Statistical Scraping
Interest Group

WEB-FOSS-NL

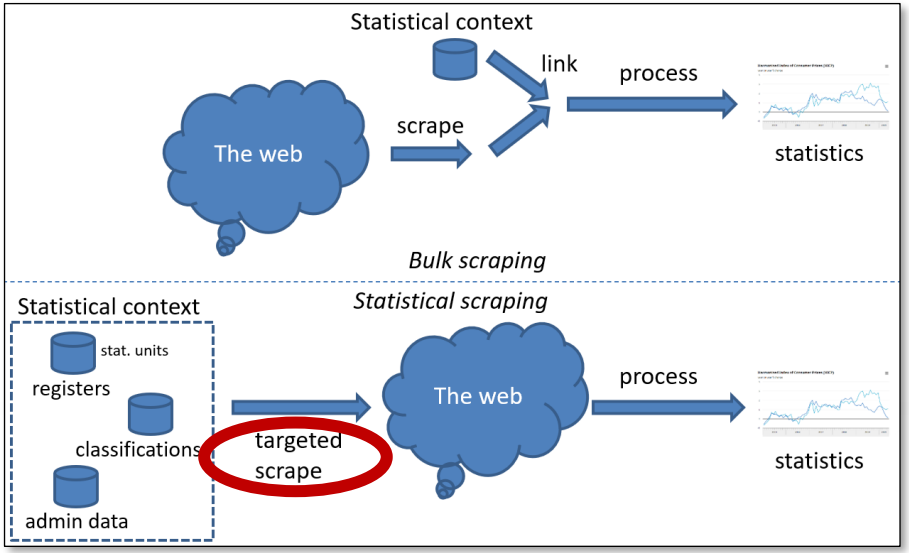
Statistical Scraping explored in the Netherlands

Olav ten Bosch, Luuk Haarmans, Dominik Blatt, Femke Bosman, Jacco Daalmans

SSIG1 – Statistical Scraping Interest Group meeting 1, 16-17 Sep. Vienna

Statistical scraping: high level view (2024)

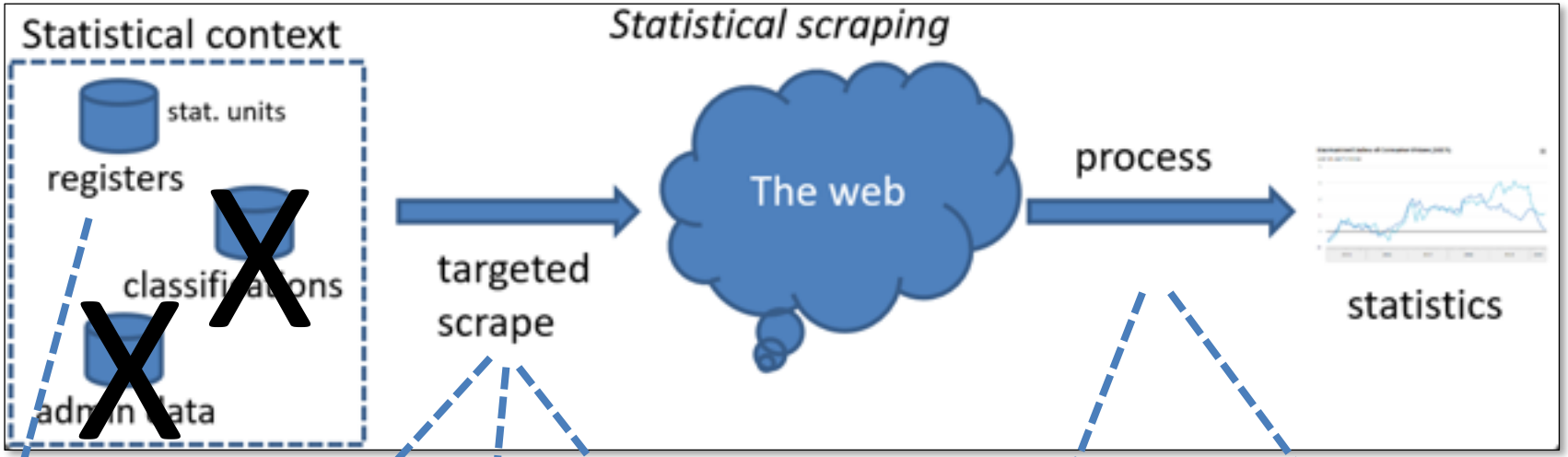
Def 1.1: *Statistical scraping is the use of online data starting from a-priori information in the respective statistical domain keeping a clear relation with the statistical context.*



Olav ten Bosch, Alexander Kowarik, Sonia Quaresma, David Salgado, Arnout van Delden, (2024), *Statistical scraping: informed plough begets finer crops*, European Conference on Quality in Official Statistics, Estoril, Portugal

link

Statistical Scraping concept applied to JV (1)



Statistical business register (NL)

