FINDING, ASSESSING, AND INTEGRATING STATISTICAL SOURCES FOR DATA MINING

Karin Becker[1], Xiaojie Tan[2],

Shiva Jahangiri[3], Craig Knoblock[3]

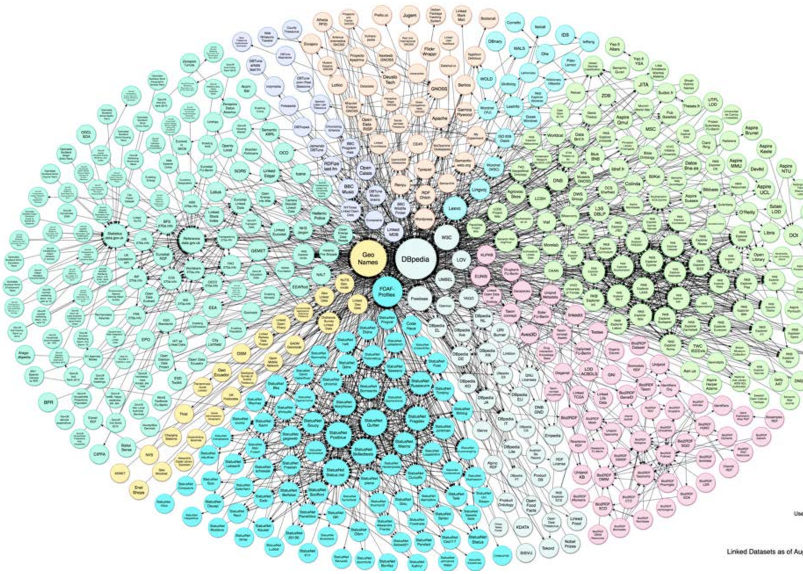
[1] Instituto de Informática – Universidade Federal do Rio Grande do Sul - Brazil

[2]School of Information Management – University of Nanjing - China

[3] Information Sciences Institute, University of Southern California - USA

Introduction

- The number of government statistical datasets in the LOD is increasing (300% in the last census)
- Enriched statistical data can be used to build analysis models
- Growing opportunity to use the LOD as a primary data source for knowledge discovery
- Cube vocabulary is a de *facto standard* for representing multi-dimensional data (indicators)



Introduction

- Existing tools support querying and visualization cubes
  - Assumes the cube datasets are given
  - Integration is mostly left to the user
- Our goal:
  - Mechanisms for finding and integrating cube datasets that contain compatible indicators
  - Data selection and preprocessing steps of knowledge discovery process

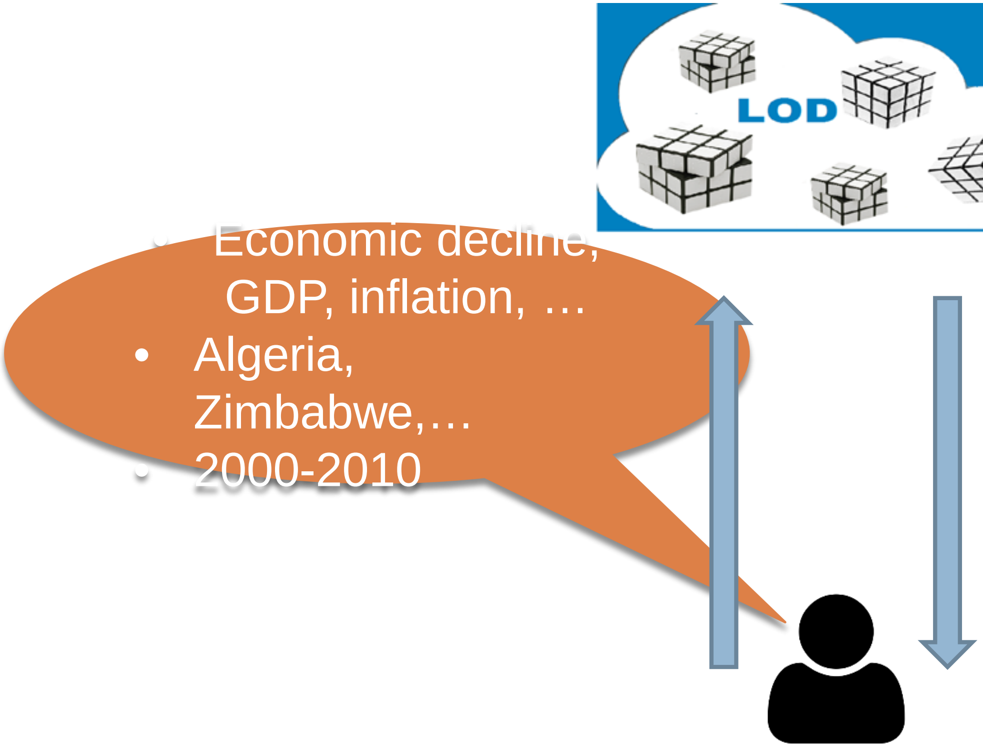
Scenario: Peacebuilding

- Predict Fragile States Indicator “Economic Decline”
  - influenced by inflation, GDP, unemployment , etc.
- Data is available as open data in different portals

