Data Science in Medicine

Lecture 8: Introduction to Graph Data and Ontologies

Dr Areti Manataki

Usher Institute
The University of Edinburgh

Data integration

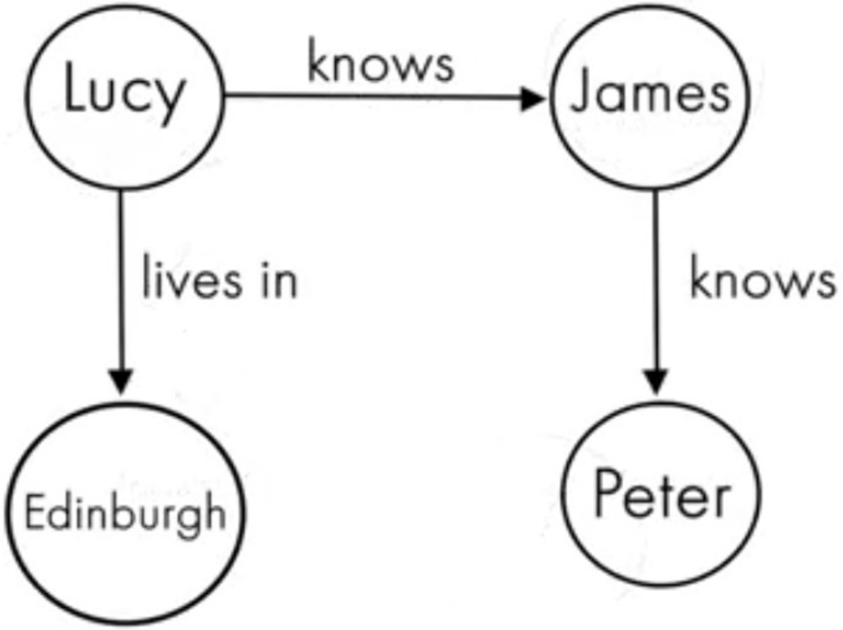
- The analysis of genomic, imaging or other types of data allows us to investigate different facets of human health.
- But in order to gain a comprehensive understanding of human health, we need to integrate such data.

Challenges to data integration

1. Biomedical and healthcare datasets sit in silos.
2. Linking entities between different datasets is not a trivial task.
   - In Scotland, we use CHI Numbers to uniquely identify patients.
   - But how about sharing data between different countries?
3. Ambiguity around the meaning of different terms.

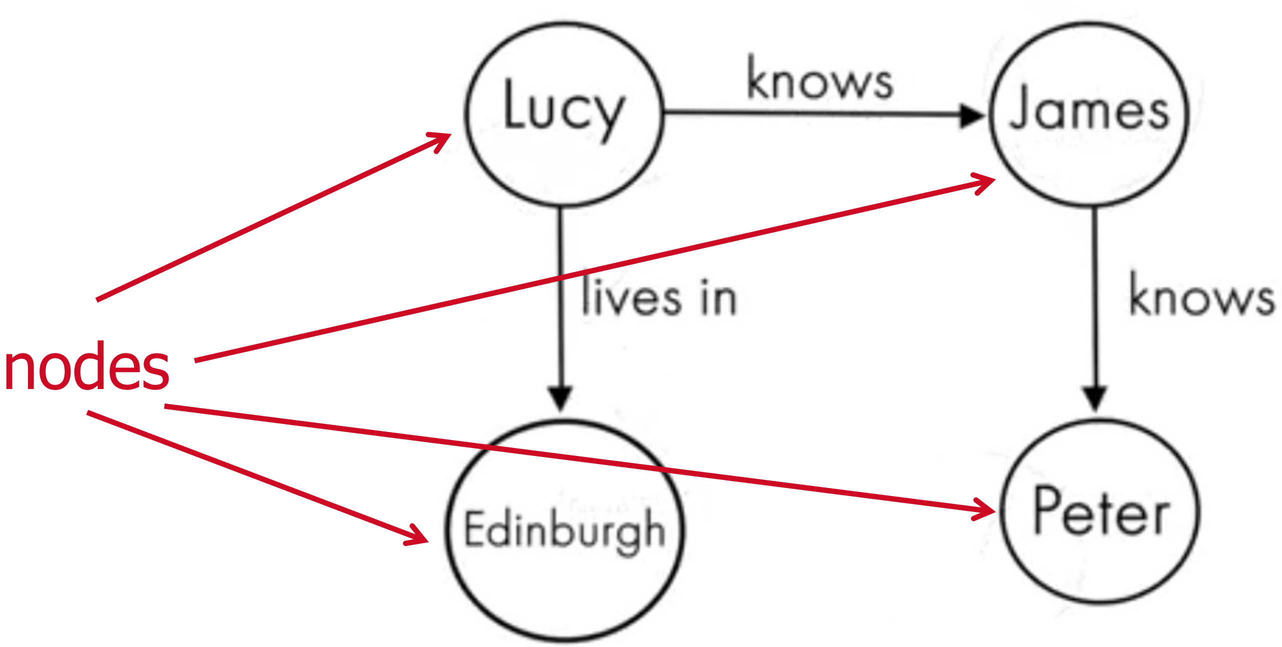
Graph databases

- Graph Databases use graph structures with nodes, edges and properties to represent and store data.



