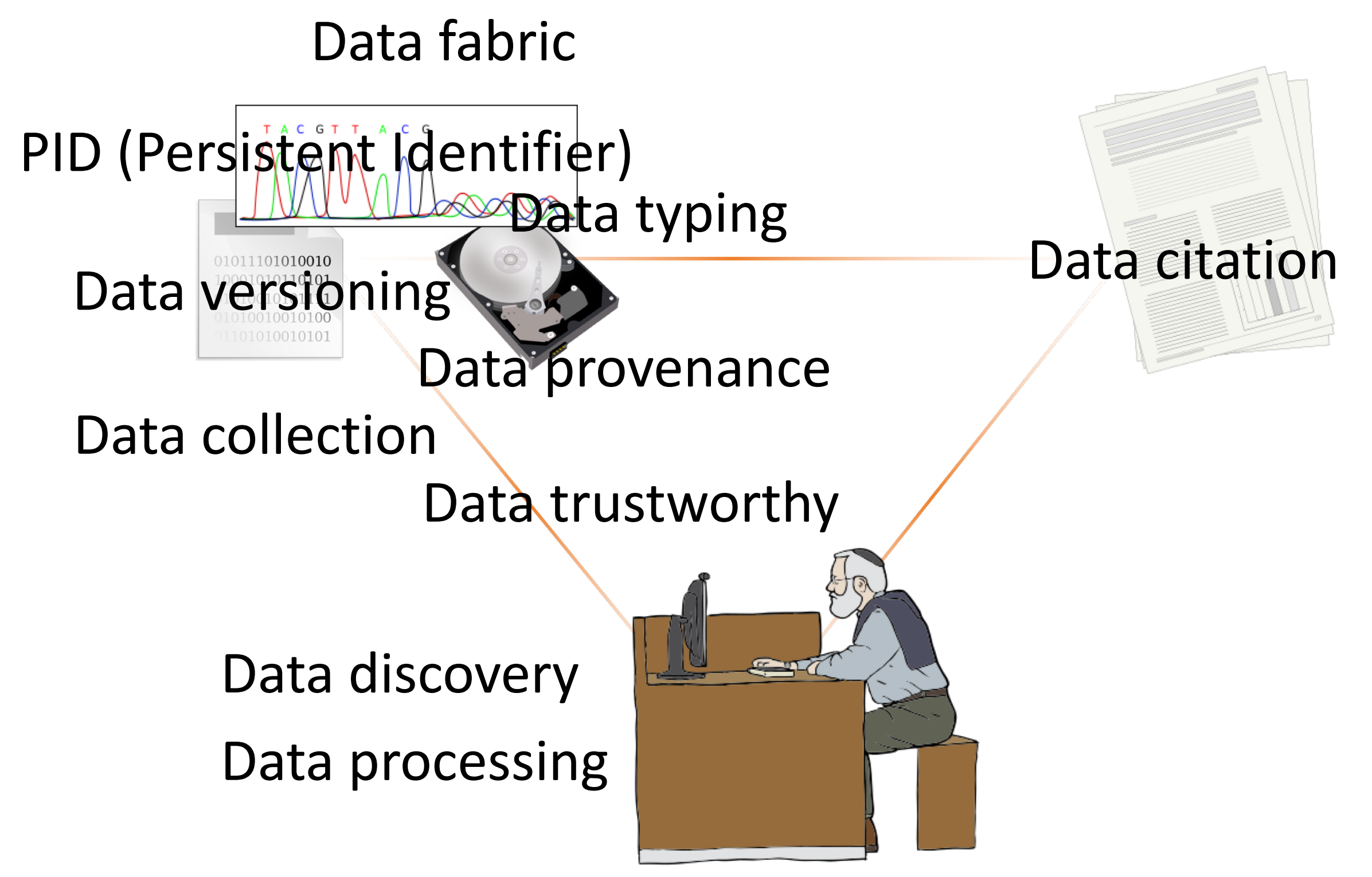
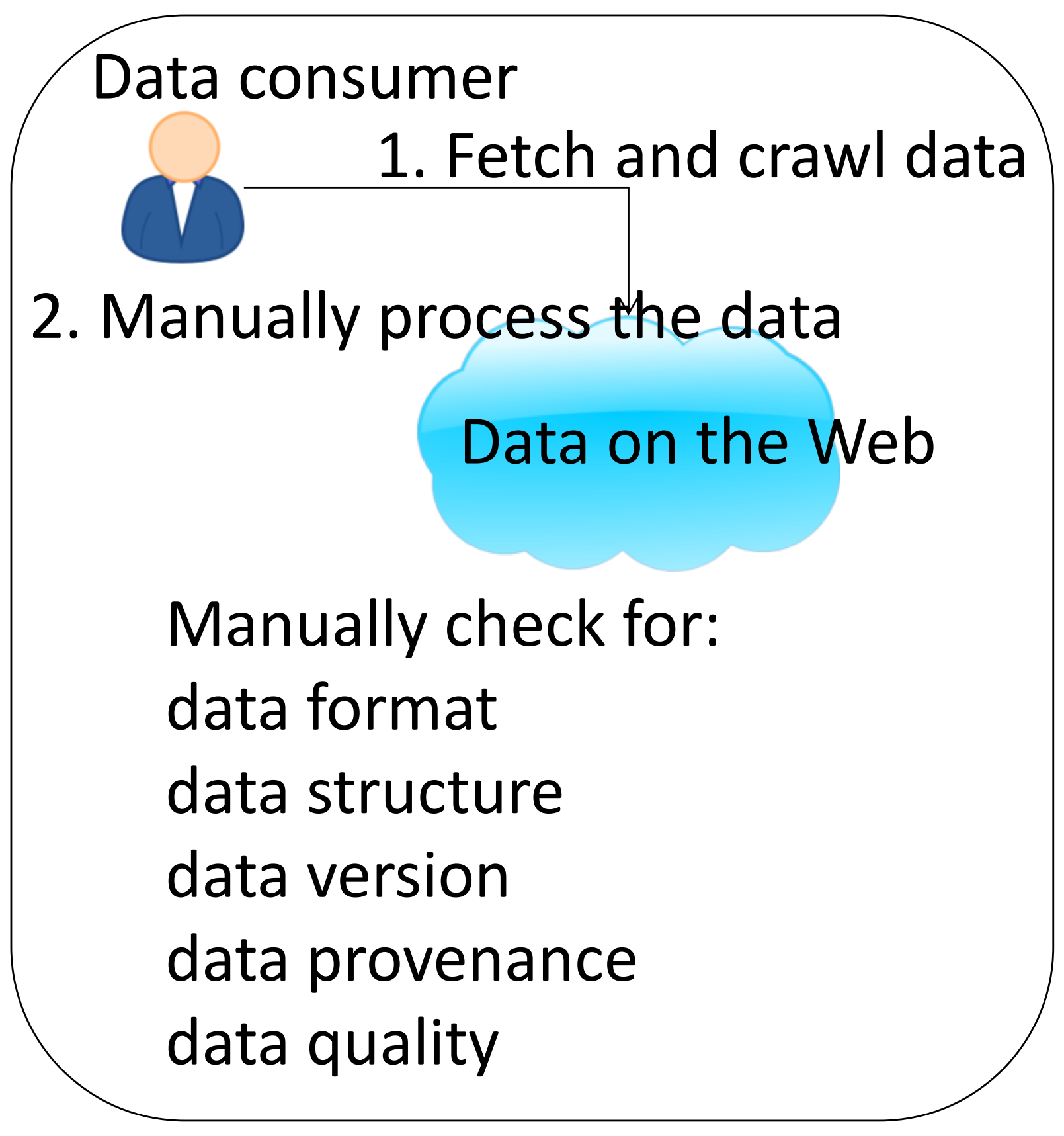
Making data typing efforts or automatically detecting data types for automatic data processing?