Finding
Understanding
Proprietary APIs and Formats
Integrating

- Laborious, time consuming, error-prone

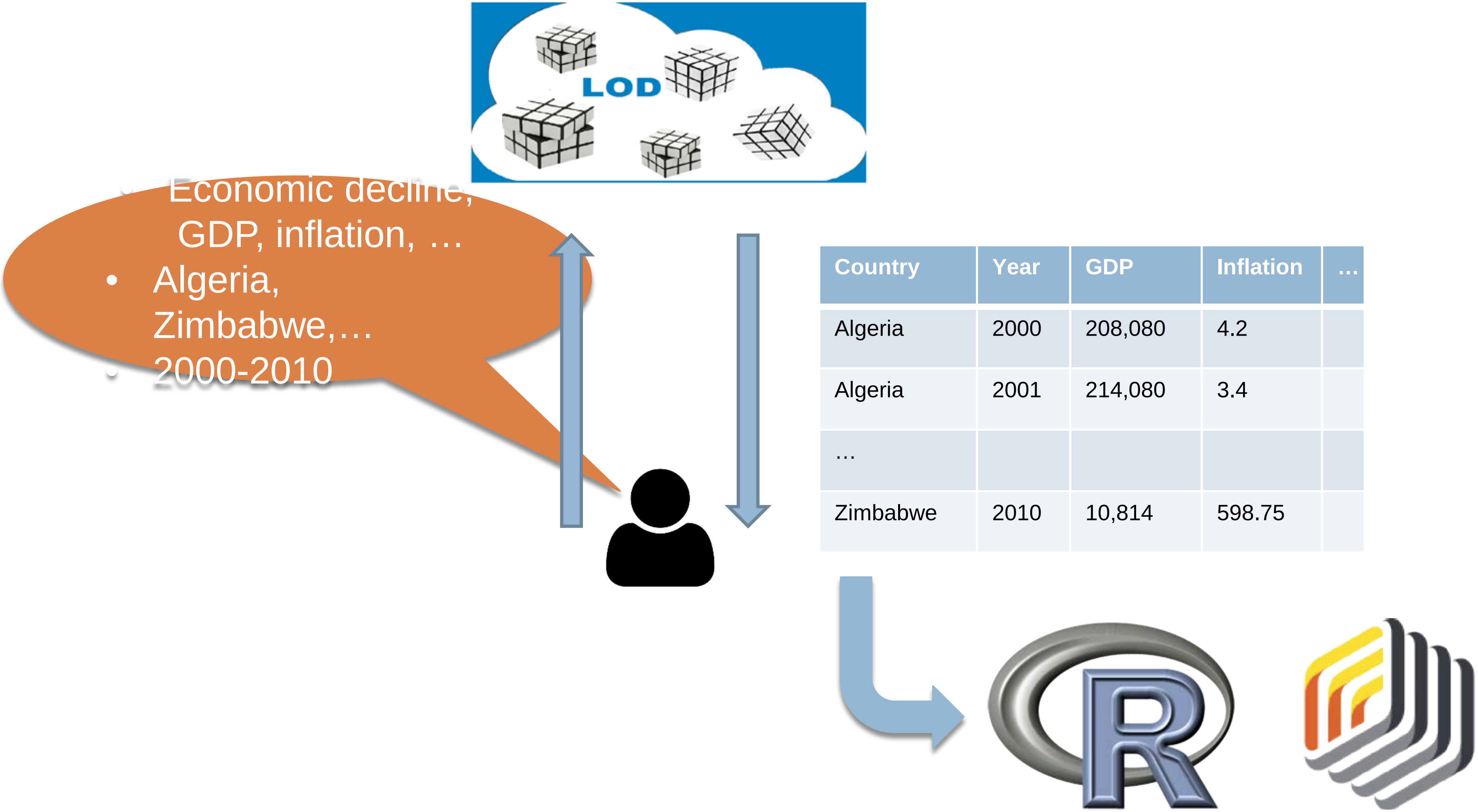
Proposed Approach



- Economic decline, GDP, inflation, …
- Algeria, Zimbabwe,…
- 2000-2010

Country	Year	GDP	Inflation	…
Algeria	2000	208,080	4.2	
Algeria	2001	214,080	3.4	
…				
Zimbabwe	2010	10,814	598.75	