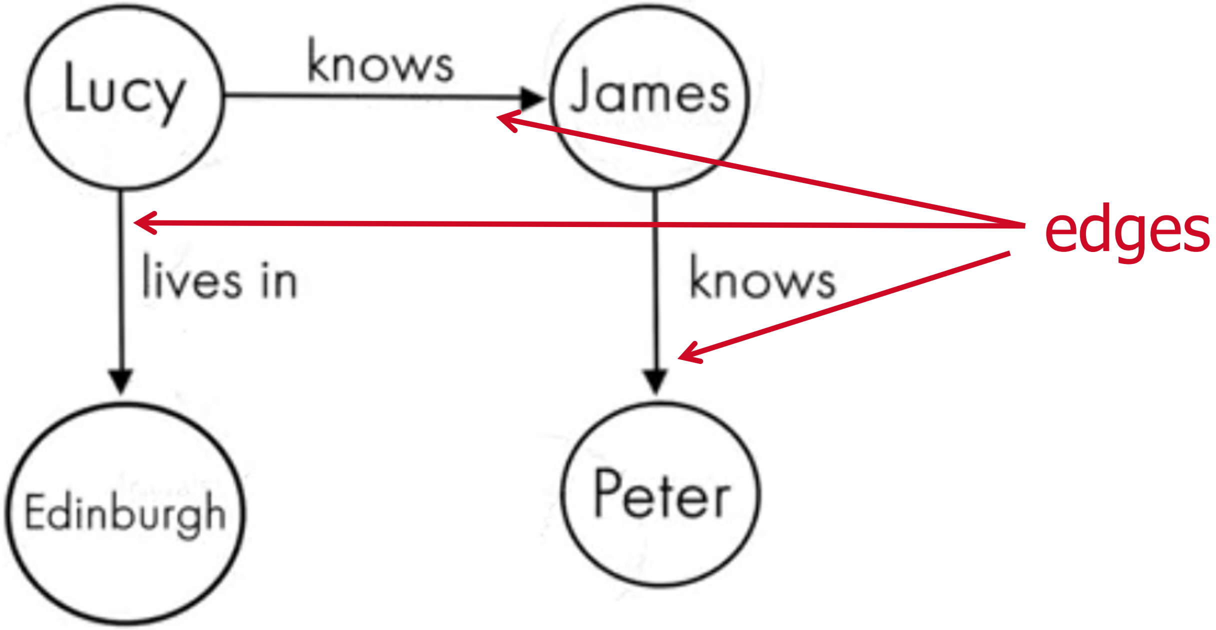
Graph databases

- Graph Databases use graph structures with nodes, edges and properties to represent and store data.