Research data consumer on the cloud where very large collections of distributed scientific data are provided



Two major processes in the scientific data use
Data discovery
Data processing

Manual data processing



Manually check for:
data format
data structure
data version
data provenance
data quality

Mostly common, but …

The procedure, which may be peculiar to each discipline, is a process of craftsmanship and too much time consuming task.

The data consumer needs to understand the semantics of data structure in domain dependent schemes and choose ordinarily a community standard of tools on a specific computational environment to process the data.

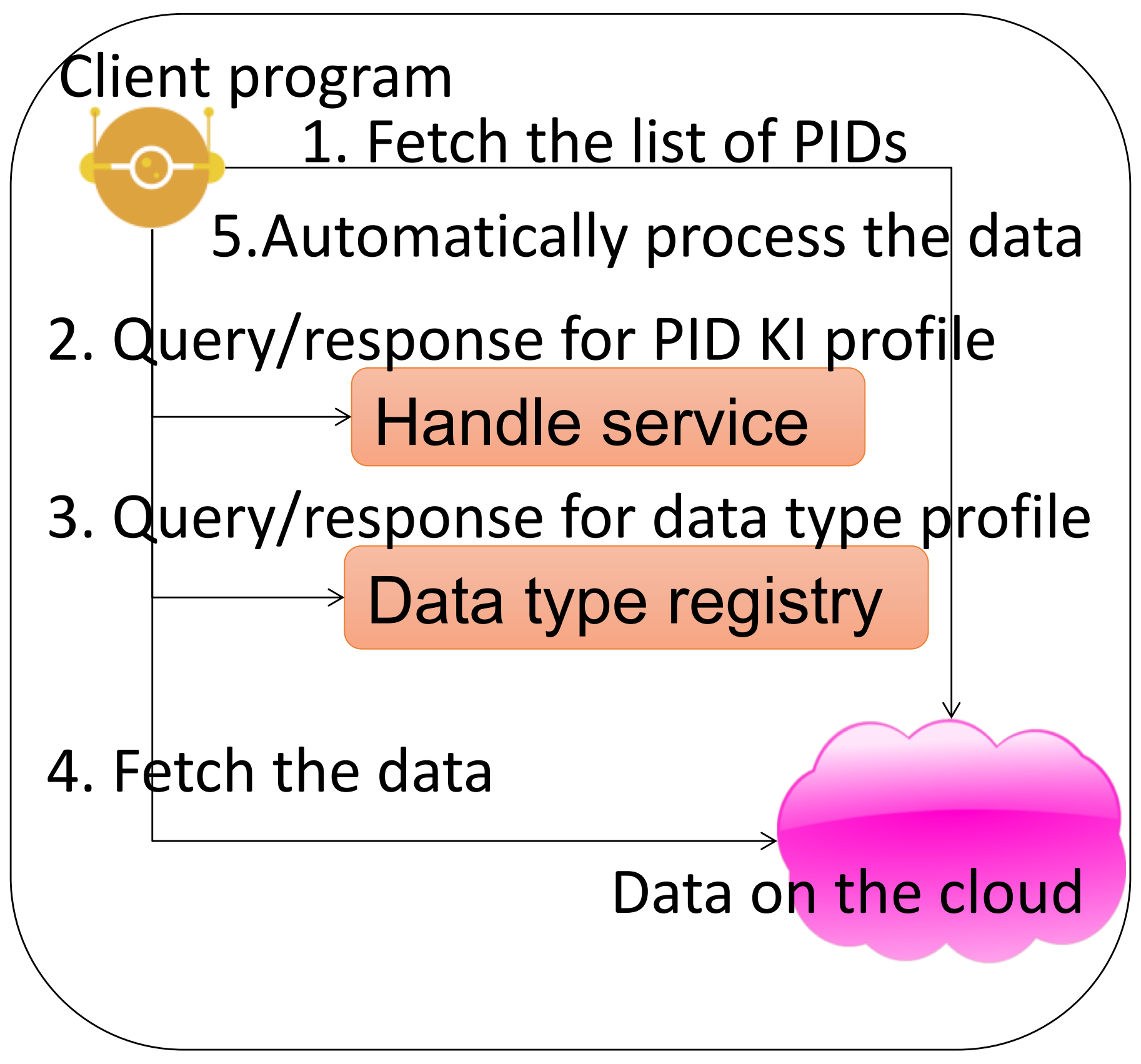
It seems to be difficult for outsiders of the expertise to do the same things.

