using AI beyond what is allowed in the syllabus, may be considered fraud and will be reported to the Examination Board.
- Keep a simple “logbook” documenting which AI tools you have used and for what purposes.

We will share several course-specific chatbots with you on Tilburg University’s chatbot platform Tilly.

Assessment (6 ECTS version!)

- **Computer exam** (60%, 120 minutes) → on campus, with internet
 - Mix of open + closed questions
 - Example questions will be shared on Canvas; exam Q&A in week 6 + during the final lecture in week 7
- **Team project (40%)** → submission on Canvas + self- and peer-assessment

Repeating this course for 3 ECTS? Please register on Canvas as such (6 ECTS Canvas page, group: 3 ECTS repeater)

Tips & tricks

- Familiarize yourself with this course (e.g., syllabus) on Canvas
- Early weeks are toughest, but skills build quickly
- Collaborate and help each other
- Recall that coding can be frustrating & tiring + you're learning a new language (Python) → take breaks, get our support

→ **Quick support?** Make use of our chatbots.

Steps of escalation

1. Use chatbot
2. Check course site
3. Google / StackOverflow
4. Ask classmates
5. Feedback sessions
6. If urgent → contact me

WhatsApp



+31 13 466 8938 Email = slower

What's in for you?

- Research skills (relevant & rigorous work)
- Entrepreneurial skills (data-based business)
- Coding showcase

Any questions so far?

Next steps

- Find a group and register by week 2 (4–5 students, mix skills!)
- Complete software installation & obtain licenses (includes premium Datacamp access)
- Most up-to-date info: Canvas → (weekly) modules

After a well-deserved break → **Python bootcamp**.