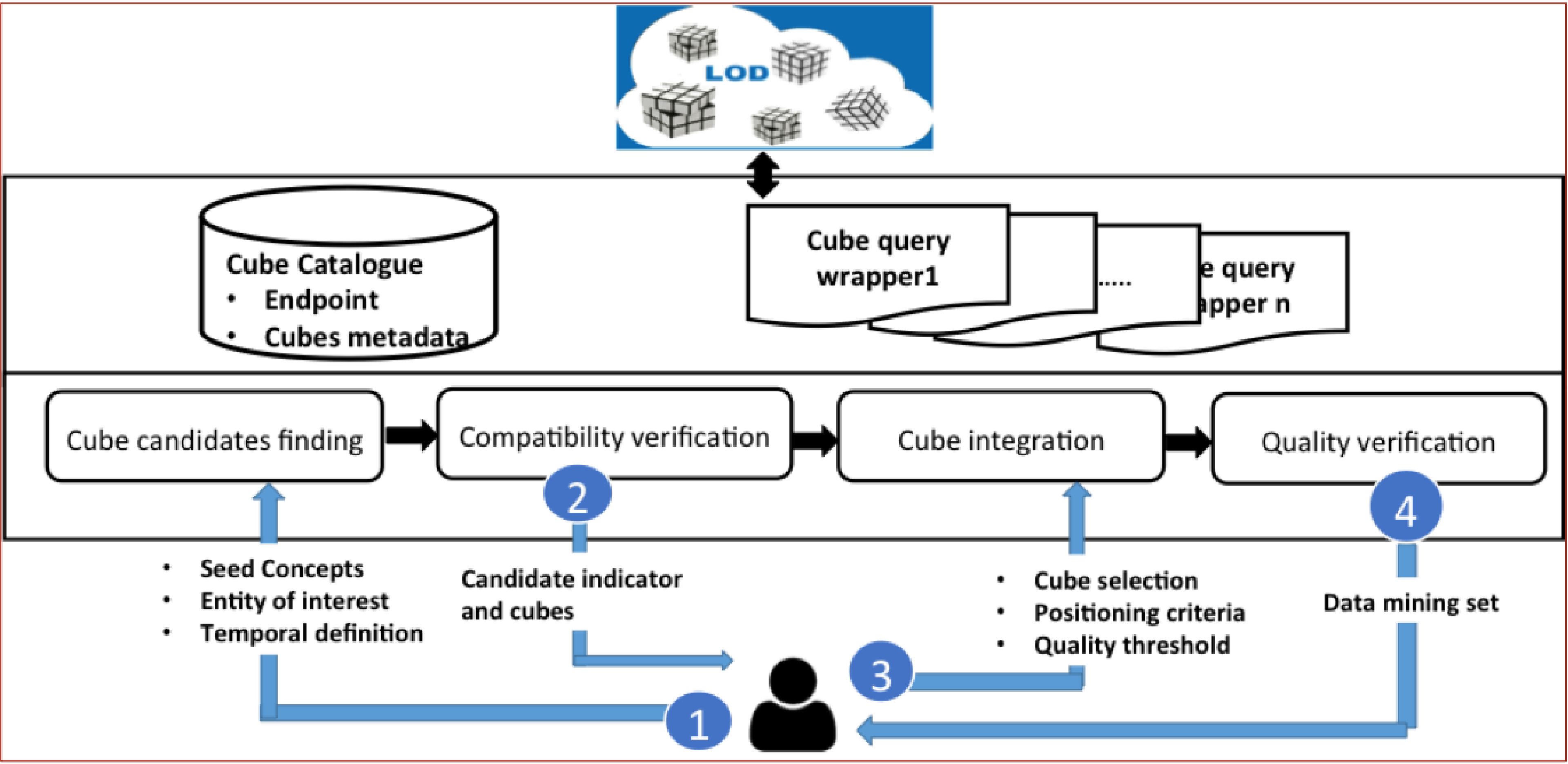
Proposed Approach



- Economic decline, GDP, inflation, …
- Algeria, Zimbabwe,…
- 2000-2010

Country	Year	GDP	Inflation	…
Algeria	2000	208,080	4.2	
Algeria	2001	214,080	3.4	
…				
Zimbabwe	2010	10,814	598.75	

