UNIVERSITÄT
LEIPZIG

DATA PREPARATION & CLEANING

Chapter 2: Data Extraction

Victor Christen

AGENDA

- General

Characteristics of Data Extraction

- Internal vs External
- ETL vs ELT
- Data Warehousing
- Extraction methods
  - API
  - Web Crawler & Scraping
  - Content specific extraction
  - Extraction from unstructured data
- Summary

DATA EXTRACTION - GENERAL

“Data extraction is the process of **collecting** or **retrieving** disparate types of data from a **variety of sources,** many of which may be **poorly organized** or **completely unstructured**.”[1]

- **Internal** vs **external** data source
- **ETL** vs **ELT**
- **Structured**
  - Standardized format, well defined access (query language)
- **semi/unstructured**
  - Websites, scanned and OCR-processed text documents, transcripted Audio Reports, etc.

Extraction methods depend on the data format and the type of data sources

[1] https://www.talend.com/resources/data-extraction-defined/

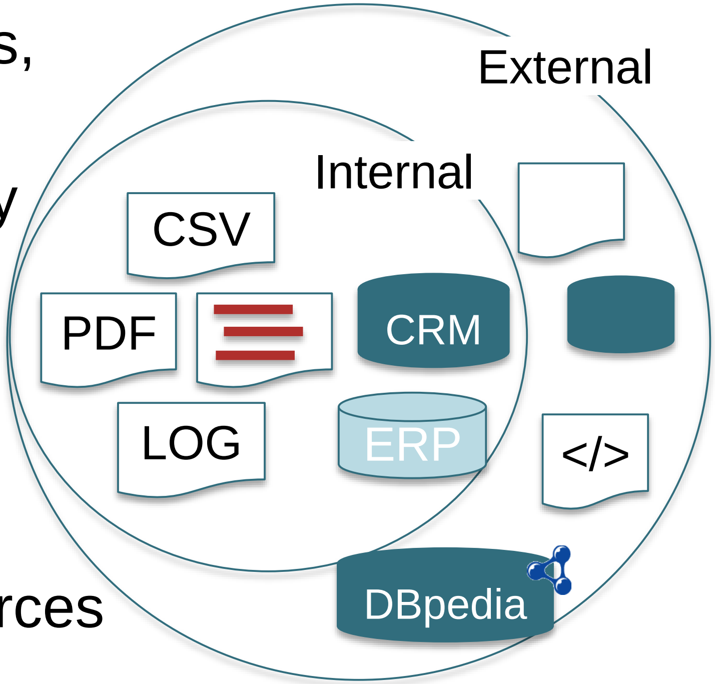
EXTERNAL VS INTERNAL DATA SOURCES

Internal

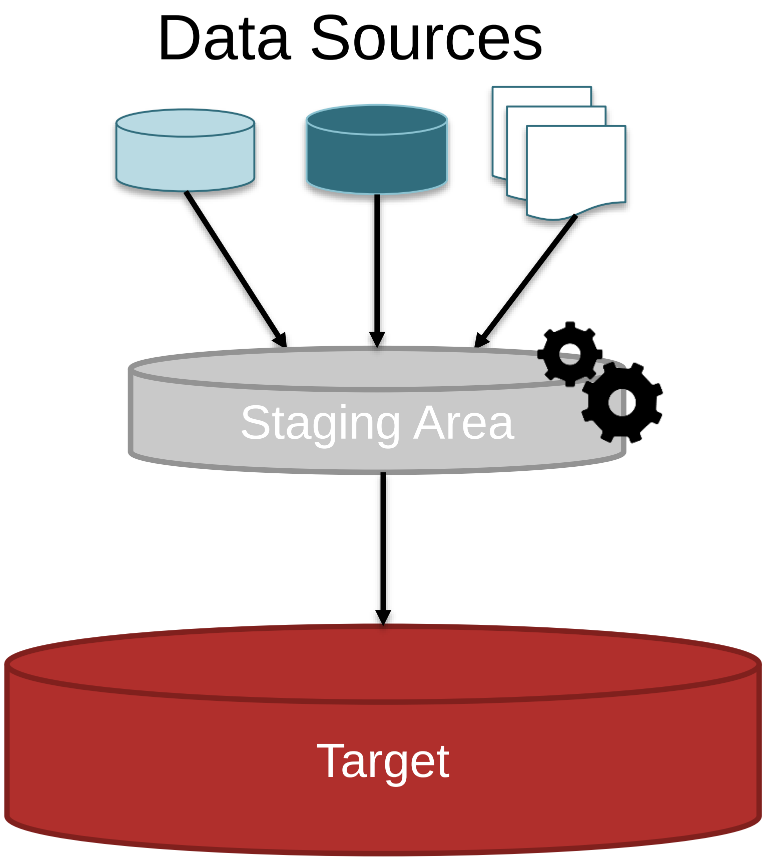
- Databases (CRM, ERP), documentations, Log-files, Emails
- View on internal processes of a company
- Completely autonomous w.r.t. the data source
- flexible access possibilities