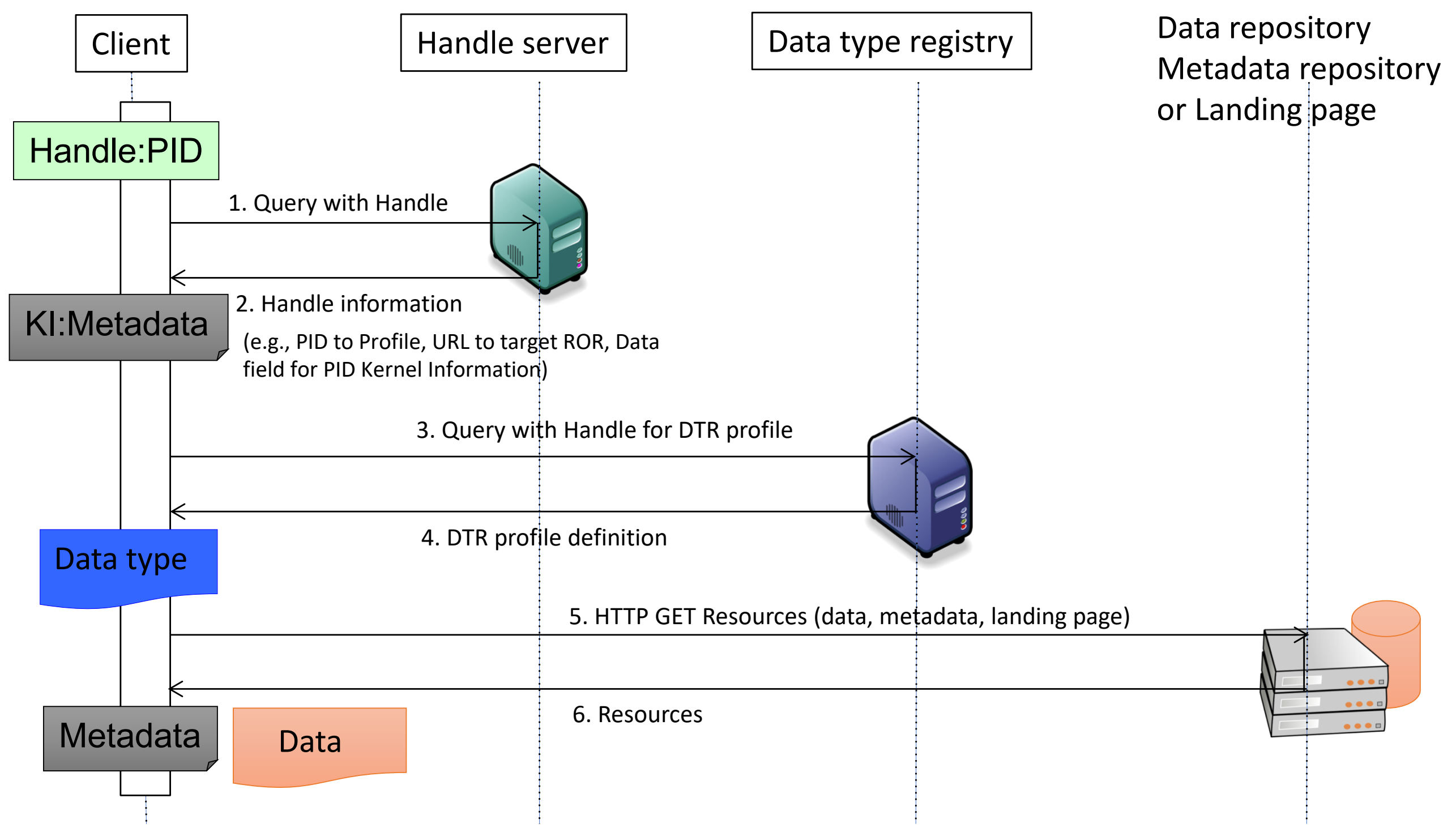
Data processing paradigm shift: from manual to automatic

For the future, we aim at an automatic way

Data consumer scenario

Automatic data processing


Client program
1. Fetch the list of PIDs
5.Automatically process the data
2. Query/response for PID KI profile
Handle service
3. Query/response for data type profile
Data type registry
4. Fetch the data
Data on the cloud


Client
Handle server
Data type registry
Data repository
Metadata repository
or Landing page
Handle:PID
1. Query with Handle
KI:Metadata
2. Handle information
(e.g., PID to Profile, URL to target ROR, Data field for PID Kernel Information)
3. Query with Handle for DTR profile
4. DTR profile definition
Data type
5. HTTP GET Resources (data, metadata, landing page)
6. Resources
Metadata
Data

If the data is not typed, the data consumer still have to process the data in a manual way or use a data type detection tool although current detection technology is immature.

Data provider scenario

In parallel, we need data typing

Step 1

❑ Data providers build their data in a community standard

- ❖ The data is packed in a commonly used format, i.e. XML, JSON, netCDF, CSV as well as application dependent such as Microsoft EXCEL format.
- ❖ Some data are shipped with a document describing data meaning, data types, and data format.