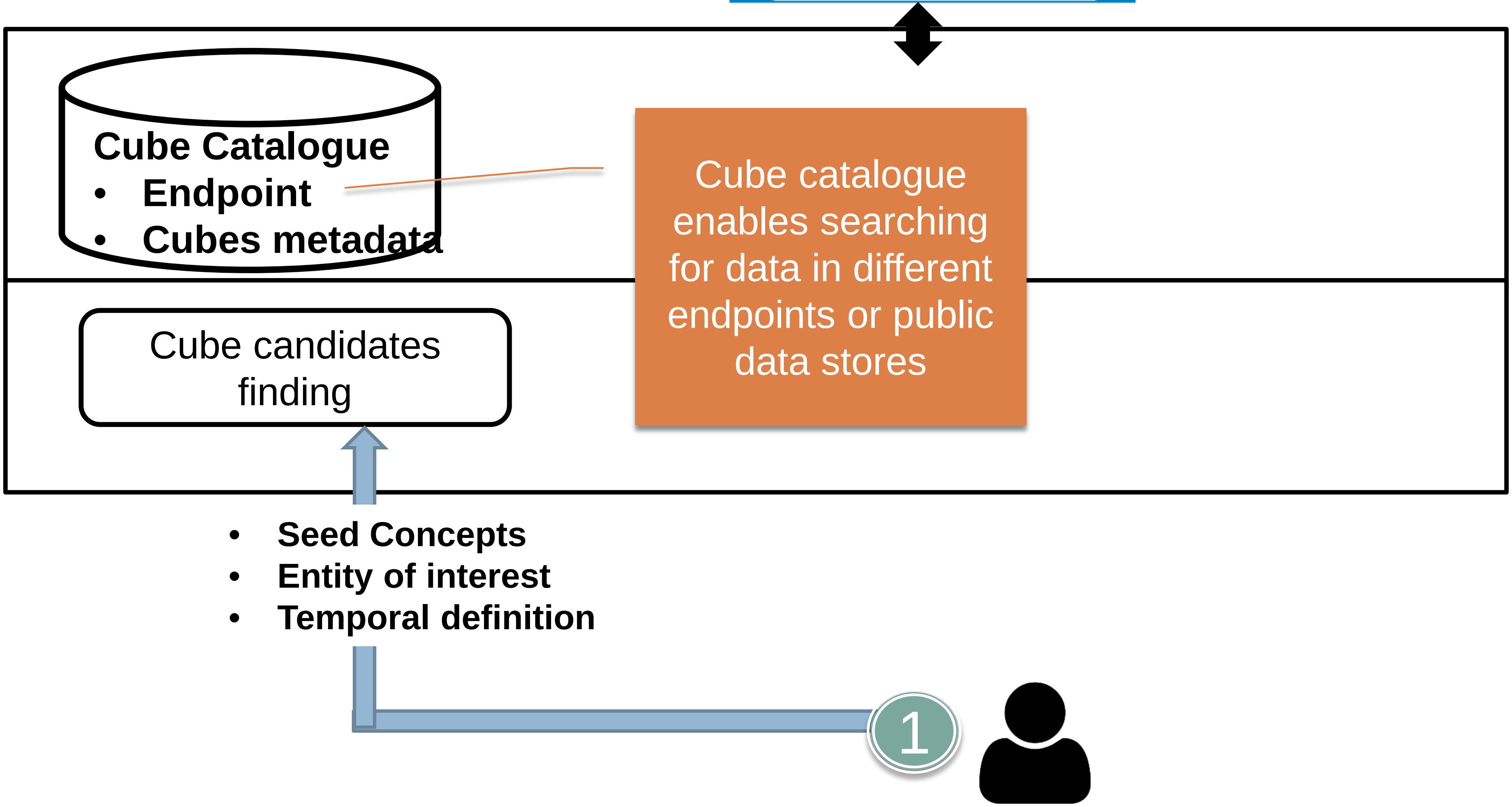
Proposed Approach



Cube Vocabulary in Practice

- Standard concepts, but different modeling styles
- Data Definition Structure (DSD) should provide the explicit definition of measures and dimensions in cube datasets
  - Often not the case
- Semantics associated at different levels, using different properties
  - Cube constructs are not exploited to their full potential
  - Many cubes are straightforward conversions of SDMX representations

Where to find?



