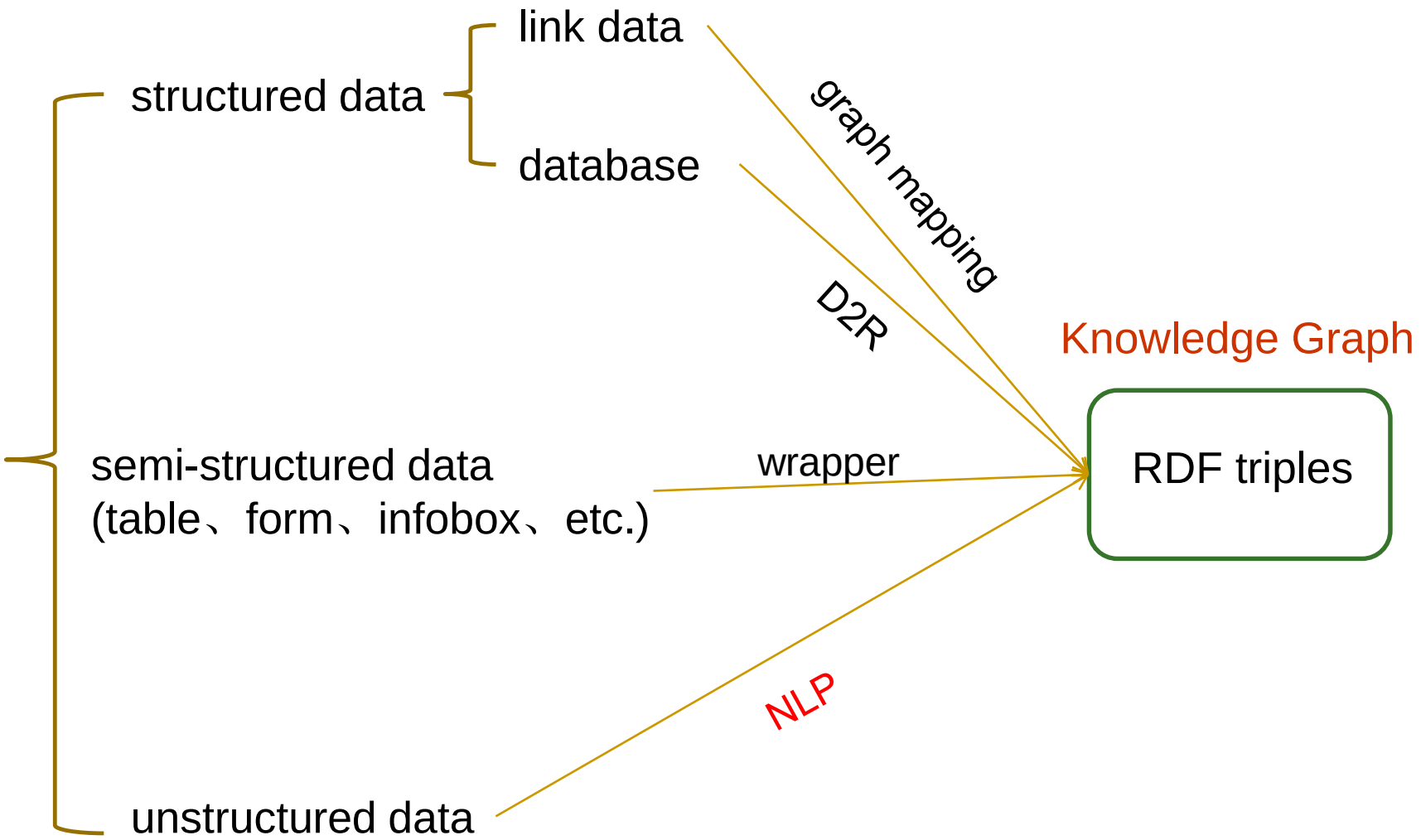
Knowledge Extraction & Mining

Hai Wan
School of Data and Computer Science
SUN YAT-SEN UNIVERSITY

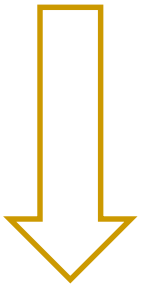
Thanks for Haofen Wang

Knowledge Extraction



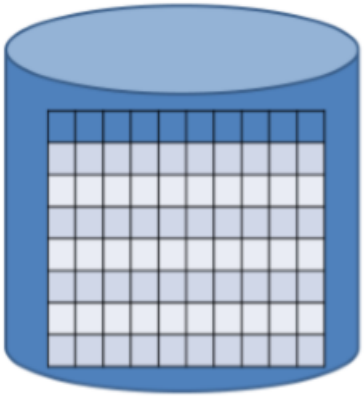
Knowledge Extraction

extracting triples from unstructured data(documents) is the most difficult



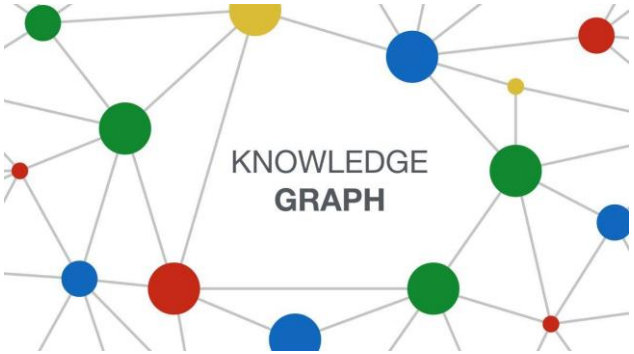
extracting triples from (semi-)structured data is relatively simpler

Knowledge Extraction



schema