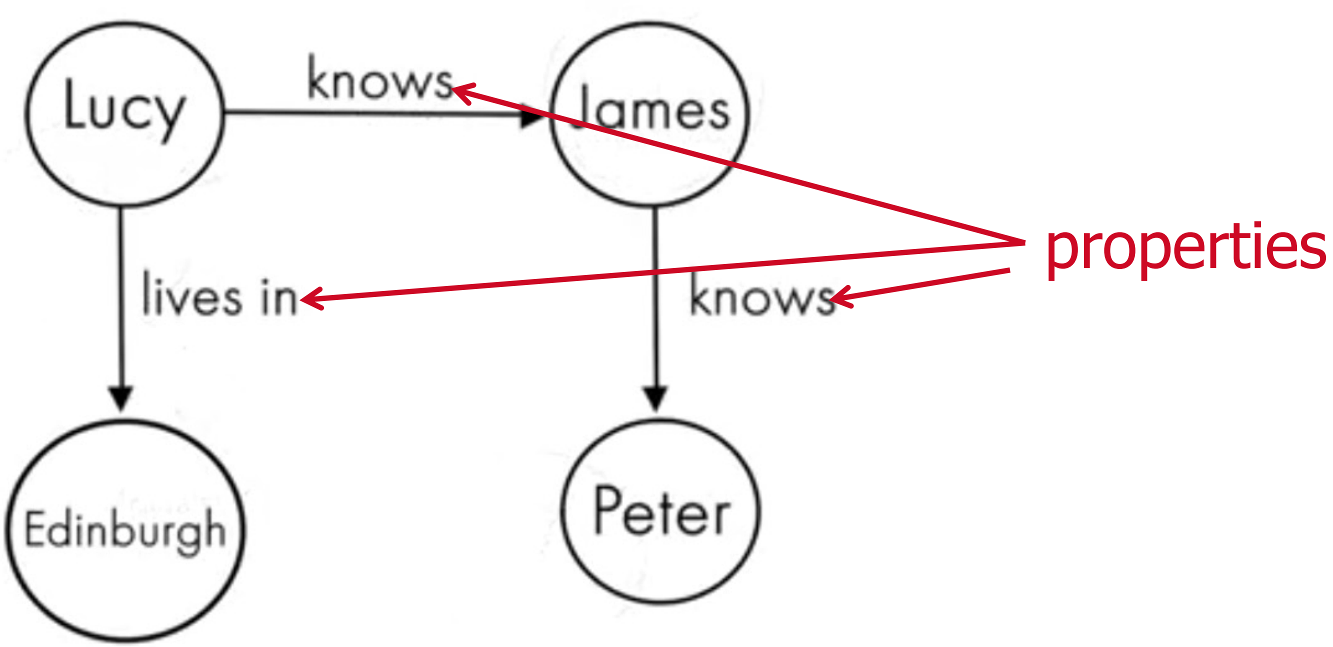
Graph databases

- Graph Databases use graph structures with nodes, edges and properties to represent and store data.



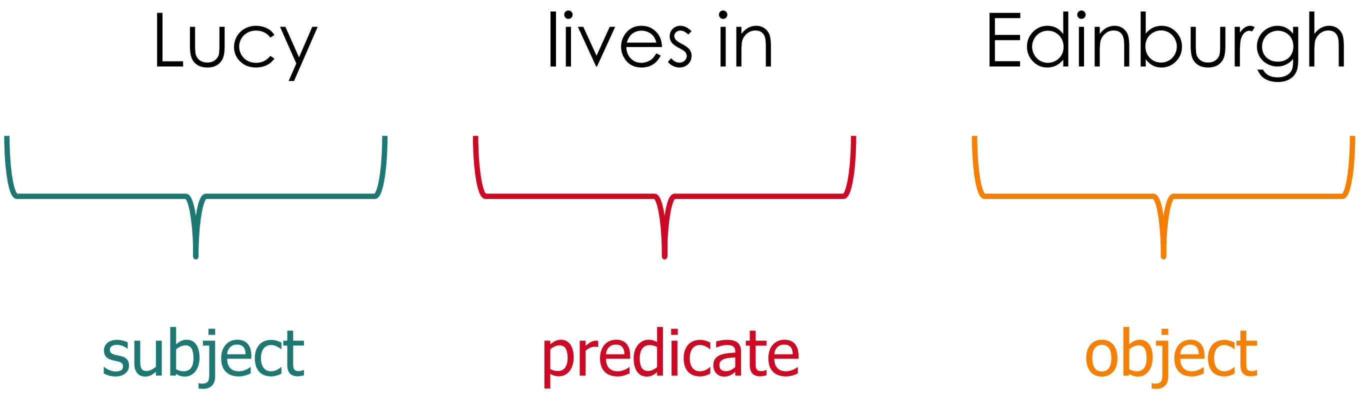
Graph databases

- Graph Databases use graph structures with nodes, edges and properties to represent and store data.



The RDF graph data model

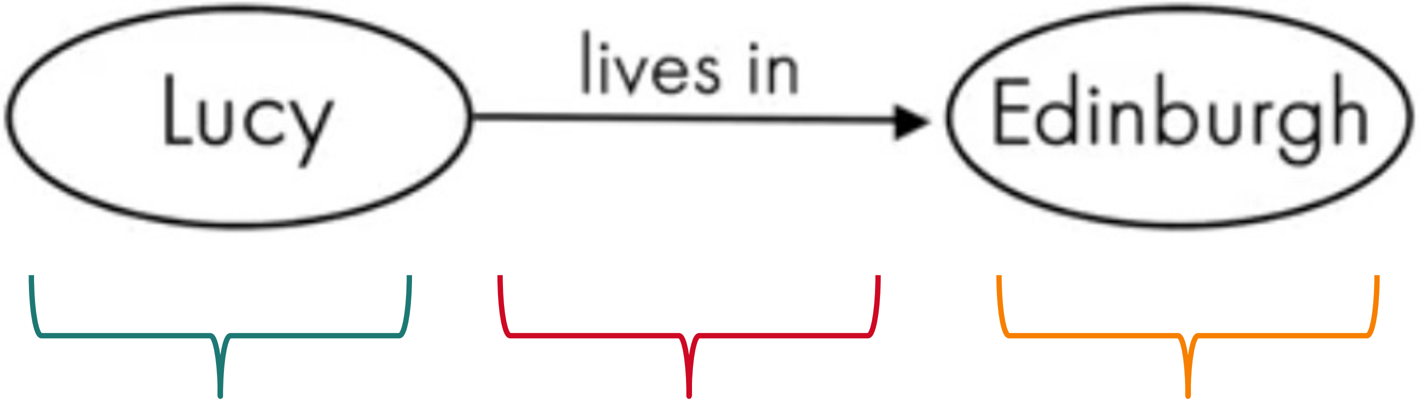
- Data is represented in the form of triples, i.e. statements consisting of a subject, a predicate and an object.