Sample

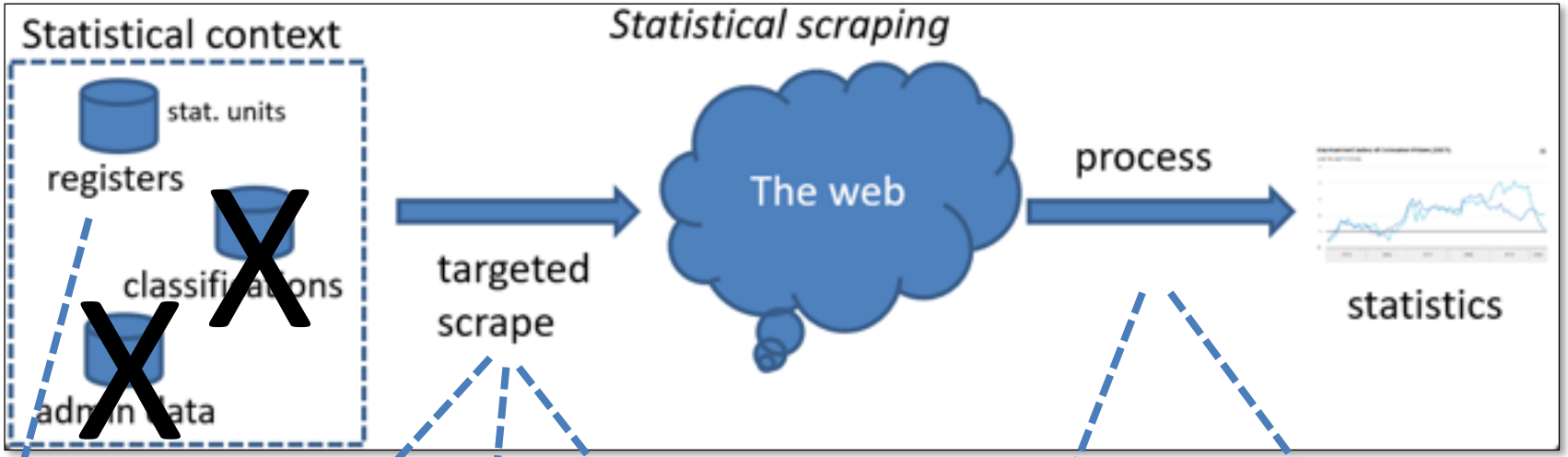
URL finding

Focused scrape

Interpret Text mining / LLMs

Experimental stats

Statistical Scraping concept applied to JV (2)



Statistical business register (ABR)

Sample (WP2)

URL finding (WP2)

Focused scrape (WP3)

Interpret Text mining / LLMs (WP3)

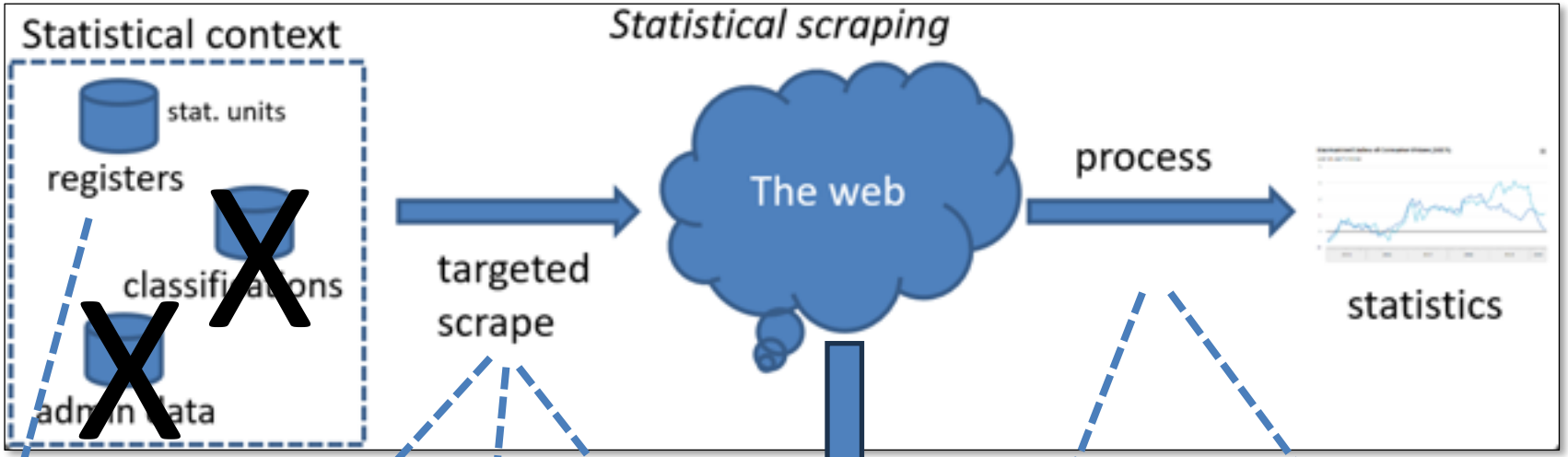
Experimental stats (WP3)

Validation

JVS

Official stats

Statistical Scraping concept applied to JV (3)



Statistical business register (ABR)

Sample (WP2)

URL finding (WP2)

Focused scrape (WP3)

Interpret Text mining / LLMs (WP3)