External

- Context information for internal data sources
- Websites, API, SPARQL-endpoints,…
- Access potentially limited by the functionality of an API, availability of resources



ETL VS ELT



Extraction

- Extract data from data sources

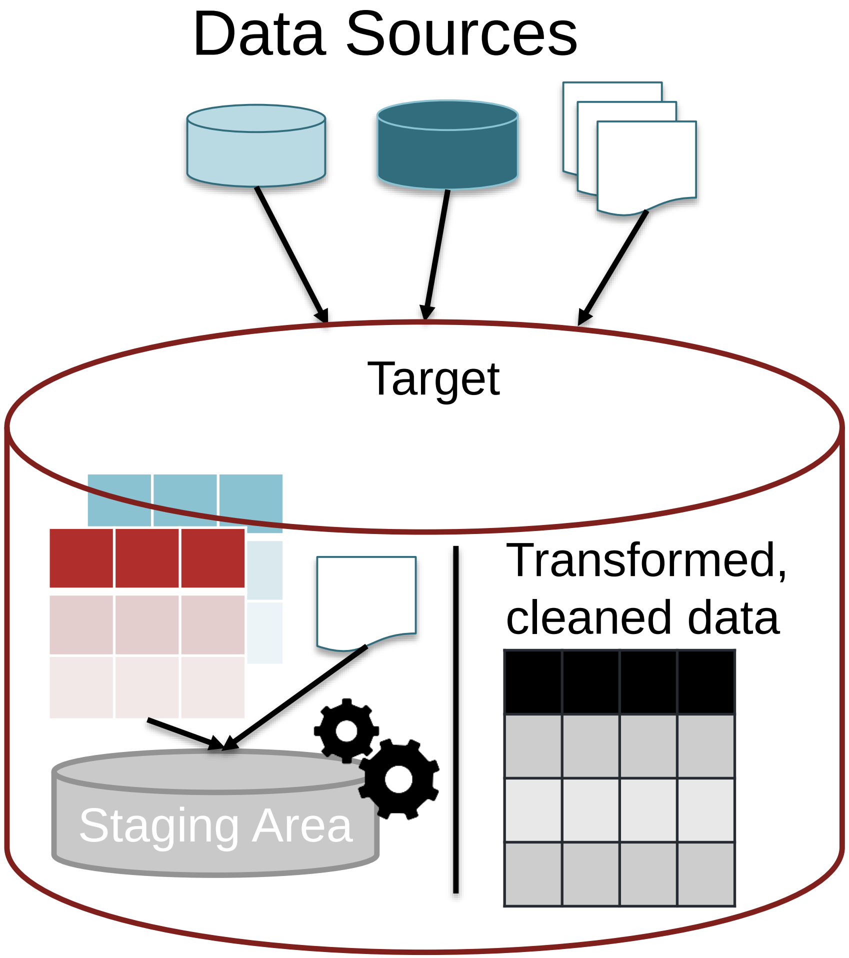
Transformation

- Transform, clean and integrate data

Loading

- Load transformed, cleaned and integrated data in target database for analysis

ETL VS ELT (2)



Extraction

- Extract data from data sources

Loading

- Load the extracted unprocessed data in the target data source, e.g., **Data Lake**

Transformation

- Transform and clean only the **relevant** data **before** analysis

ETL VS ELT – PROS & CONS

ETL

Target analysis tasks are known, e.g. product sales, human resources, production rate

+ Transformation before enables fast analysis

– Not scalable for large and complex transformation tasks

– Not flexible regarding changed analysis goals

ELT

Direction of analysis is known but not in detail

+ Flexible regarding analysis tasks and data evolution

+ Store all data → flexible regarding new analysis tasks

○ High performance using cloud infrastructure by using additional resources on demand

ETL - DATA WAREHOUSING

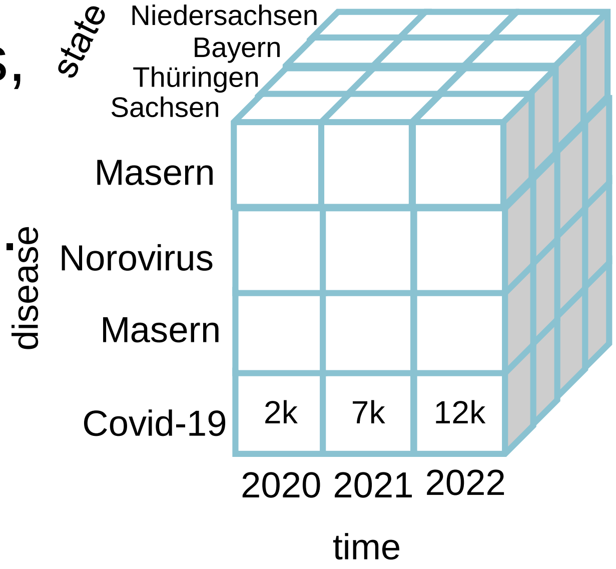
- Specialized database to support company decisions
- Disconnected from operational systems
- Periodic ETL process
  - Support time dependent analysis
  - Initial loading after that mostly reading operations
- Focus on specific target analysis such as product sales, citation numbers, …
  - Described by different dimensions