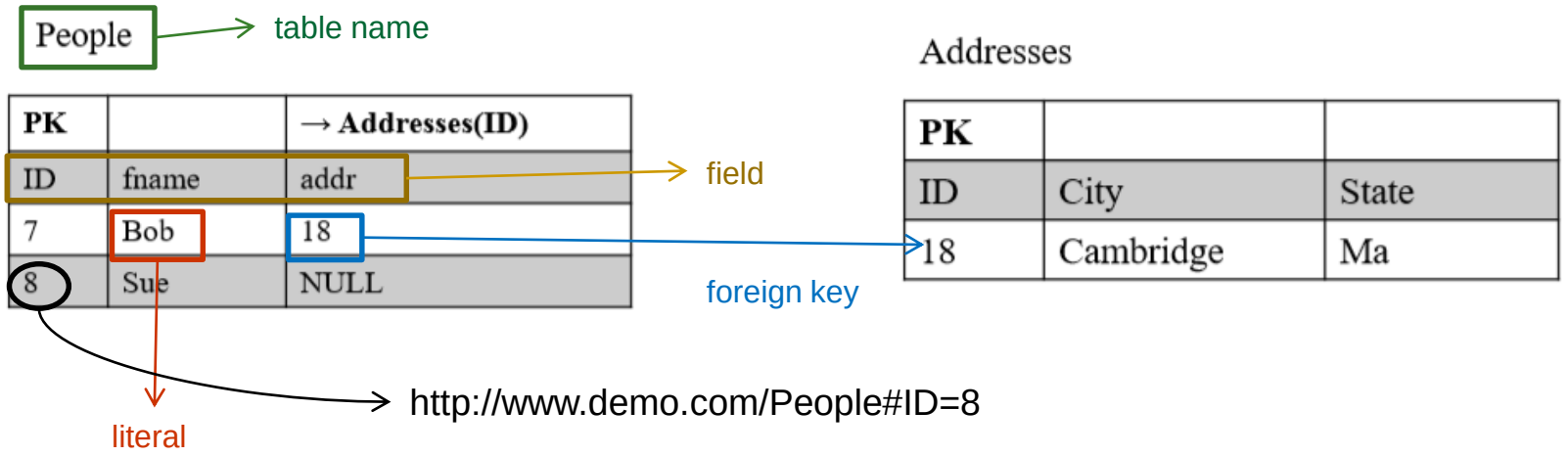
mapping



※ Direct Mapping
※ R2RML

Direct Mapping

- convert relational data into RDF, by making explicit the semantics encoded in the relational schema.
- Mapping Rules:
  - Table Name -> Class Name
  - Field -> Property
    - Field Value is literal -> Data Property
    - Field Value is foreign key -> Object Property
  - Each row is a resource -> use the primary key + table name to create URI of this resource



R2RML

- R2RML is a language for specifying mappings from relational to RDF data.