Experimental stats (WP3)

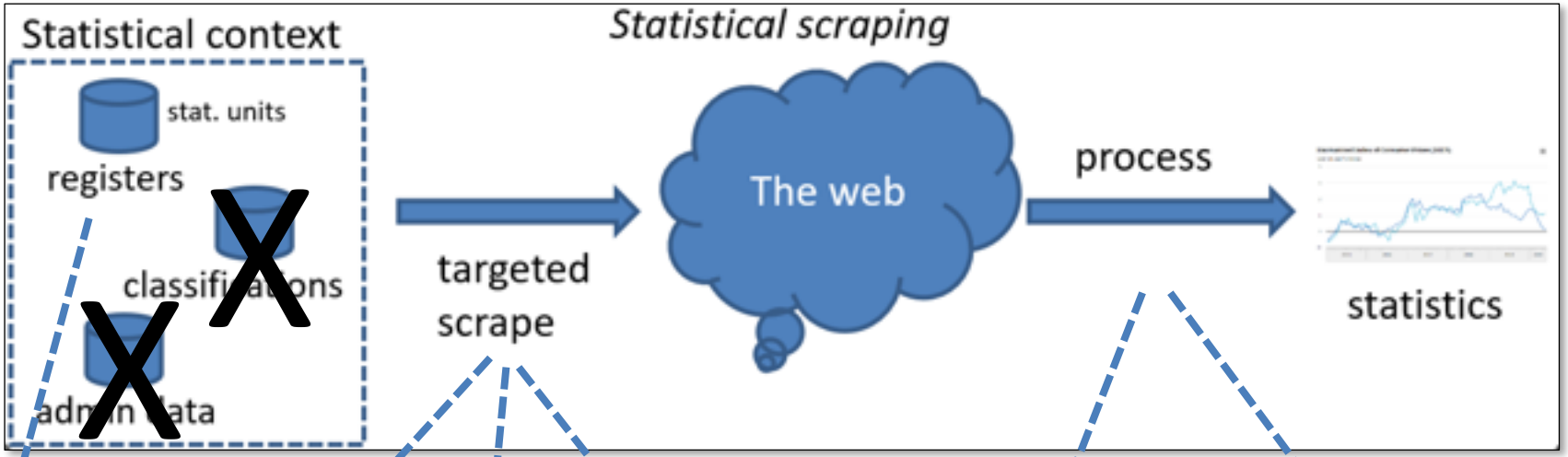
Validation

JVS

Validation/editing !!!

Official stats

Statistical Scraping concept applied to JV (4)



Statistical business register (ABR)

Sample (WP2)

URL finding (WP2)

Focused scrape (WP3)

Interpret Text mining / LLMs (WP3)

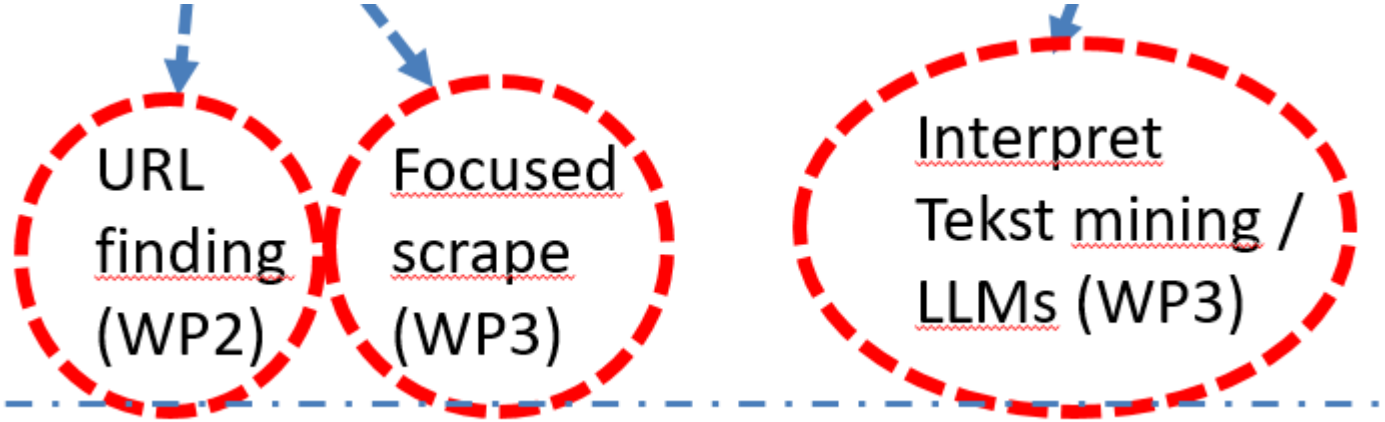
Experimental stats (WP3)