Reverse ETL

- Use Analysis result in operational databases

DATAWAREHOUSE - COMPONENTS

- Data-Cube
  - Consisting of Dimensions and measures
- Measures are numerical values (counts, sales, etc.)
- Context information by dimensions (e.g. Time, place, name of disease)
- Dimensions represent coordinates for measures
- Hierarchical structure of dimensions possible



DATAWAREHOUSING - OPERATIONS

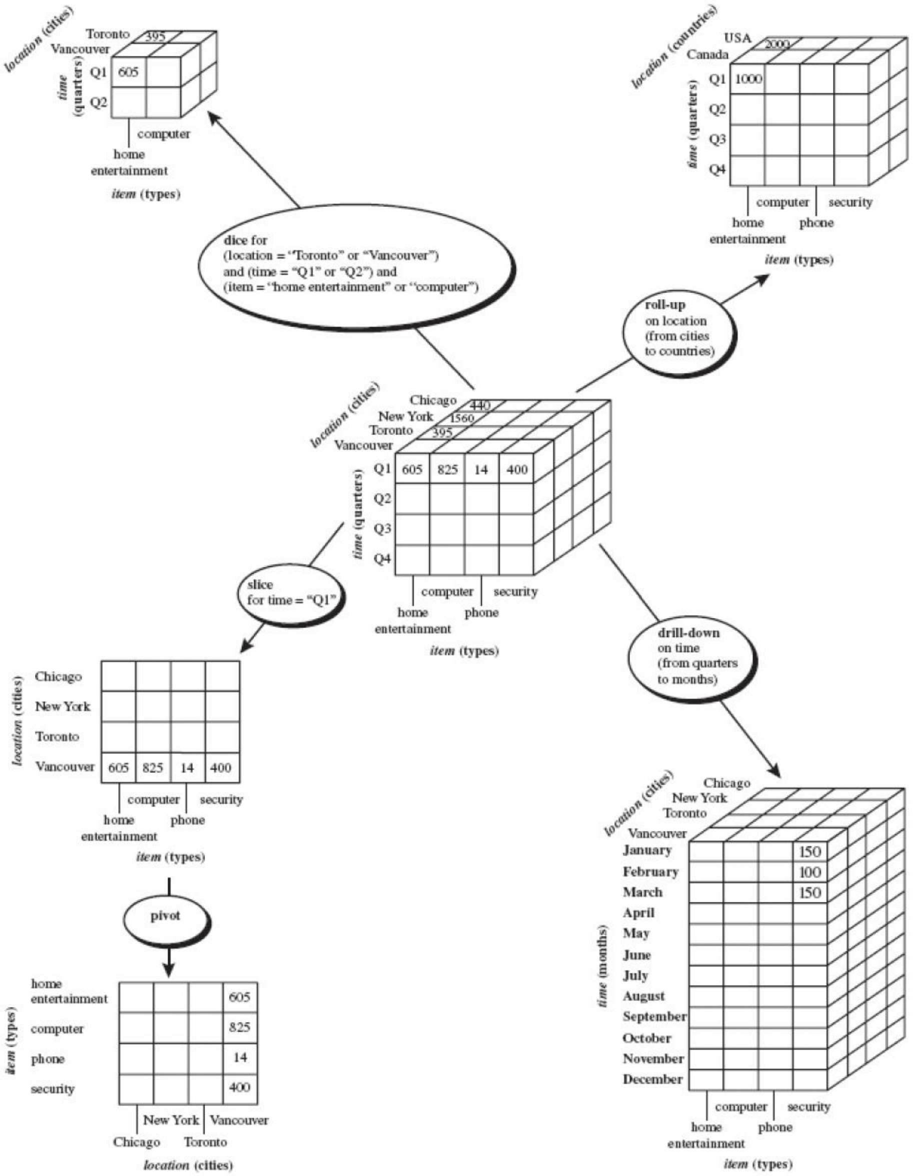
Operators for analysis

- Roll-up (Aggregation of fine-granular hierarchies)
- Drill-down or Roll-down (detailed view)
- Slice and Dice (Restriction of one resp. multiple dimensions)
- Pivot (Switch orientation of dimensions)

Applications

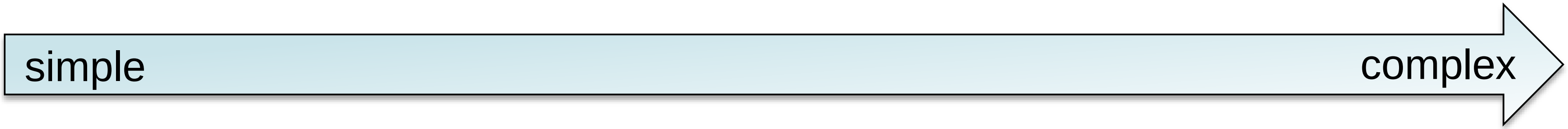
- Generation of statistics, reports, charts, etc.)
- Base to generate Data Mining models

DATAWAREHOUSING - OPERATIONEN



Quelle: Han and Kamber, DM Book, 2 nd Ed. (Copyright © 2006 Elsevier Inc.)