Lucy lives in Edinburgh

subject predicate object

RDF triple visualisation

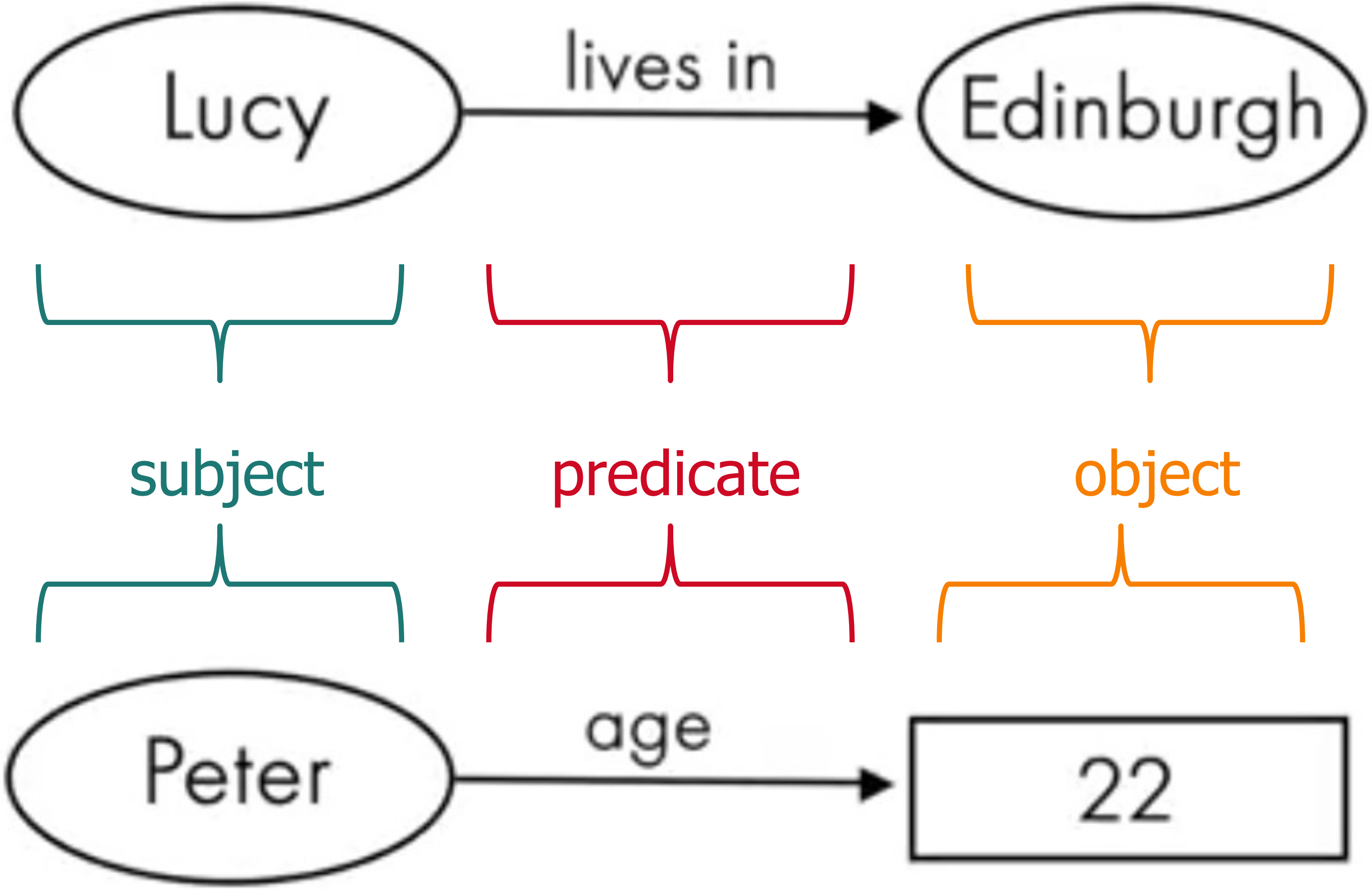


subject

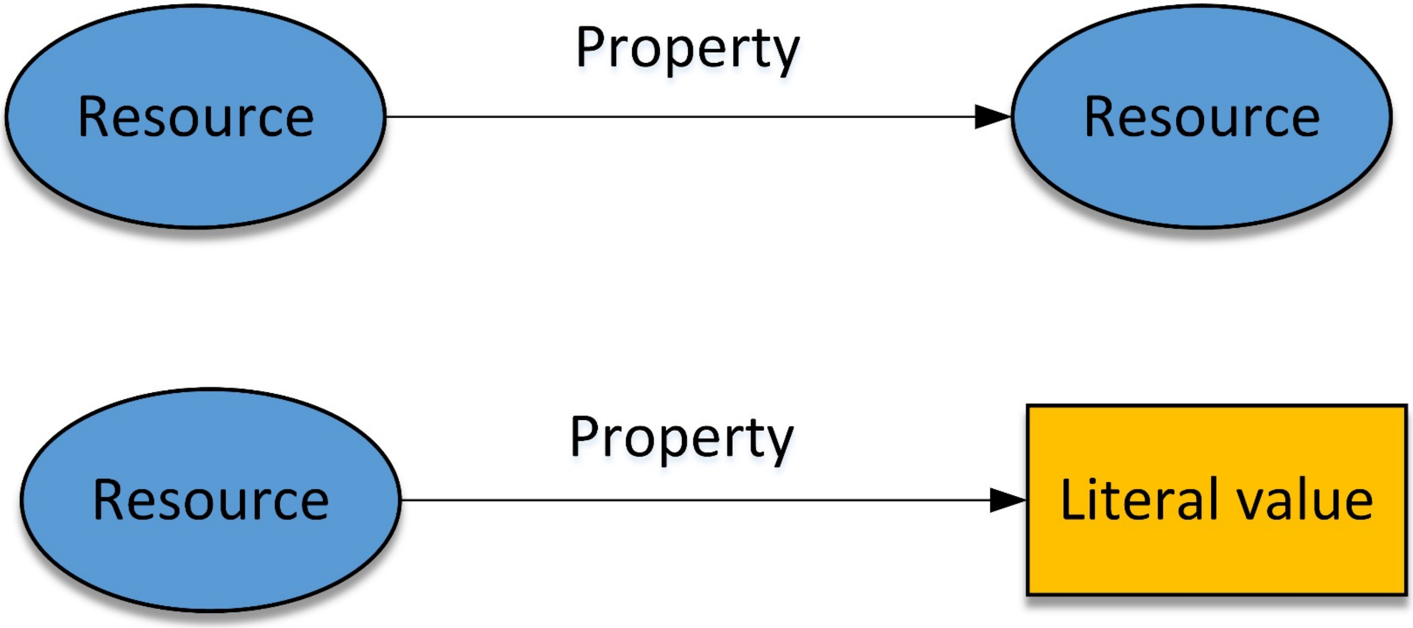
predicate

object

RDF triple visualisation



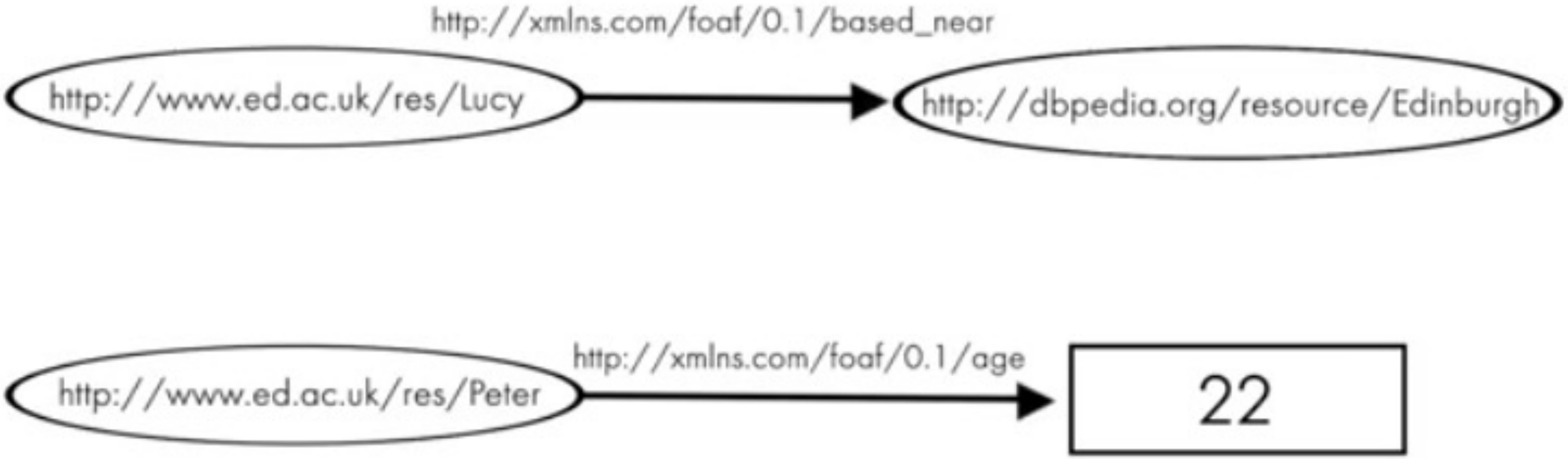
RDF triple visualisation



Unique identifiers: URIs

- In RDF, we use URIs (Uniform Resource Identifiers) to uniquely identify concepts and entities.
- Examples:
  - http://dbpedia.org/resource/Edinburgh
  - http://xmlns.com/foaf/0.1/age
- URIs are used for both resources and properties.