Validation

JVS

Official stats

Iteration plan 1st year

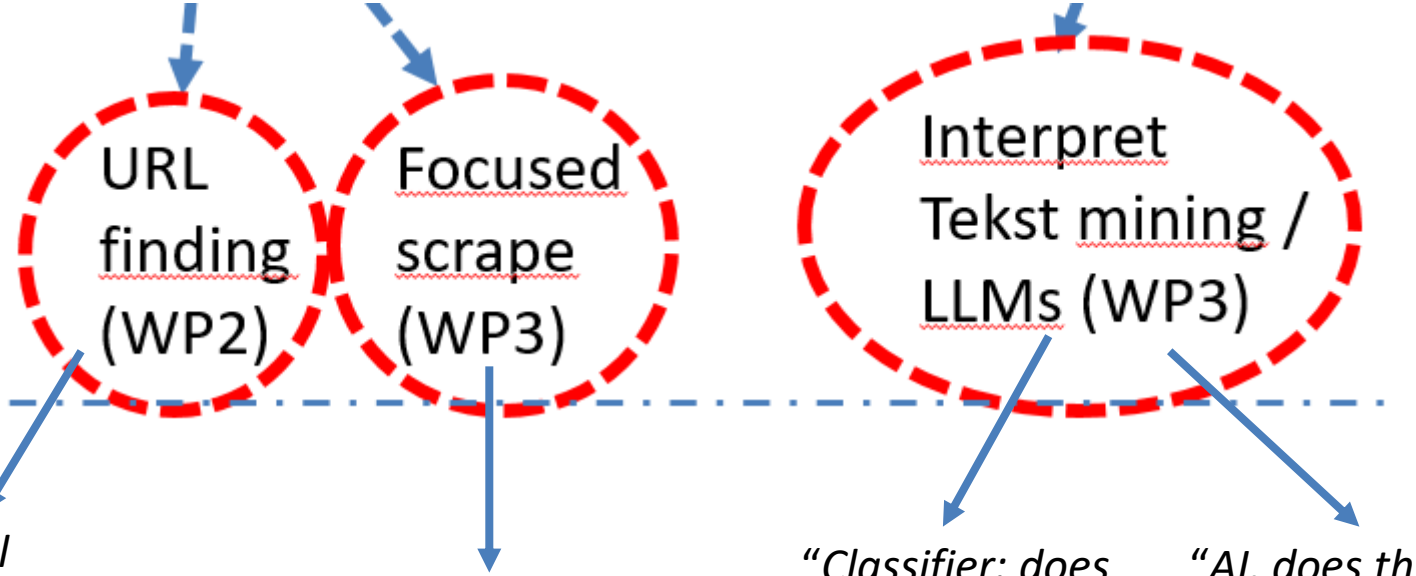


- Phase 1: develop building blocks ***independently***, make them runnable on fresh web data or fake data
- Note: WEB-**FOSS**-NL: **F**ree and **O**pen **S**ource **S**oftware
- Phase 2: Build an experimental process ***chain*** of building blocks
- Phase 3: test run on small subset from Statistical Business Register

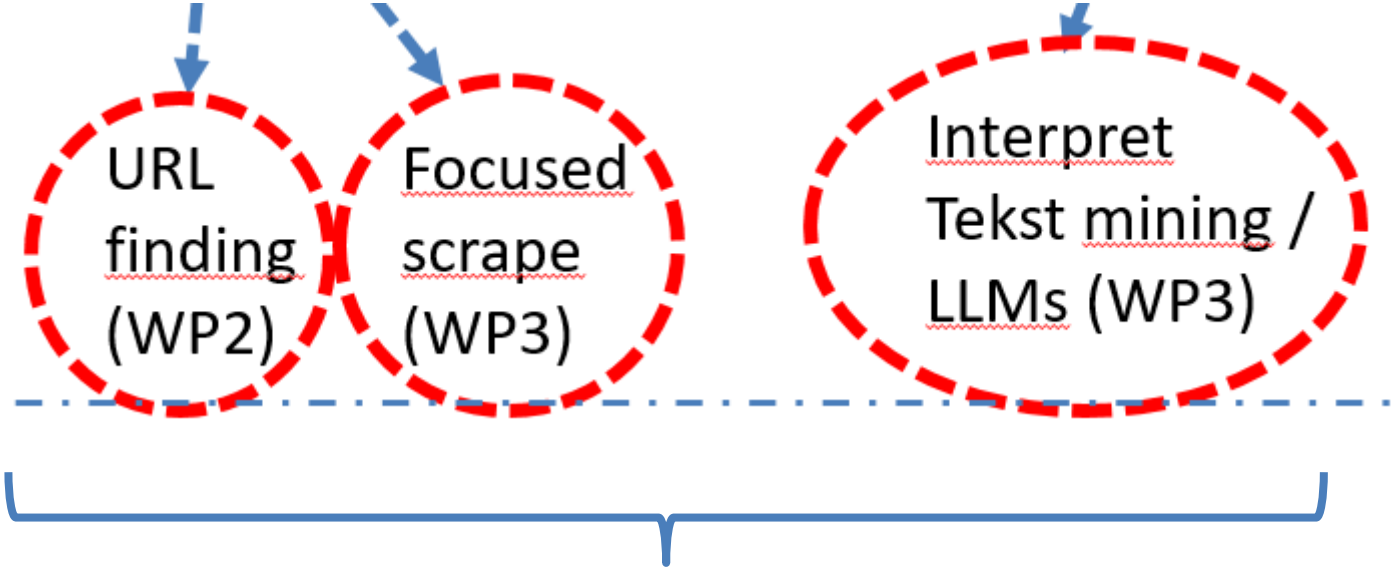
AI (?)