TYPES OF EXTRACTION METHODS



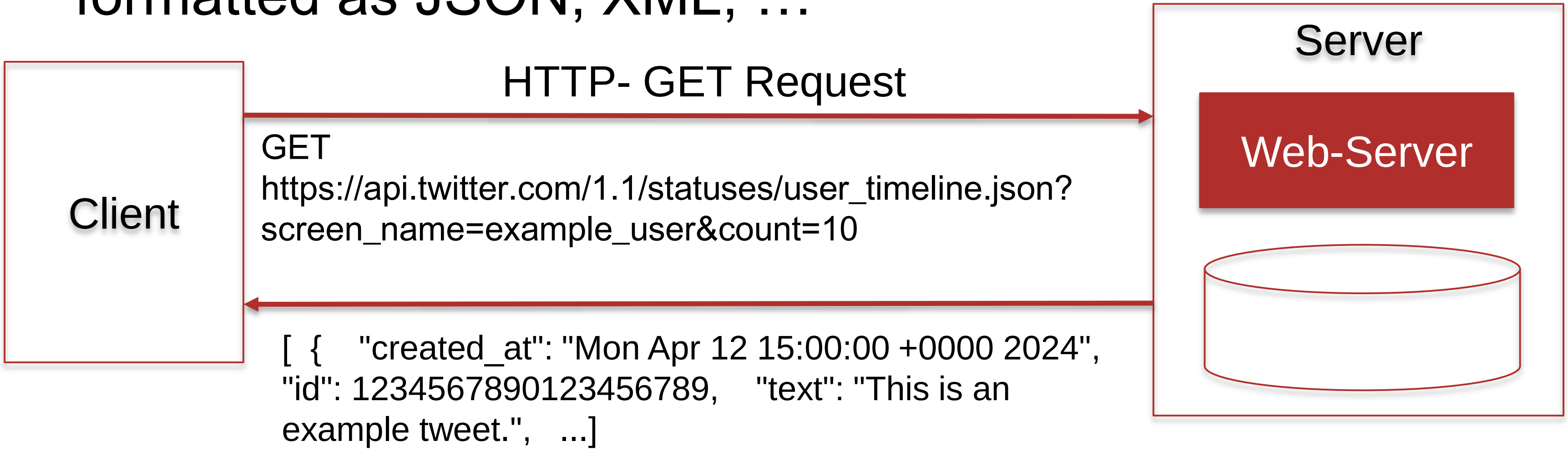
- Using a predefined Query Language
  - SQL, Cipher, SPARQL,
- Process relational data, e.g. CSV/TSV following a strict schema using program specific libraries such as Pandas, tool connectors

- Structure is defined by the data itself, e.g., XML/JSON
- Arbitrary set of attributes
- Potentially complex extraction methods required depending on the content of an attribute
- Query Language, API, Web Crawler/Scraping

- Data is completely unstructured, e.g. free formatted text, Images, Audio
- Requires Advanced methods such as Named Entity Recognition, OCR, speech-to-text
- Content specific libraries → tables

REST - APPLICATION PROGRAMMING INTERFACE – REST API

- Website providers enable access to their resources by a REST-API
- Predefined set of access functionalities called by HTTP requests to retrieve the current state of resources
- Server responds by returning a semi-structured result formatted as JSON, XML, …



Client

HTTP- GET Request

```
GET
https://api.twitter.com/1.1/statuses/user_timeline.json?
screen_name=example_user&count=10
```

Server

Web-Server

```
[  {    "created_at": "Mon Apr 12 15:00:00 +0000 2024",
"id": 1234567890123456789,    "text": "This is an
example tweet.",    ...]
```

API - GEOCODING

- Extraction of geo coordinates from address data for the visualization and analysis of location data
- Reverse determination of the corresponding address with regard to a geo coordinate

Applications

- Epidemiological research to determine disease spread clusters or analyze geographically related health problems