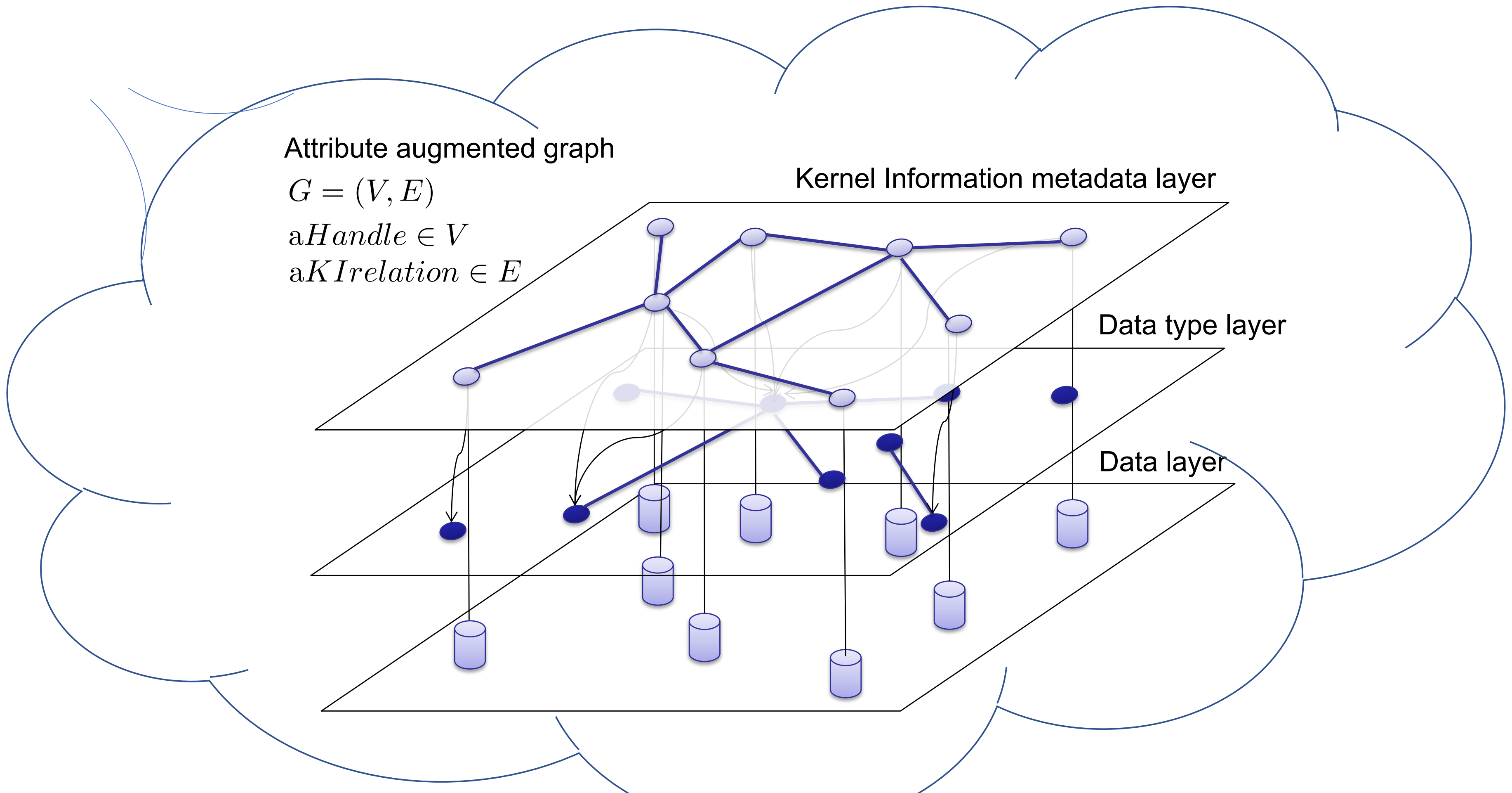
Step 2

❑ Data providers use more complicated data format to assert data types

- ❖ A set of Handle server of DOI objects with Kernel Information profile and Data Type Registry is a recommended candidate for a variety of domain community to assert their data types in addition to their data sources in a community standard format.
- ❖ On the other hand, linked data community uses RDF/XML, JSON-LD and other linked data formats, or a kind of mixture format of data type and value.
  - Common vocabularies are provided in a public server, e.g. schema.org.

Well-organized data cloud structure

Kernel information connects data and data types



This poster is presented for

11th Research Data Alliance Plenary Meeting
Berlin, Germany
21-23 March, 2018

Acknowledgements: The authors are thankful for RDA KI WG discussions.

Authors

Kei Kurakawa, National Institute of Informatics
(orcid.org/0000-0002-7031-1846, e-mail: kurakawa@nii.ac.jp)
Takayuki Sekiya, The University of Tokyo
Yasumasa Baba, The Institute of Statistical Mathematics