Some free thinking

Applying AI, scenarios (1)



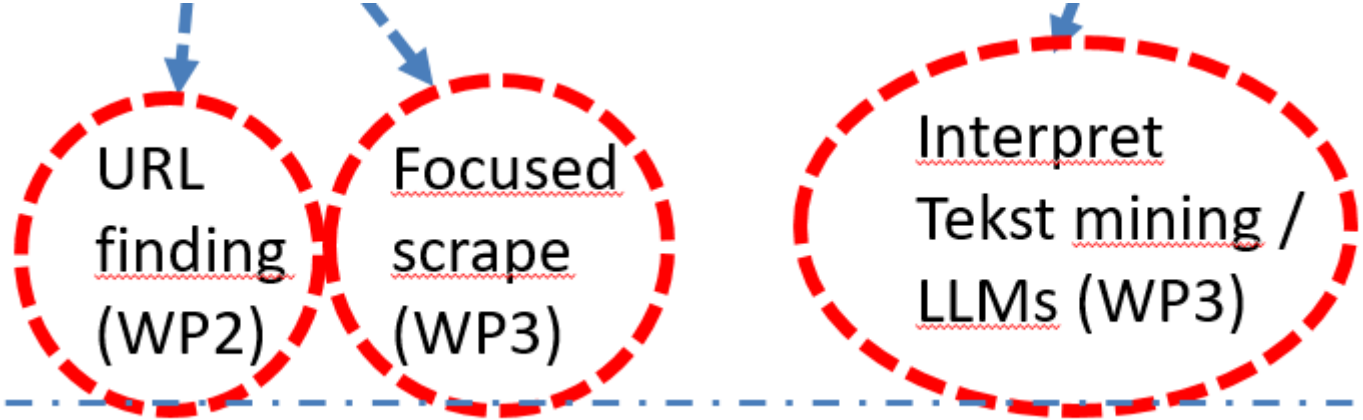
Applying AI, scenarios (2)



"Hey AI, tell me if enterprise X has a JV at this moment?"

(and how many and what types)

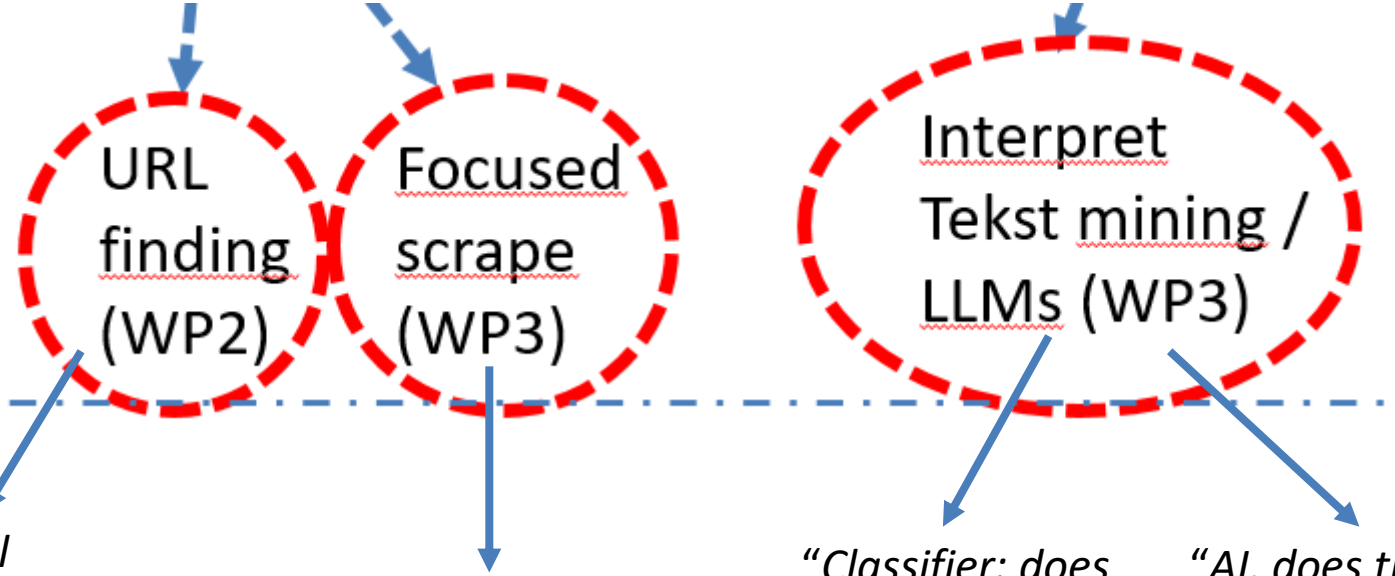
Applying AI, scenarios (3)



~~*"Hey AI, Just give me the latest JV statistics"*~~

Oh no! ☺ ☹ ☹

Applying AI, scenarios, reatively safe



"AI, list all websites where enterprise X publishes JVs"

"AI, which links on this home page are most likely to contain JVs?"