GEOCODING - METHODS

- Use of a reference data set with corresponding mapping → often not available

API

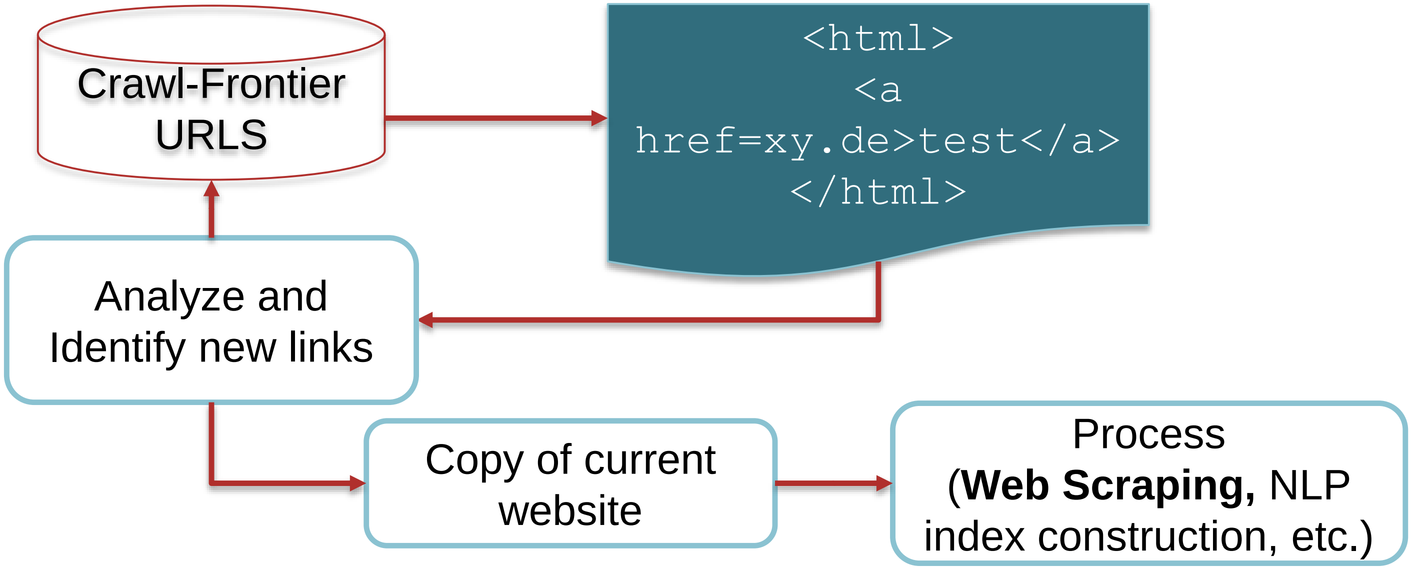
- Standardized access to geo location data
- e.g. GeoPy, geocoder
- Exact match
  - 1:1 mapping address-coordinate
- Fuzzy Match
  - Specification of a region if there is no exact match with address

DATA EXTRACTION FROM WEBSITES

Web- Crawler

- Collect & Process **recursively** links and the content from websites based on the hyperlink structure

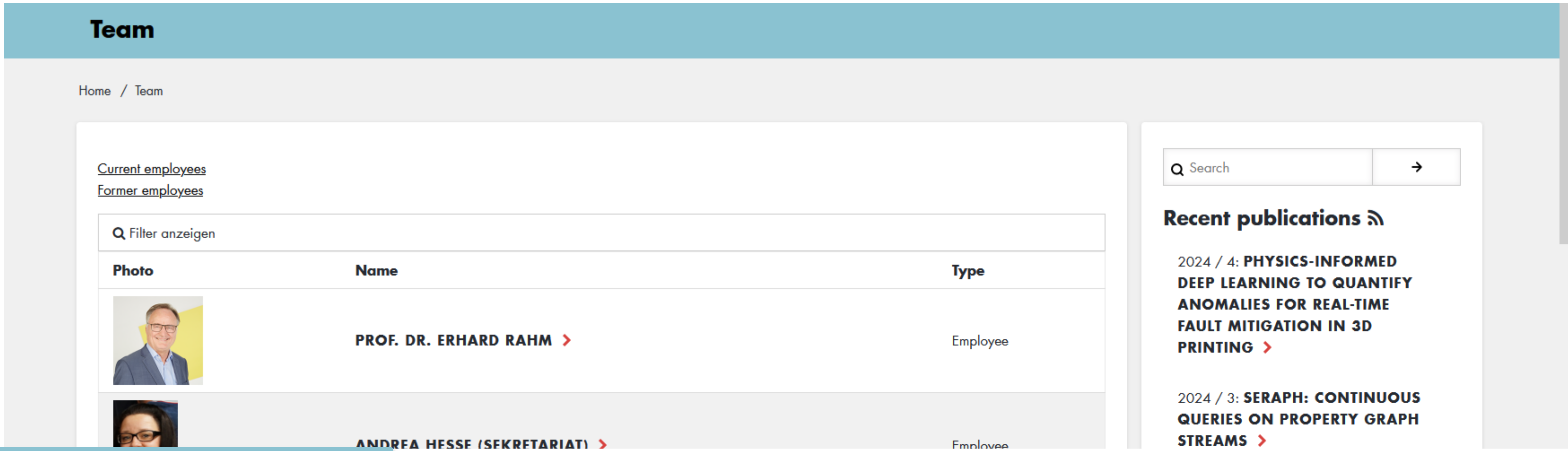
Priority-Queue
Score based on the
Up-to-dateness



```
<html>
<a
href=xy.de>test</a>
</html>
```