Unique identifiers: URIs


http://xmlns.com/foaf/0.1/based_near
http://www.ed.ac.uk/res/Lucy
http://dbpedia.org/resource/Edinburgh
http://xmlns.com/foaf/0.1/age
http://www.ed.ac.uk/res/Peter
22

How to use URIs

- 1st approach (recommended): use existing URIs
  - DBPedia (http://dbpedia.org) is a very good source of URIs.
    - Every resource that is the subject of a page in Wikipedia has a corresponding URI in DBpedia.
    - URI forEdinburgh:

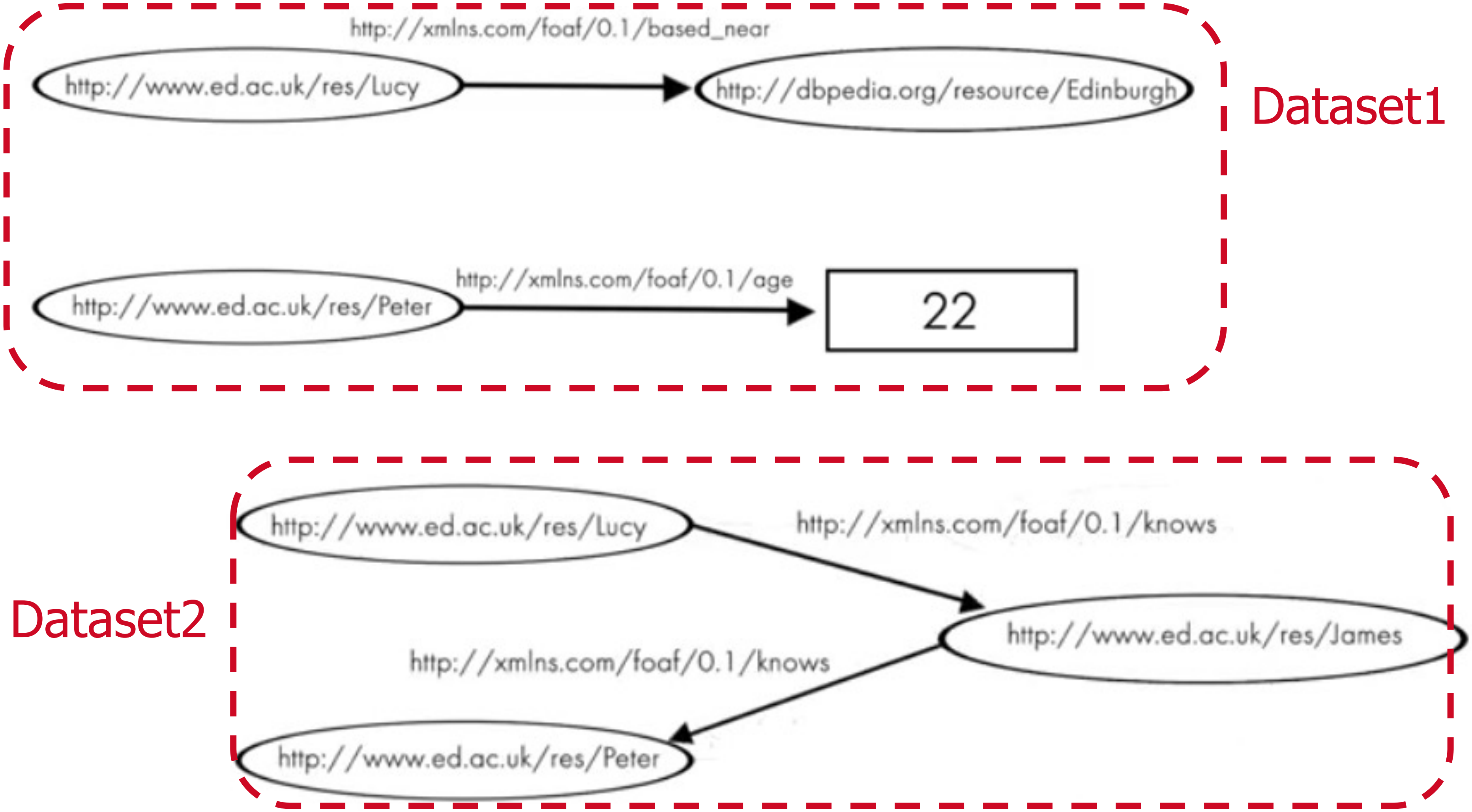
      http://dbpedia.org/resource/Edinburgh
- 2nd approach: create your own URIs
  - If you don't own a domain name, you can use http://example.com/

    http://example.com/id/EwanMcGregor
  - Keep it simple

Merging RDF data is easy!