"Classifier: does this text represent a JV?"

"AI, does this web page contain JV(s)?"

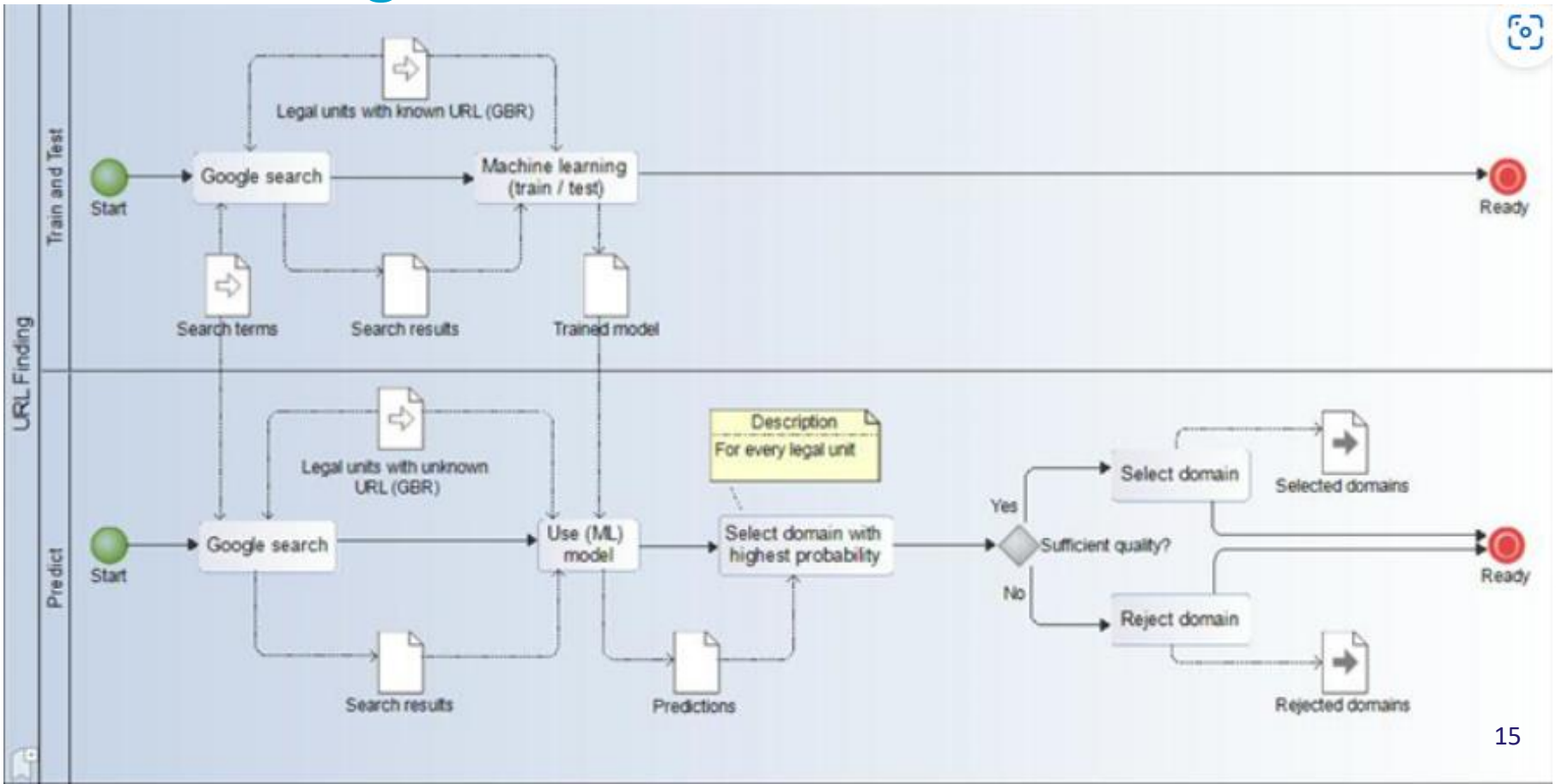
Test / validation sets

1. Whatever ML/AI scenarios, we need test and validation sets
2. Can we use the output of the WIN hackathon?
3. Can we use other JV texts?
   Public JV texts?
4. Any other ideas?