Cube Catalogue

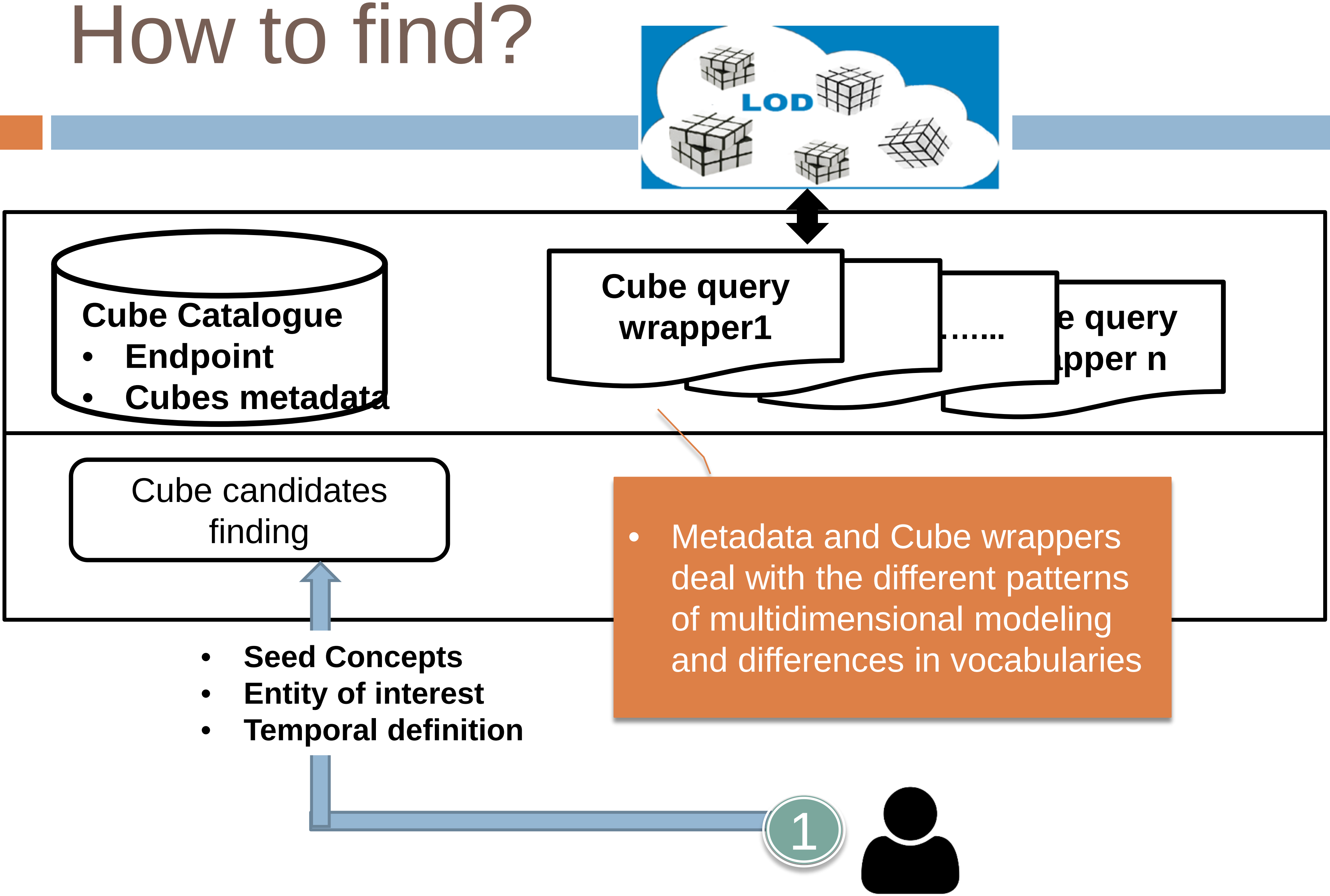
- Endpoint
- Cubes metadata

Cube catalogue enables searching for data in different endpoints or public data stores

Cube candidates finding

- Seed Concepts
- Entity of interest
- Temporal definition

How to find?



LOD

Cube Catalogue
- Endpoint
- Cubes metadata

Cube query wrapper1

…….