- Periodical execution to notify changes of collected websites

WEB SCRAPING



Team

Home / Team

Current employees
Former employees

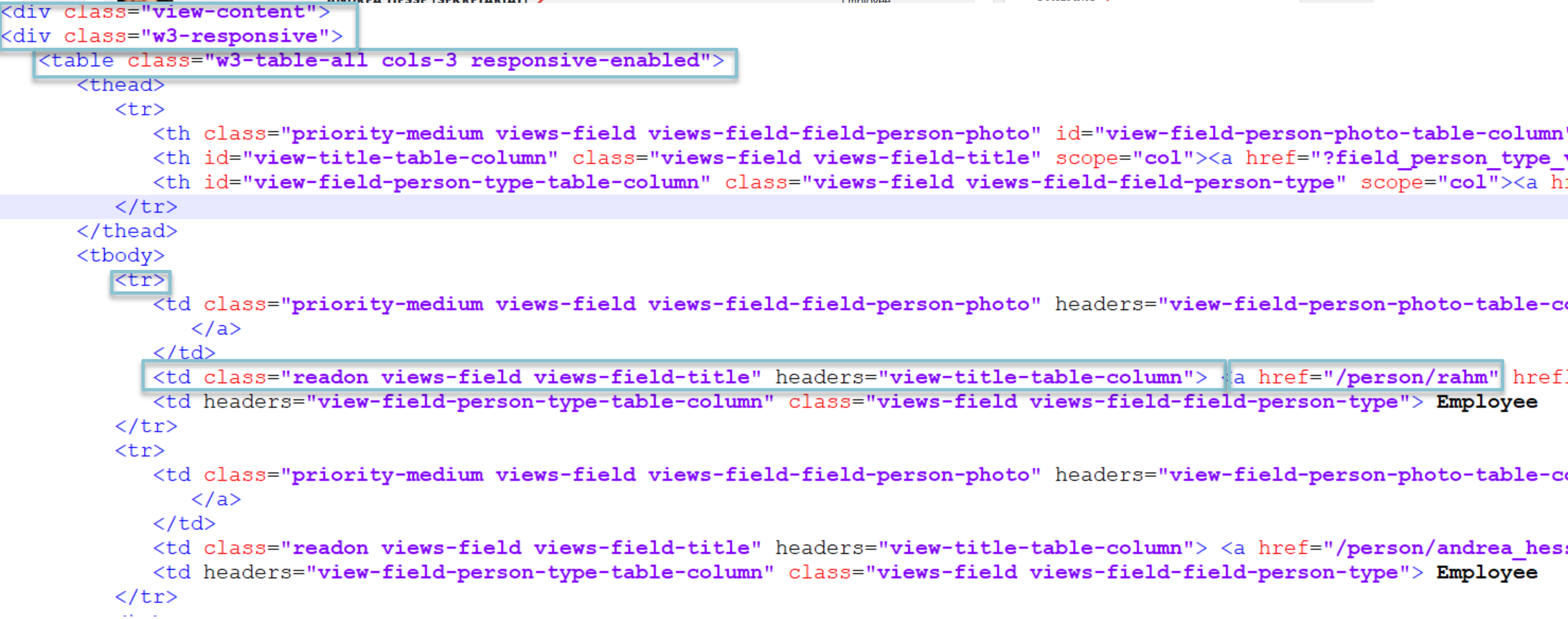
Filter anzeigen

Photo	Name	Type
	PROF. DR. ERHARD RAHM ›	Employee
	ANDREA HESSE (SEKRETARIAT) ›	Employee

Recent publications

2024 / 4: **PHYSICS-INFORMED DEEP LEARNING TO QUANTIFY ANOMALIES FOR REAL-TIME FAULT MITIGATION IN 3D PRINTING ›**

2024 / 3: **SERAPH: CONTINUOUS QUERIES ON PROPERTY GRAPH STREAMS ›**



```
<div class="view-content">
<div class="w3-responsive">
   <table class="w3-table-all cols-3 responsive-enabled">
      <thead>
         <tr>
            <th class="priority-medium views-field views-field-field-person-photo" id="view-field-person-photo-table-column"
            <th id="view-title-table-column" class="views-field views-field-title" scope="col"><a href="?field_person_type_
            <th id="view-field-person-type-table-column" class="views-field views-field-field-person-type" scope="col"><a h
         </tr>
      </thead>
      <tbody>
         <tr>
            <td class="priority-medium views-field views-field-field-person-photo" headers="view-field-person-photo-table-co
               </a>
            </td>
            <td class="readon views-field views-field-title" headers="view-title-table-column"> <a href="/person/rahm" hrefl
            <td headers="view-field-person-type-table-column" class="views-field views-field-field-person-type"> Employee
         </tr>
         <tr>
            <td class="priority-medium views-field views-field-field-person-photo" headers="view-field-person-photo-table-co
               </a>
            </td>
            <td class="readon views-field views-field-title" headers="view-title-table-column"> <a href="/person/andrea_hess
            <td headers="view-field-person-type-table-column" class="views-field views-field-field-person-type"> Employee
         </tr>
```

WEB SCRAPING - LIBRARIES

- Extraction of specified elements via libraries, e.g. Scrapemark[1], Scrapy[2], BeautifulSoup[3] and the definition of website-specific element patterns

```
from bs4 import BeautifulSoup
import requests

webFile = requests.get("https://dbs.uni-leipzig.de/de/person")
soup = BeautifulSoup(webFile.content, 'html.parser')
first_table = soup.select_one("div.w3-responsive table.w3-table-
all.cols-3.responsive-enabled")
list = soup.select(
"tr:has(> td.readon.views-field.views-field-title) a")
```