A mapping takes a logical table as input , i.e.,

- a database table
- a database view, or
- an SQL query

using rule: Triples Map

Output is a set of triples

R2RML

- Triples Maps: triples are produced by subject maps、predicate maps、object maps

People

PK		→ Addresses(ID)
ID	fname	addr
7	Bob	18
8	Sue	NULL

Addresses

PK		
ID	City	State
18	Cambridge	Ma

※ The subject IRI is generated from the primary key (ID) column by the template

http://www.demo.com/People/{ID}

※ The predicate IRI is the constant

http://www.demo.com/fname、 http://www.demo.com/addr

※ The object is literal or an subject IRI

fname -> Bob、Sue

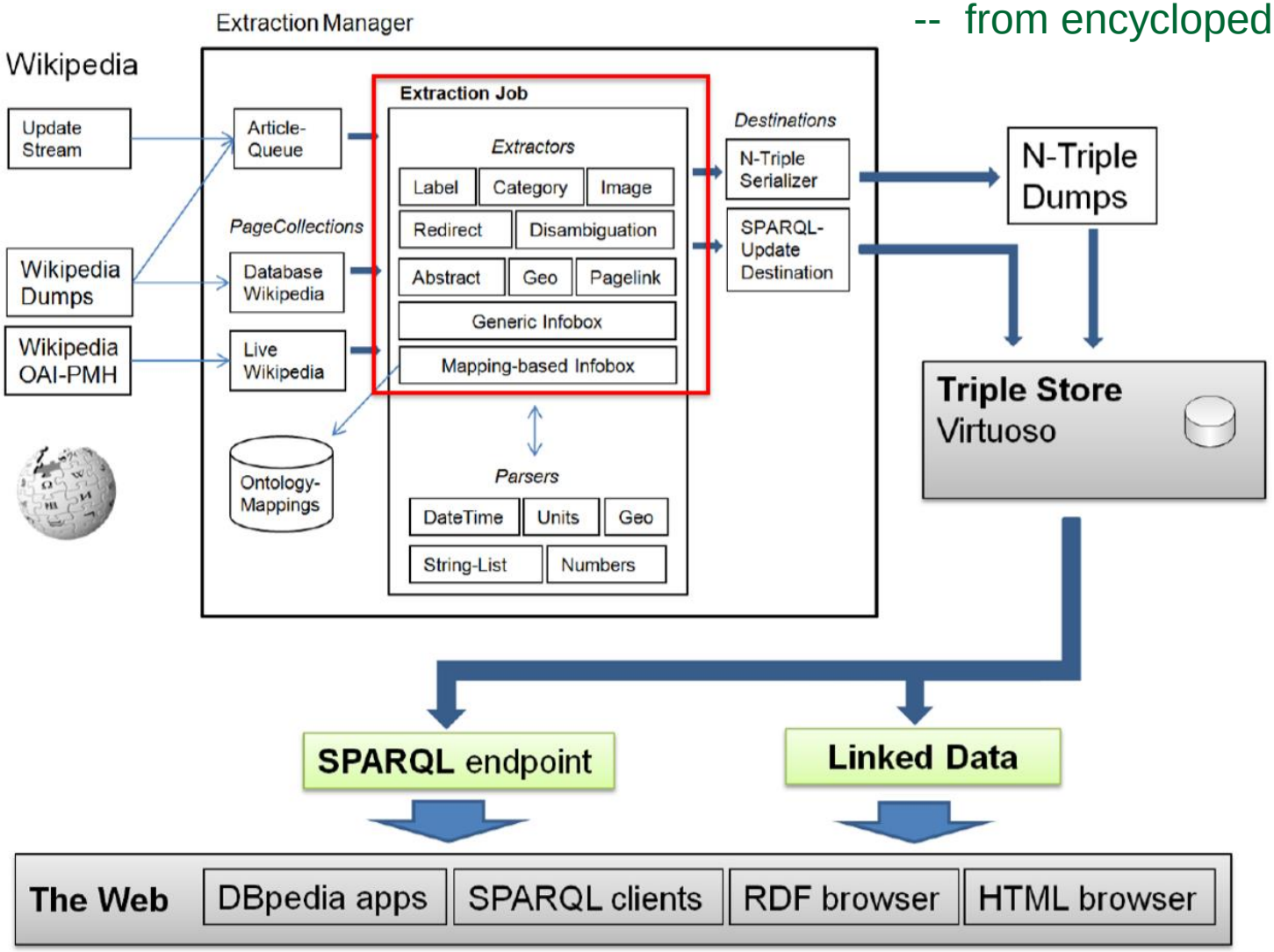
addr -> http://www.demo.com/Address/{ID}