e query
pper n

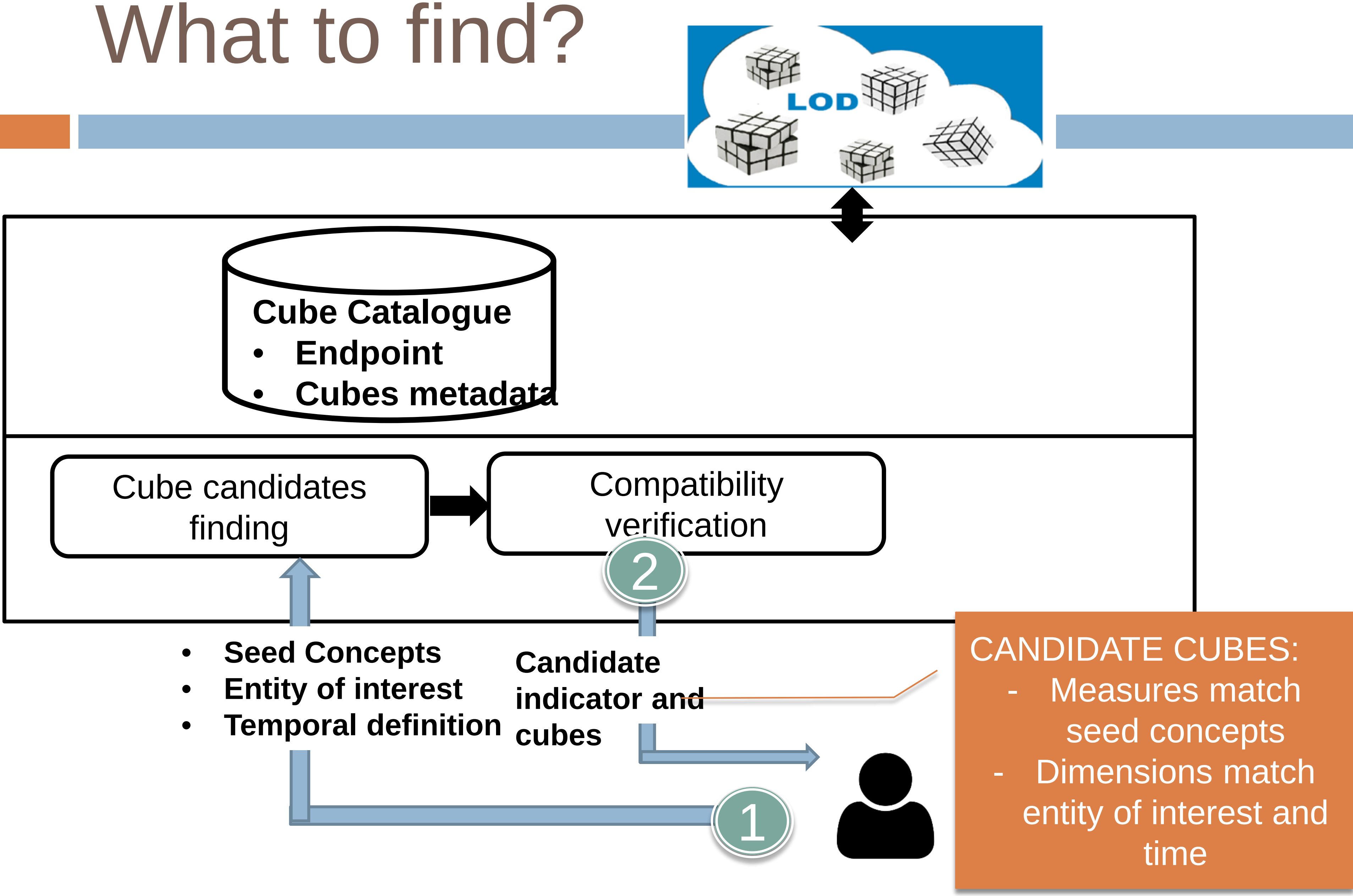
Cube candidates finding

- Seed Concepts
- Entity of interest
- Temporal definition

1

- Metadata and Cube wrappers deal with the different patterns of multidimensional modeling and differences in vocabularies

What to find?

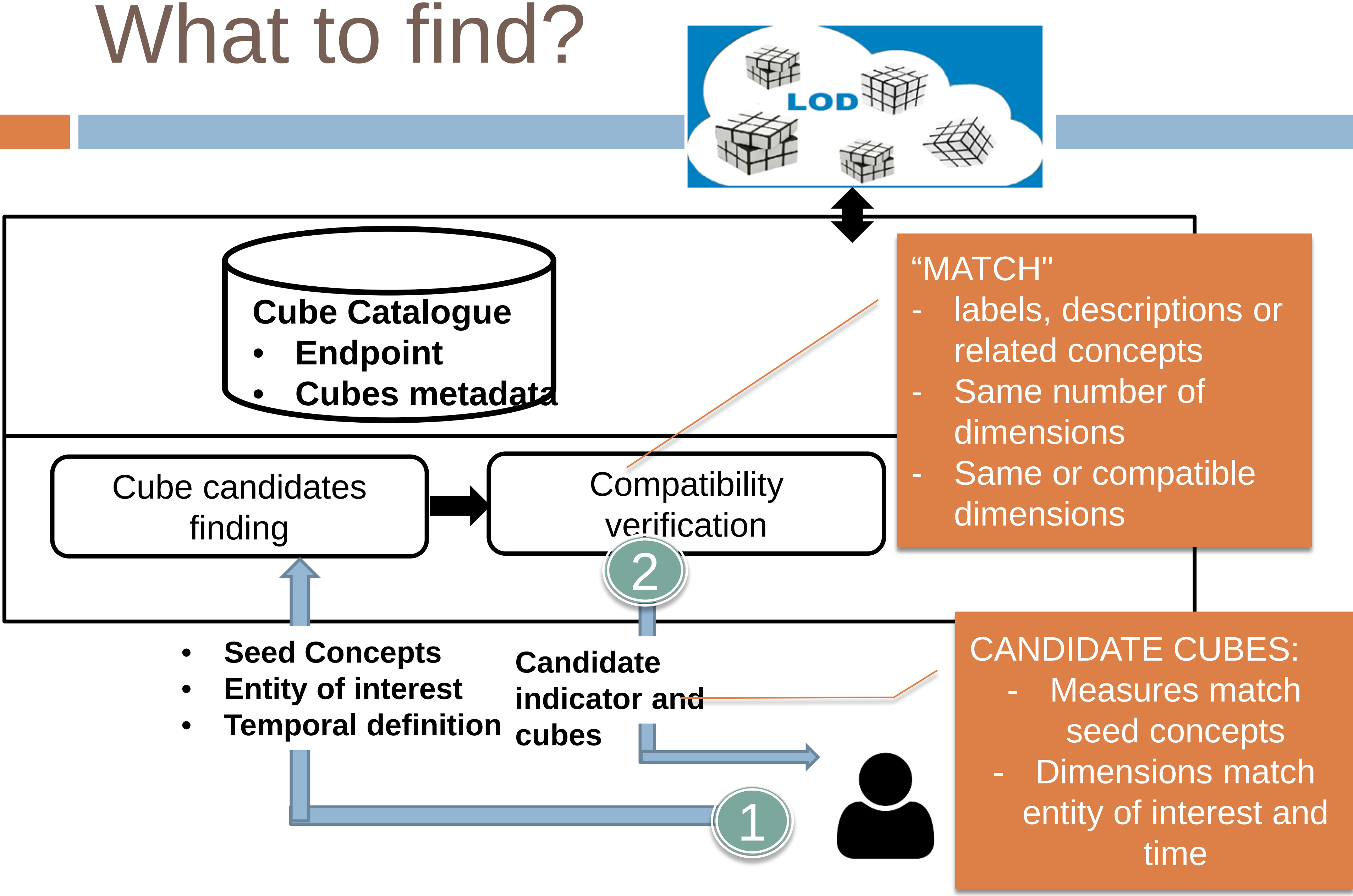


- Seed Concepts
- Entity of interest
- Temporal definition

Candidate indicator and cubes

CANDIDATE CUBES:
- Measures match seed concepts
- Dimensions match entity of interest and time

What to find?