- `soup.tag` returns the selected value with tags, without by using `.string`
- `element.find_all(element_name)` returns a list of specified elements regarding the selected ancestor element

[1] https://github.com/arshaw/scrapemark
[2] https://scrapy.org/
[3] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

WEB SCRAPING – CHALLENGES & RESTRICTIONS

- Evolution of websites → adaption of patterns
- Classical Scraping limited to static content, JavaScript content require evaluation
- Anti-Scraping mechanisms have to be considered and must not be circumvented
  - Access limitation (number of requests for a certain time interval)
  - Captcha
  - IP-Blocker

Legal restrictions

- Guarantee of copyrights depending on the use case → research vs commercialised application

WEB SCRAPING - TOOLS

User-Friendly Interface

- Specify relevant elements through a graphical user interface

Proxy Support

- Avoid IP bans by proxies and rotation mechanism

JavaScript Rendering

Extraction of JavaScript generated content requires rendering functionalities

Scheduled Scraping

- Definition of crawling/scraping intervals

Compliance and Ethical Considerations

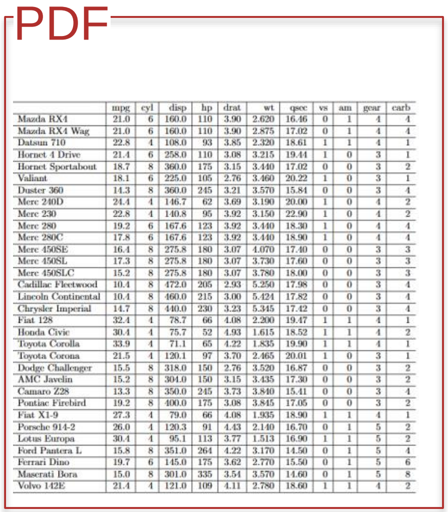
- Ensure compliance with website terms of service, legal regulations, ethical considerations, e.g., respecting robots.txt directives, avoiding excessive requests

EXTRACTION FROM SPECIFIC CONTENT TYPE

Tabular data

- Programming language specific libraries ,e.g., `tabula-java, tabula-py`
  - Automatic extraction of table data in a PDF-document
  - specification of pages, area, output options

PDF

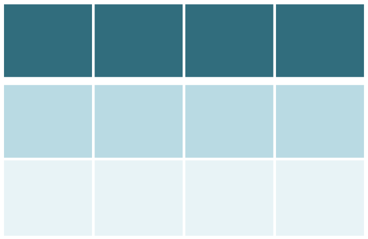


	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
Cadillac Fleetwood	10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
Lincoln Continental	10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
AMC Javelin	15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
Camaro Z28	13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
Porsche 914-2	26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

```
import tabula
pdf_path =
"https://github.com/chezou/tabula-
py/raw/master/tests/resources/data.pdf"

dfs = tabula.read_pdf(pdf_path,
stream=True)
# read_pdf returns list of DataFrames
```

Programming language specific table representation



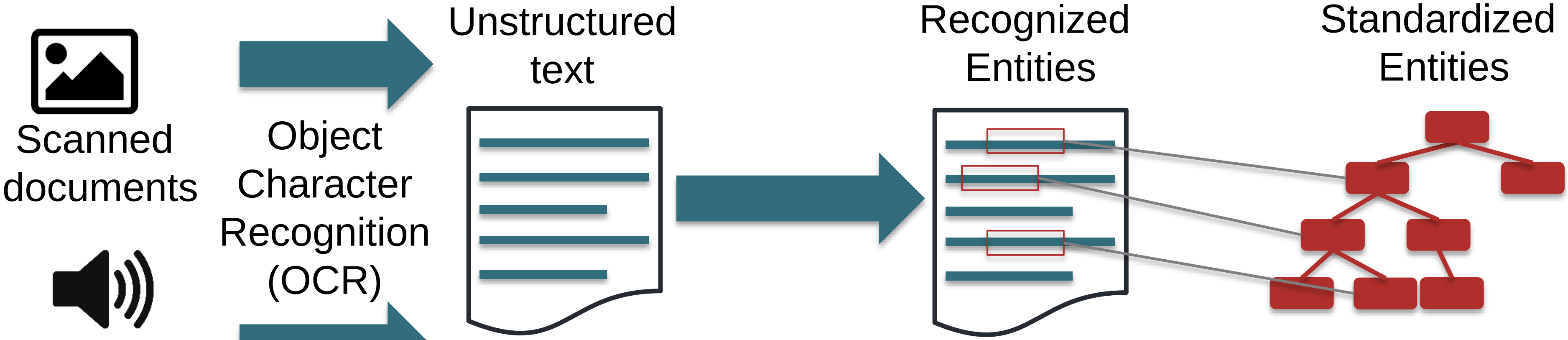
EXTRACTION FROM UNSTRUCTURED DATA

Manual

- Handcrafted rules and background knowledge, e.g. Extracting author information from publications
- High quality of extracted data being correct
- Potential high number of missing data
  - Depending on the heterogeneity
- Very sensitive regarding changing content structure → High adaption effort