R2RML

- Tools
  - Ontop
  - D2RQ

Semi-structured Data Extraction

-- from encyclopedias



Semi-structured Data Extraction

-- from encyclopedias

Berners-Lee in 2014

Born	Timothy John Berners-Lee 8 June 1955 (age 62)[1] London, England, UK
Other names	TimBL TBL
Education	Emanuel School
Alma mater	The Queen's College, Oxford (BA)
Occupation	Computer scientist
Spouse(s)	Rosemary Leith (m., 2014) Nancy Carlson (m., 1990; div., 2011)
Children	2
Parent(s)	Conway Berners-Lee Mary Lee Woods
Awards	Turing Award (2016) Queen Elizabeth Prize (2013) OM (2007) KBE (2004) FRS (2001)[2] FREng (2001) FRSA (2001) DFBCS (1995) See full list of honours
Website	www.w3.org/People/Berners-Lee
Scientific career	
Institutions	World Wide Web Consortium University of Oxford University of Southampton Plessey

infobox template names=instance types, **(rdf: type)**

```
{{pp-move-indef}}
{{pp-semi-vandalism|small=yes}}
{{Infobox person
 name = Sir Tim Berners-Lee
 honorific_suffix = {{postnominals|country=GBR|OM|KBE|FRS|FREng|FRSA|FBCS}}
 image = Sir Tim Berners-Lee (cropped).jpg
 image_size = 220px
 caption = Berners-Lee in 2014
 alt = blond man in his fifties wearing a blue suit, light blue shirt, and blue
 birth_name = Timothy John Berners-Lee
 birth_date = {{birth date and age|1955|6|8|df=y}}<ref name="whoswho"/>
 birth_place = [[London]], England, UK
 education = [[Emanuel School]]
 alma_mater = [[The Queen's College, Oxford]] (BA)
 awards = {{Plainlist|
* [[Turing Award]] (2016)
* [[Queen Elizabeth Prize]] (2013)
* [[Member of the Order of Merit|OM]] (2007)
* [[Knight Commander of the Order of the British Empire|KBE]] (2004)
* [[Fellow of the Royal Society|FRS]] (2001)<ref name=frs/>
* [[Fellow of the Royal Academy of Engineering|FREng]] (2001)
* [[Fellow of the Royal Society of Arts|FRSA]] (2001)
* [[Distinguished Fellow of the British Computer Society|DFBCS]] (1995)
* [[Awards and honours presented to Tim Berners-Lee|See full list of honours]]}}
 spouse = {{Plainlist|
```

infobox properties=instance properties,

(dbpedia:property/[propertyName])

Semi-structured Data Extraction

-- from encyclopedias

- Generic Infobox Extraction
  - Do not handle synonym property
    - birthDate & dateOfBirth are different

- Mapping-based Infobox Extraction
  - Predefined ontology, properties; and judge the extracted properties
    - predefined property birthDate
      - birthDate
      - dateOfBirth