Cube Catalogue
- **Endpoint**
- **Cubes metadata**

Cube candidates finding

Compatibility verification

- **Seed Concepts**
- **Entity of interest**
- **Temporal definition**

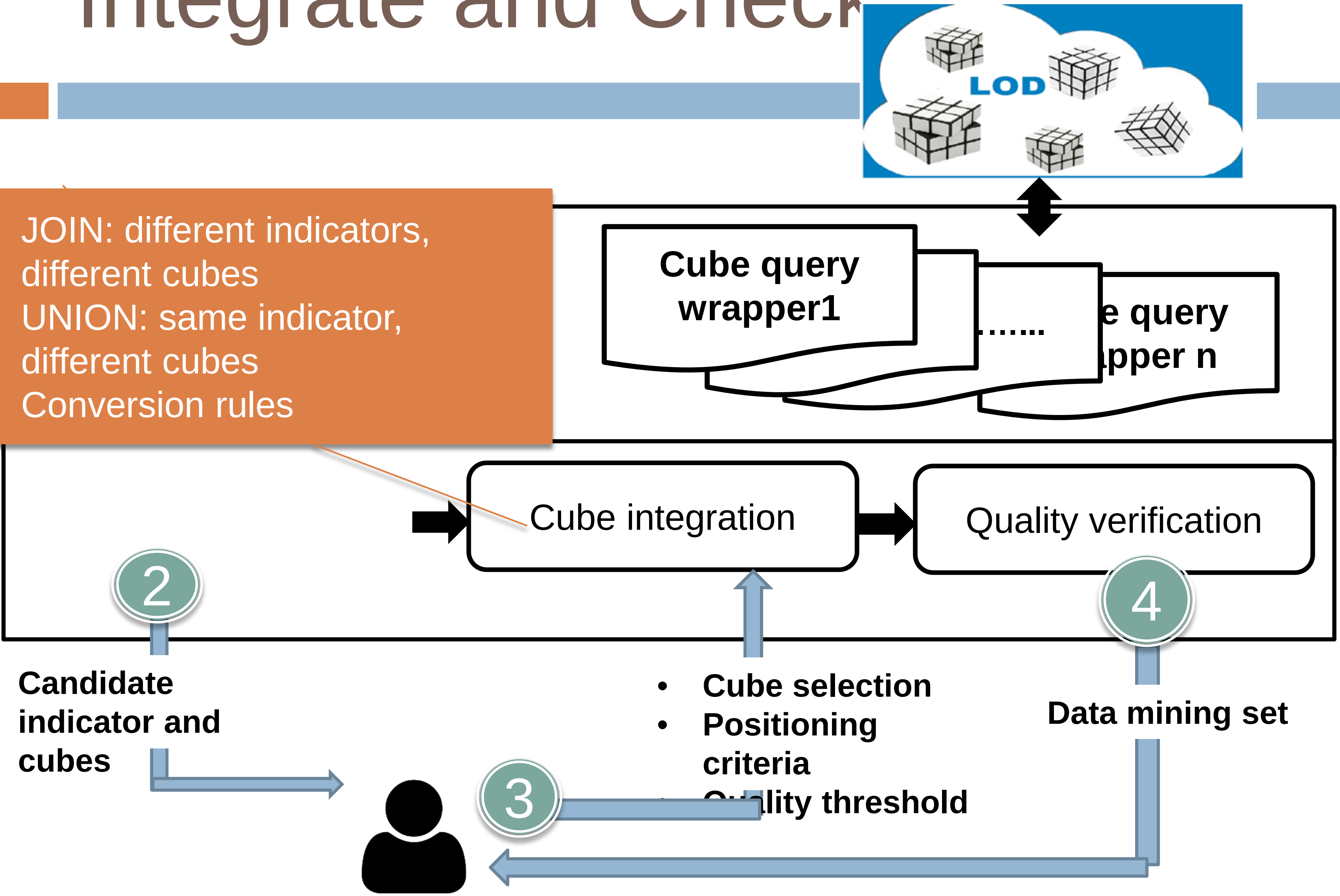
Candidate indicator and cubes

"MATCH"
- labels, descriptions or related concepts
- Same number of dimensions
- Same or compatible dimensions

CANDIDATE CUBES:
- Measures match seed concepts
- Dimensions match entity of interest and time

Integrate and Check

- JOIN: different indicators, different cubes
- UNION: same indicator, different cubes
- Conversion rules



Cube query wrapper1

…….