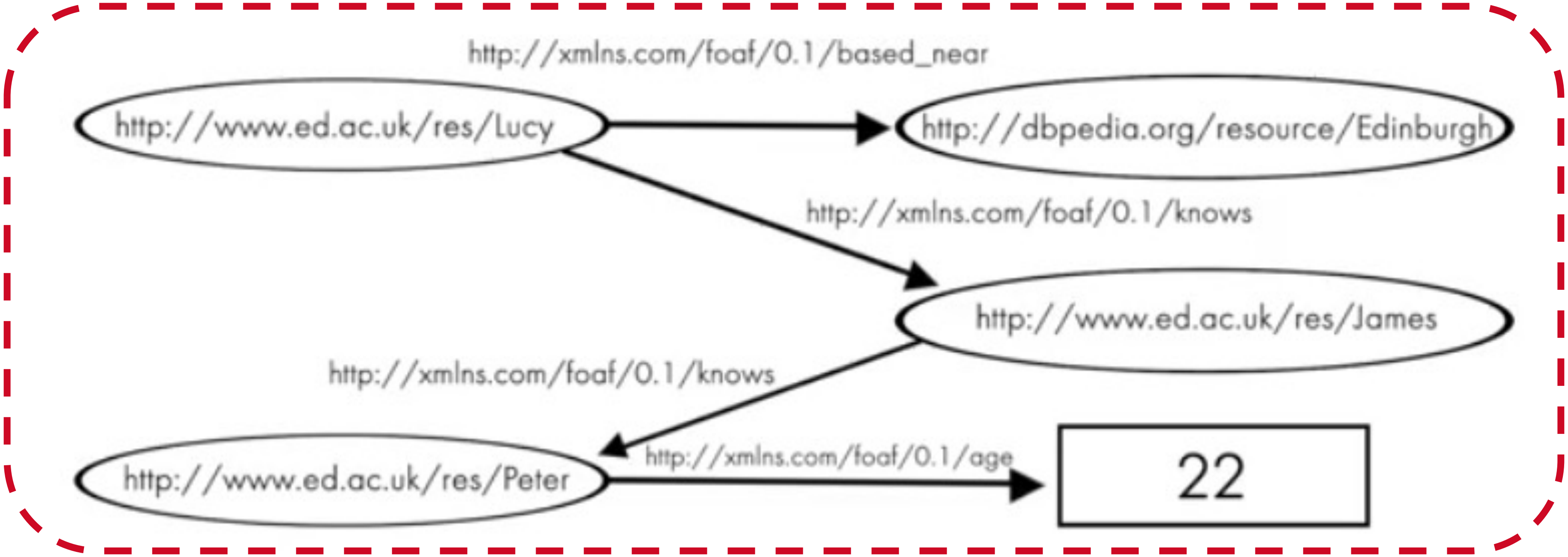
- By uniquely identifying resources with the use of URIs, we can easily link data about the same resource.
- Merging different RDF datasets is simply a matter of bringing the two sets of RDF statements together.

Merging RDF data is easy!


http://xmlns.com/foaf/0.1/based_near
http://www.ed.ac.uk/res/Lucy
http://dbpedia.org/resource/Edinburgh
http://xmlns.com/foaf/0.1/age
http://www.ed.ac.uk/res/Peter
22
Dataset1
http://www.ed.ac.uk/res/Lucy
http://xmlns.com/foaf/0.1/knows
http://www.ed.ac.uk/res/James
http://xmlns.com/foaf/0.1/knows
http://www.ed.ac.uk/res/Peter
Dataset2

Merging RDF data is easy!

Dataset3 = Dataset1 + Dataset2


http://xmlns.com/foaf/0.1/based_near
http://www.ed.ac.uk/res/Lucy
http://dbpedia.org/resource/Edinburgh
http://xmlns.com/foaf/0.1/knows
http://www.ed.ac.uk/res/James
http://xmlns.com/foaf/0.1/knows
http://www.ed.ac.uk/res/Peter
http://xmlns.com/foaf/0.1/age
22

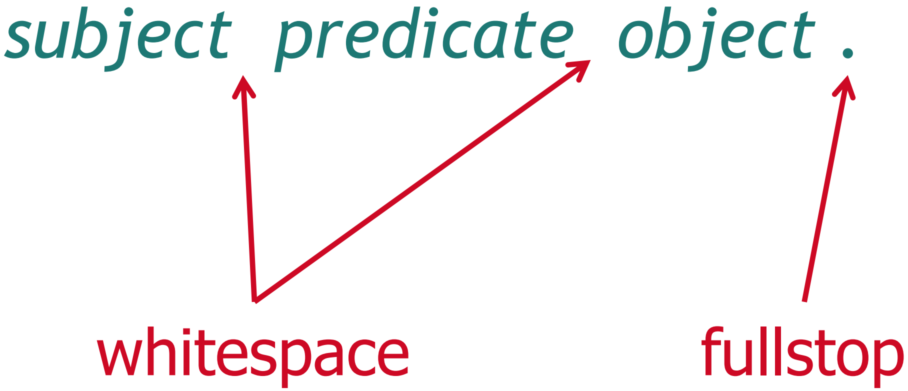
Writing RDF statements in Turtle

- Turtle (Terse RDF Triple Language): One of the most popular forms of syntax for expressing RDF.
- General form:

  subject predicate object .