AUTOMATIC EXTRACTION

– Combine multiple extraction methods tailored to the certain task



Named Entity Recognition (NER)

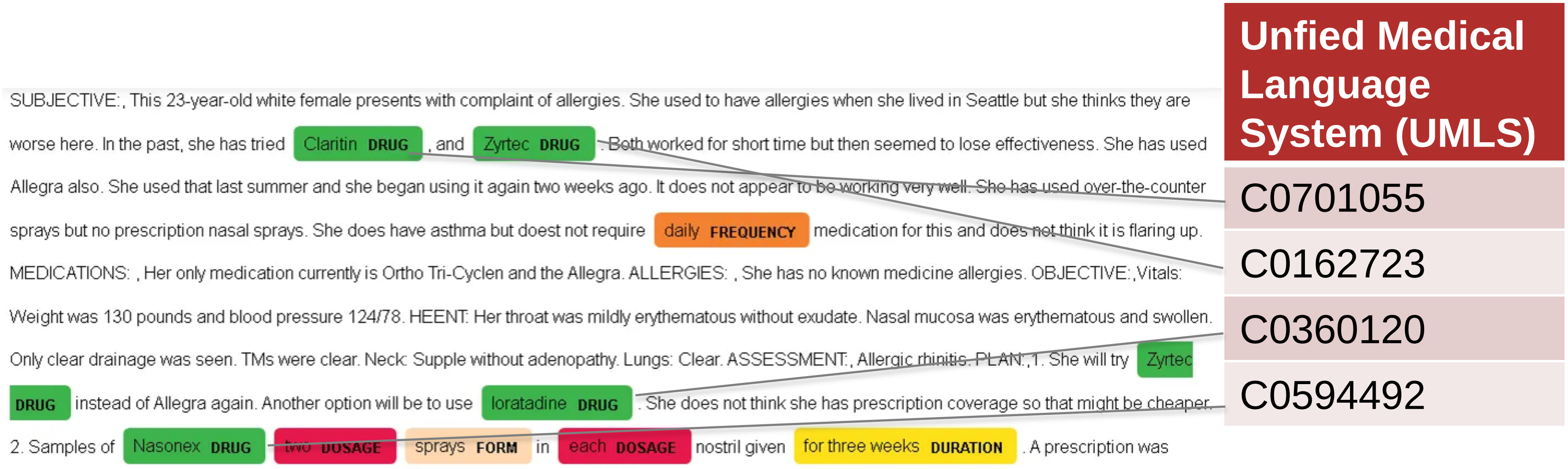
- Identification of text boundaries of an named entity
- Utilizing domain-specific NER-model trained on a corpus from the same domain

Entity Linking

- Link identified mentions to standardized entities of a common knowledge base such as ontology

AUTOMATIC EXTRACTION - EXAMPLE

- Processing medical documents such as electronic health record
- Enhance comparability by NER & EL



Standardized Entities

Unfied Medical Language System (UMLS)

C0701055

C0162723

C0360120

C0594492

SUBJECTIVE:, This 23-year-old white female presents with complaint of allergies. She used to have allergies when she lived in Seattle but she thinks they are worse here. In the past, she has tried Claritin DRUG , and Zyrtec DRUG . Both worked for short time but then seemed to lose effectiveness. She has used Allegra also. She used that last summer and she began using it again two weeks ago. It does not appear to be working very well. She has used over-the-counter sprays but no prescription nasal sprays. She does have asthma but doest not require daily FREQUENCY medication for this and does not think it is flaring up. MEDICATIONS: , Her only medication currently is Ortho Tri-Cyclen and the Allegra. ALLERGIES: , She has no known medicine allergies. OBJECTIVE:,Vitals: Weight was 130 pounds and blood pressure 124/78. HEENT: Her throat was mildly erythematous without exudate. Nasal mucosa was erythematous and swollen. Only clear drainage was seen. TMs were clear. Neck: Supple without adenopathy. Lungs: Clear. ASSESSMENT:, Allergic rhinitis. PLAN:,1. She will try Zyrtec DRUG instead of Allegra again. Another option will be to use loratadine DRUG . She does not think she has prescription coverage so that might be cheaper. 2. Samples of Nasonex DRUG two DOSAGE sprays FORM in each DOSAGE nostril given for three weeks DURATION . A prescription was

SUMMARY

- Internal vs External data sources
- ETL vs ELT
  - ETL: for small projects, complex transformation tasks potentially decrease the efficiency depending on the amount of data and heterogeneity, e.g., Data warehousing
  - ELT: transformation on demand considering a subset, High scalability using cloud environments, e.g., Data Lakes
    - increase resources on demand
- Data Warehousing
- Various extraction methods depending on the data source and format
  - Querying, processing relational data
  - API data extraction by predefined functions, e.g., Geocoding
  - Web crawler/web scraping
  - Content depending extraction methods, tabular data
  - Unstructured data requires complex and sophisticated methods such NER and entity linking