Building blocks in progress: URLfinding

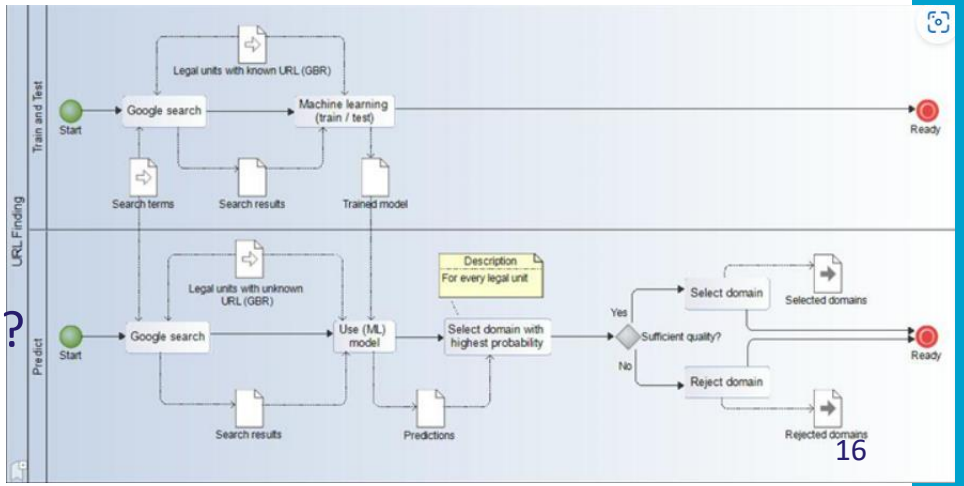
URL finding

https://github.com/SNStatComp/urlfinding



URL finding (2)

- https://github.com/SNStatComp/urlfinding
- Generic software for finding websites of enterprises using search
- 5 years old; refactoring; more simple
- Retraining of ML model Google API
- Other search engines? (DuckduckGo?)
- Open web search project EU?: https://openwebsearch.eu/
- Other web sources? spatial data (maps)?
- Your URL finding software (ideas)?



Building blocks in progress: Focused scrape

Goal

- **Scrape only** (all) the pages of a website that are likely to contain a **certain type of content**
- Our focus in this project: **Online job advertisements** (OJAs) of enterprises

- **Input**: URL of an enterprise, e.g., https://www.cbs.nl
- **Output**: Set of pages with potential OJA texts

Considerations

- Work in progress https://github.com/SNStatComp/webfocusedscrape
- Developing on Onyxia https://datalab.sspcloud.fr/
- Python code
- Re-use of established libraries
- Can we use sitemaps?

General questions (1/2)

- Can we develop an algorithm that is likely to work well for **all relevant websites in a country**, e.g. The Netherlands?
- Can a general library be developed that works for multiple countries and languages? What are **country-specific challenges**?
- How rewarding is the inclusion of **parameters** that relate to known/ expected **business properties**?
  - Industry
  - Organization size
  - Digitalization

General questions (2/2)

- **Depth**: How deep do we look, and much should we punish going down an unrewarding path?
- **Breadth**: And how broadly do we start?

- What is the **ideal output for** the next step in our pipeline: the **webtextclassifiier**? Raw HTML, structured data, clean text?

Keywords in URL

- How far do we get with a keyword-based approach?
- Can we reasonably implement it?
- NL
  - based on example of 68:

    vacatures, **werkenbij**, **jobs**, werken-bij, **careers**, vacature, work-with-us, work-at, solliciteren, recruitee
  - What else?

    loopbaan, banen, jobsat, join, joinus, joinat, talent, werken, werkenmet, team, people, future, next, lifeat, inside

Keywords other languages

- Austria

  `stellen, stellenangebote, karriere, jobs, arbeitenbei, arbeiten-bei, bewerbung, mitarbeiten, karriereportal, team, menschen, talente, zukunft, inside`

- Poland

  `praca, kariera, oferty-pracy, zatrudnienie, dołącz, dolacz-do-nas, rekrutacja, aplikuj, talent, zespol, ludzie, przyszlosc, inside`

- UK

  `jobs, careers, career, join, joinus, join-at, jobs-at, work-with-us, work-at, talent, people, team, future, next, lifeat, inside, apply, opportunities, hiring, openings, vacancies`