Writing RDF statements in Turtle

- Turtle (Terse RDF Triple Language): One of the most popular forms of syntax for expressing RDF.
- General form:



subject predicate object .

whitespace

fullstop

Writing RDF statements in Turtle

- Turtle (Terse RDF Triple Language): One of the most popular forms of syntax for expressing RDF.
- General form:

  subject predicate object .
- When using URIs, these should be enclosed in angle brackets, e.g.
  <http://dbpedia.org/resource/Edinburgh>

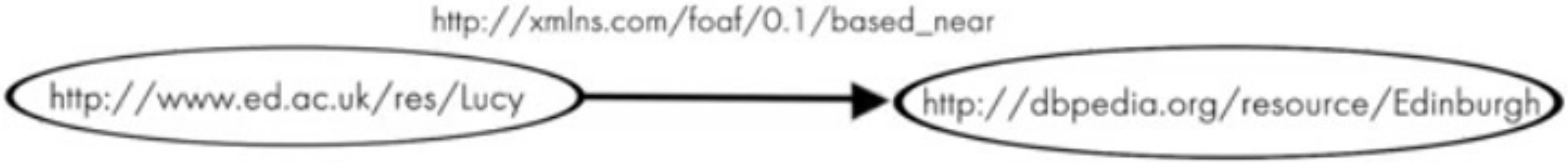
Example RDF statements in Turtle

```
<http://www.ed.ac.uk/res/Lucy>  <http://xmlns.com/foaf/0.1/based_near>  <http://dbpedia.org/resource/Edinburgh> .
```

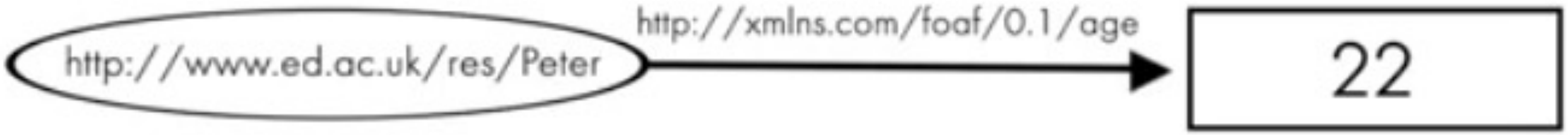
```
<http://www.ed.ac.uk/res/Peter>  <http://xmlns.com/foaf/0.1/age>  22 .
```

Example RDF statements in Turtle

<http://www.ed.ac.uk/res/Lucy> <http://xmlns.com/foaf/0.1/based_near> <http://dbpedia.org/resource/Edinburgh> .


http://xmlns.com/foaf/0.1/based_near
http://www.ed.ac.uk/res/Lucy
http://dbpedia.org/resource/Edinburgh

<http://www.ed.ac.uk/res/Peter> <http://xmlns.com/foaf/0.1/age> 22 .



Ontologies

- Ontology definition: A formal, explicit specification of a shared conceptualisation.
- Essentially, a way of encoding domain knowledge.
- Something like an enhanced dictionary, where you can look up the meaning of different concepts and find relations between them.

Ontology Components

- Possible components include:
  - Classes (e.g. Woman)
  - Individuals (e.g. Lucy)
  - Attributes (e.g. Age)
  - Relations (e.g. MotherOf)
- Ontologies often contain a class taxonomy.
- Formal definitions of classes may also be included.

Why are ontologies useful?

- They allow us to attach meanings to data.
  - e.g. when a dataset uses the term “Viral pneumonia”, we know what is meant
- They enable the standardisation of terminology.
  - e.g. the same term “Viral pneumonia” is used across different datasets for the same disease
- They allow us to infer new knowledge from existing data.
  - If we know that James is suffering from viral pneumonia, and our ontology specifies that Viral pneumonia is a subclass of Lung disease, then we can infer that James is suffering from a lung disease.

Medical Ontologies

- Gene Ontology:
  - http://www.geneontology.org/
  - It represents information about biological processes, cellular components and molecular functions.
- Disease Ontology:
  - http://disease-ontology.org/
  - It provides descriptions of human disease terms, phenotype characteristics and related medical vocabulary disease concepts.

Medical Ontologies

- SNOMED-CT:
  - https://www.snomed.org/snomed-ct
  - It is a collection of medical terms. It includes codes, terms, synonyms and definitions used in clinical documentation and reporting.
  - It is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world.

Conclusions

- Ontologies allow for a common and unambiguous understanding of different concepts in the datasets. This is crucial when sharing medical data.
- The RDF data model makes it possible to easily link data and discover previously unknown relationships between different concepts.
- The RDF data model is very flexible, allowing us to add new nodes, new kinds of relationships, and new subgraphs to an existing structure without disturbing existing functionality.