Semi-structured Data Extraction

-- from normal web pages

- manual operation
  - analyse the web page structure and code manually, then write an expression that fits the page
    - XPath expression
    - CSS selector expression
- wrapper
  - A software program that can extract data from web pages and restore them to structured data
- Automatic extraction

Semi-structured Data Extraction

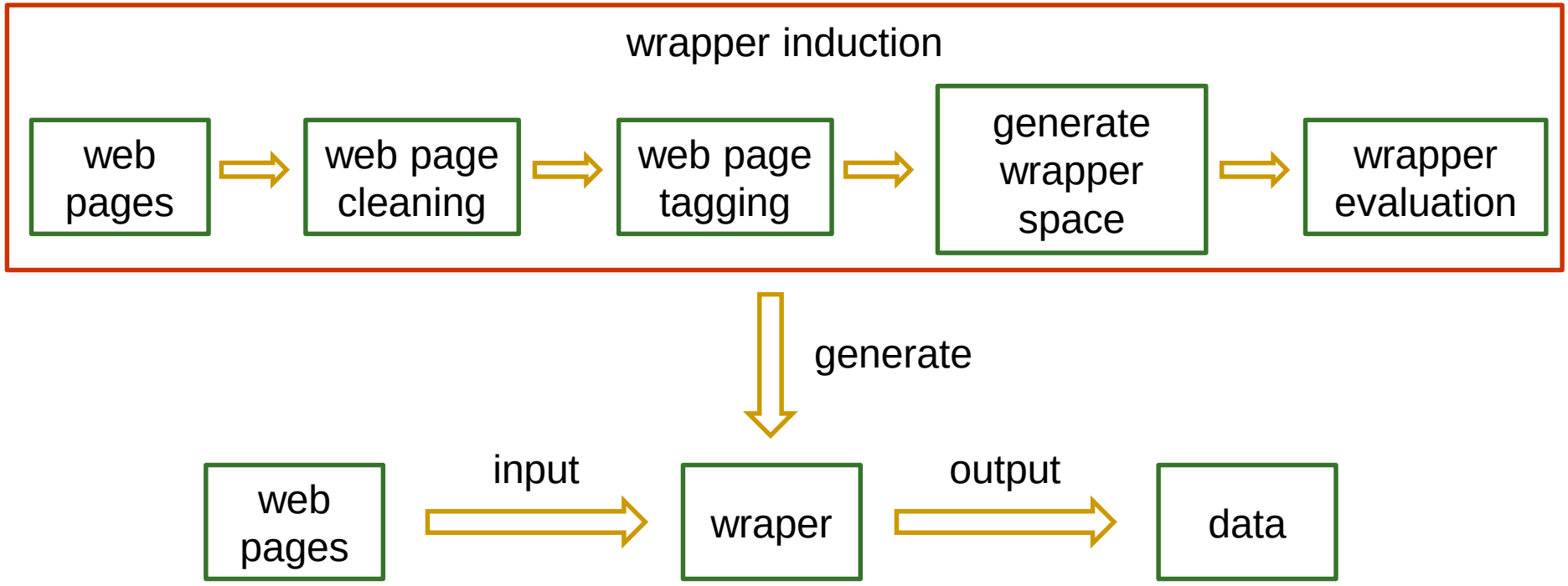
-- from normal web pages

- manual operation
  - XPath (XML Path Language)
    - It is a language used to locate a part of an XML document, so we can get the location of elements in a web page
  - CSS selector
    - Use CSS elements to locate data elements in web pages and get information about the data elements

Semi-structured Data Extraction

-- from normal web pages

- Wrapper
  - A software program that can extract data from web pages and restore them to structured data