Keyword location in URL

- Subdomain

https://careers.companyname.nl

- Dedicated domain

https://careersatcompanyname.nl

- In the URL path

https://www.companyname.nl/careers

Building blocks in progress: Text mining / LLMs

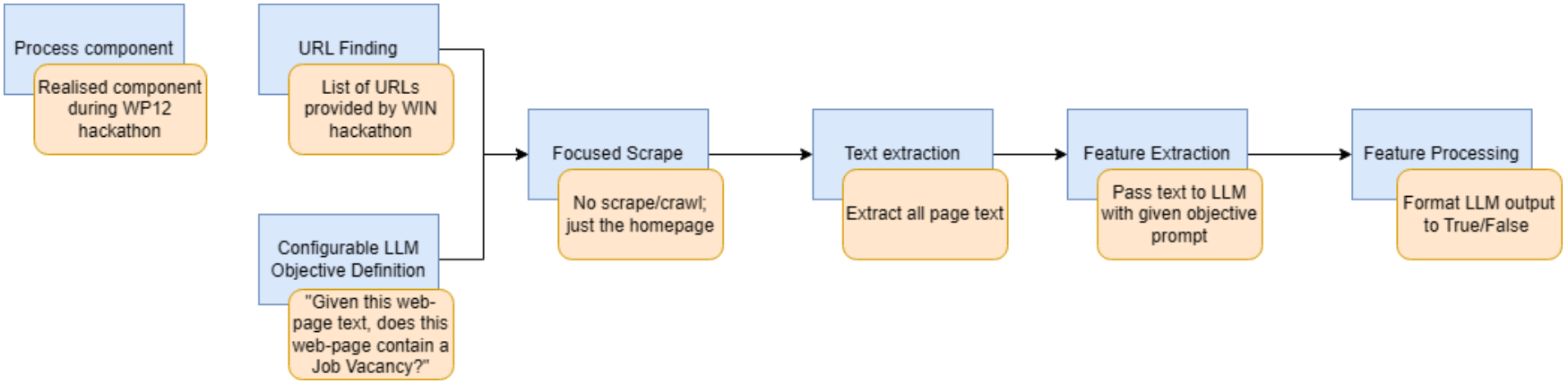
Text mining / LLMs

- Prompting:
  - “given a web page, does this page contain a JV?”
  - “given a web page, how many JVs are contained?”
  - “given a web page, how many JVs are contained and what type?”
- Building on work from earlier CBS project (LLM project)
- Zero-shot: No examples/labeled data for training
- Alternative: Feature extraction
  - Use LLMs to extract features from tekst
  - Pass features onto additional classifier models

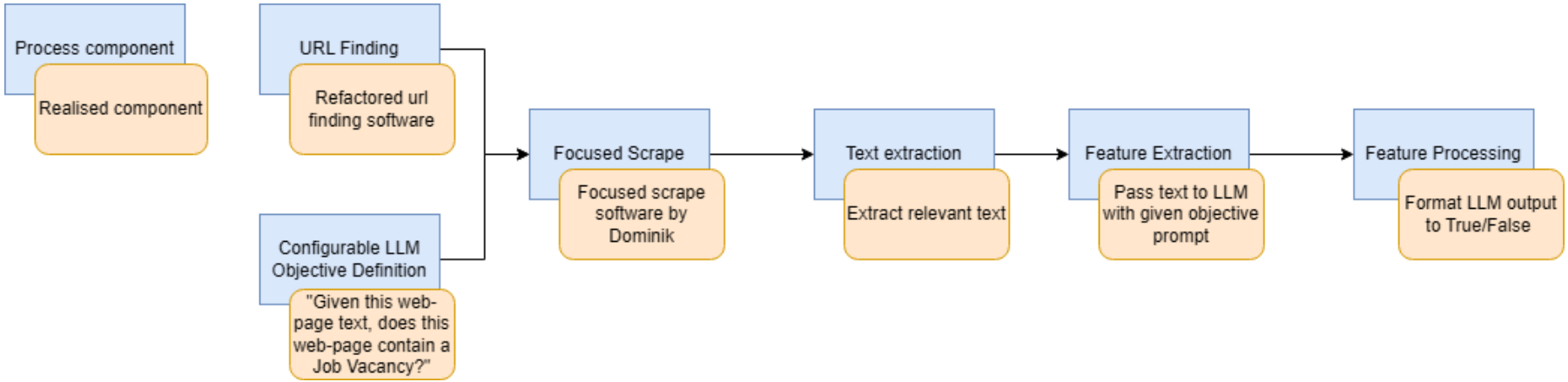
AI-ML4OS WP12 (LLM) Lisbon Hackathon

- Goal: Create a prototype for a reusable LLM application
  - Identify webpages that relate to a certain topic
    - JVs, Hotels, Sport clubs, etc.
  - Extract information that can be used later
- Use of Onyxia platform of INSEE and github for collaboration
  - WP12/WP12_hackathon/web_corner
  - Development and LLM usage
- Provided base for more generic setup for WEB-FOSS-NL: https://github.com/SNStatComp/webtextclassifier

Initial Design AI-ML4OS WP12 Hackathon



Current Design for WEBFOSS-NL



Example output

- JSON file
  - (url: label) pairs
- Feature dictionary
  - (url: feature_vector) pairs

Example output

http:/www.flextra.nl
https://www.youngcapital.nl/

31

Current challenges

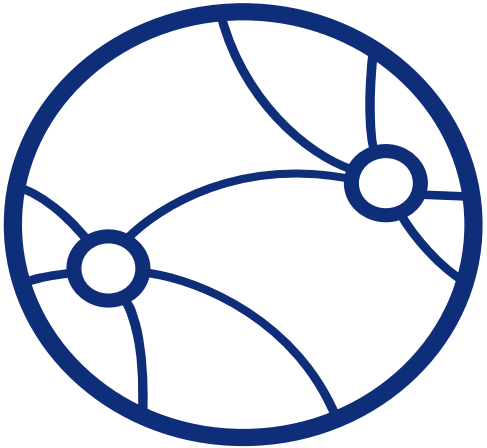
- Prompting:
  - Finding the best prompt
  - Generating consistent output in a specified format
- Can we use a multi-lingual approach?
  - Reusability
- Need for labeled training/validation data
  - Use of LLM vs more traditional ML approaches
- Scalability?
  - Might work well for 1000 pages, but also for 1000x1000?

Take aways & more to come…

SSIG1: 16-17 September in Vienna (now)

WEB-FOSS :experiment & help improve building blocks

SSIG2: March/April/May 2026 in The Hague



Statistical Scraping
Interest Group

https://github.com/SNStatComp/SSIG