e query pper n

Cube integration

Quality verification

2

Candidate indicator and cubes

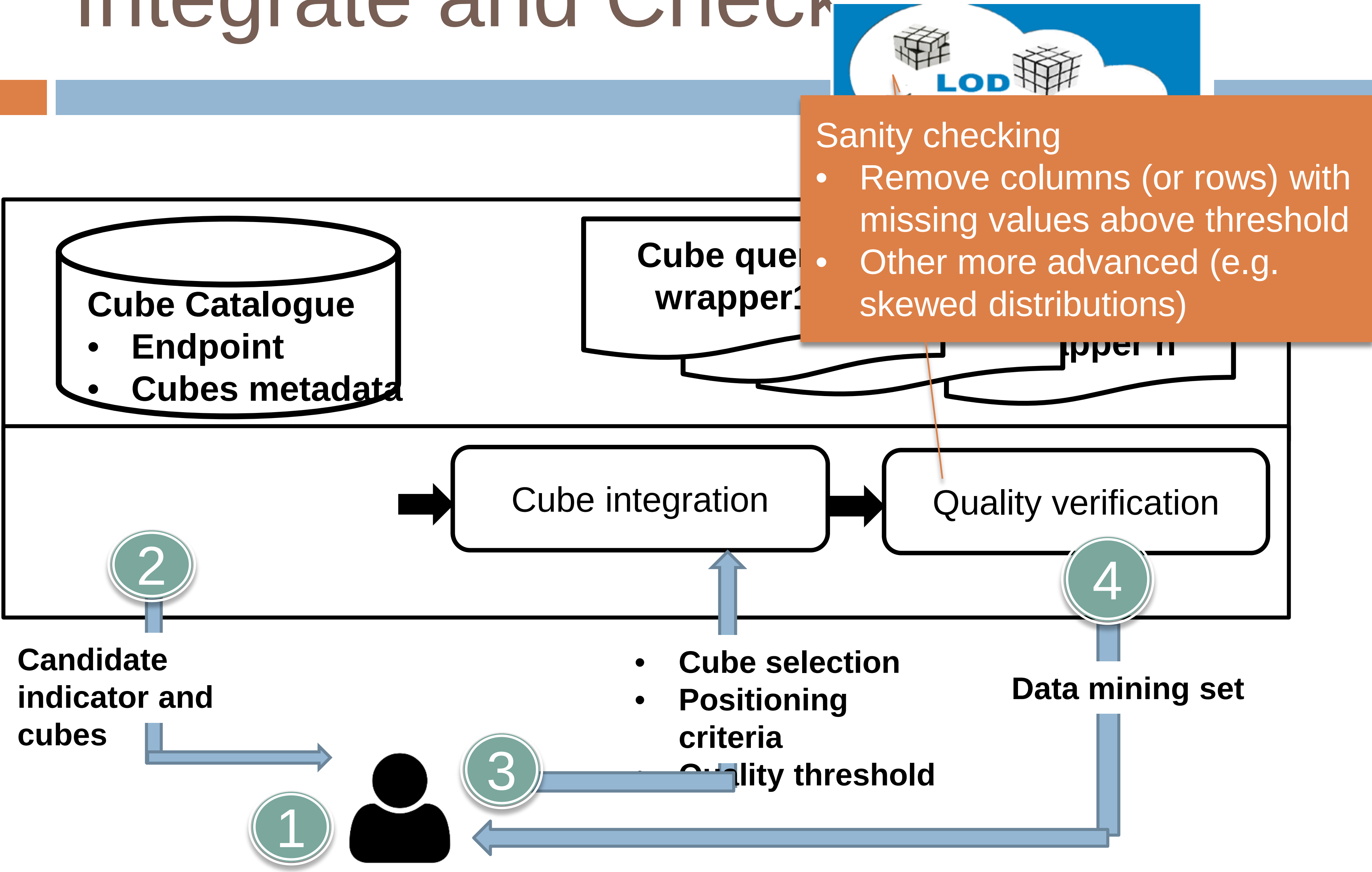
3

- Cube selection
- Positioning criteria
- Quality threshold

4

Data mining set

Integrate and Check



Sanity checking
- Remove columns (or rows) with missing values above threshold
- Other more advanced (e.g. skewed distributions)

Cube Catalogue
- Endpoint
- Cubes metadata

Cube que
wrapper1

pper n

Cube integration

Quality verification

Candidate indicator and cubes

- Cube selection
- Positioning criteria
- Quality threshold

Data mining set

Related Work

- Cube Platforms: LOD2 Statistical Workbench, OpenCube, OLAP4LD
  - Support the creation, validation, querying, and visualization of cube datasets
- LOD extension for RapidMiner
  - Set of operators for integrating data with LOD data
  - Cube retrieval operator
- Janpuangton and Shell (2015) – identification of relevant data in the LOD from seed concepts
  - Does not deal with multidimensional data
- Our work complements these works with functionality for Cube discovery and integration

Conclusions and Future Work

- Approach to
  - finding and integrating cube datasets from seed concepts
  - Assessing their capability
  - Integrating them to generate a mining dataset
- Next steps
  - Automatic generation of query wrappers
  - Exploiting the data for predicting indicators