Semi-structured Data Extraction

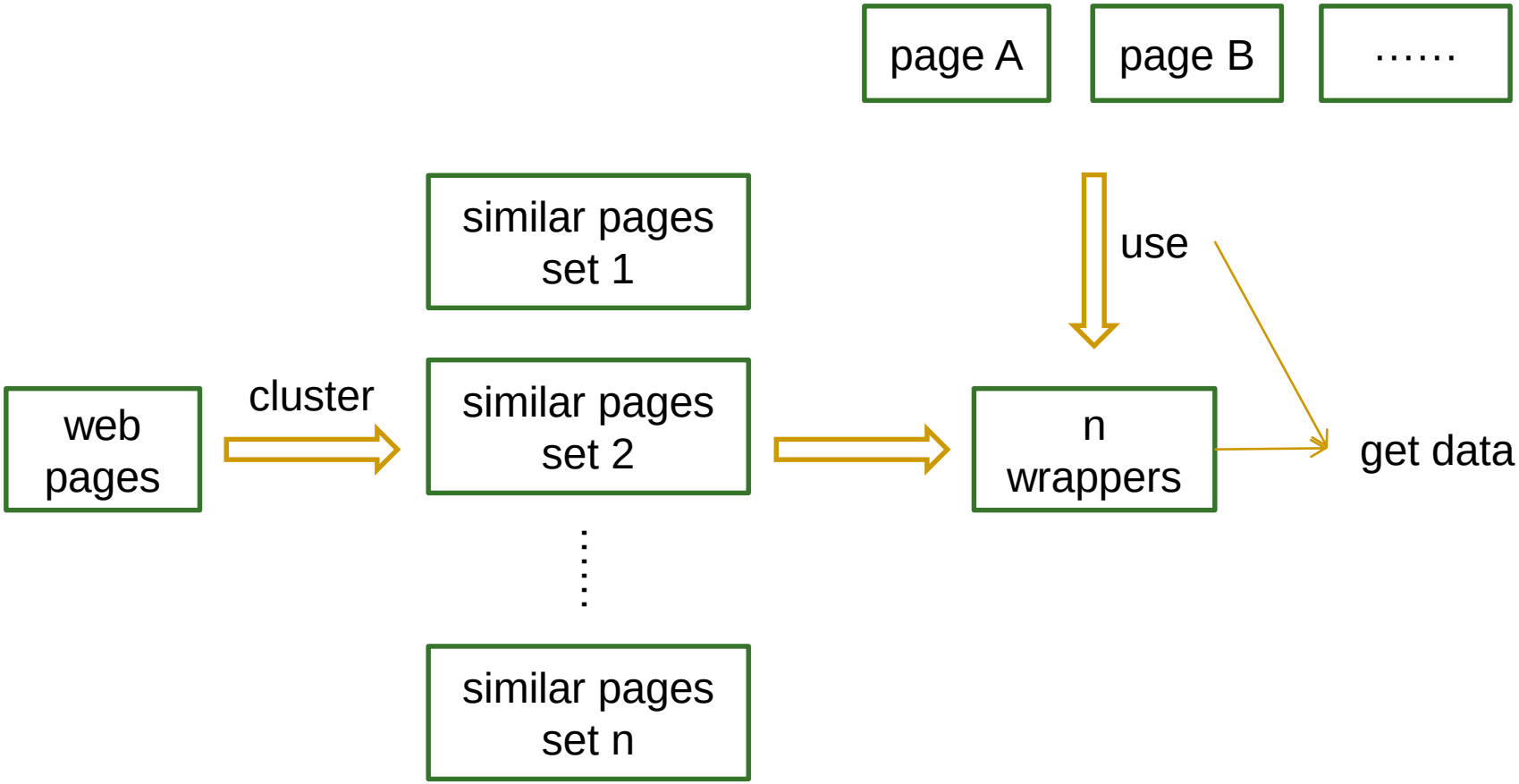
-- from normal web pages

- Wrapper induction
  - wrapper induction is based on supervised learning, which learns data extraction rules from labeled training sample sets, and is used to extract target data from other web pages with the same label or template.

Semi-structured Data Extraction

-- from normal web pages

- Automatic extraction



Semi-structured Data Extraction

-- from normal web pages

- Automatic extraction
  - wrapper training
    - Do not need any manual labeling
    - Clustering a group of web pages to divide similar web pages into several groups. Each group will get its own wrapper
  - wrapper application
    - Comparing the Web pages that need to be extracted with those that used to generate wrappers
    - Determine the classification of this new page, and then use the corresponding wrapper