

# Querying Data Provenance in the Internet of Things

Argyro Avgoustaki, Giorgos Flouris, Irini Fundulaki, Dimitris Plexousakis

{argiro,fgeo,fundul,dp}@ics.forth.gr

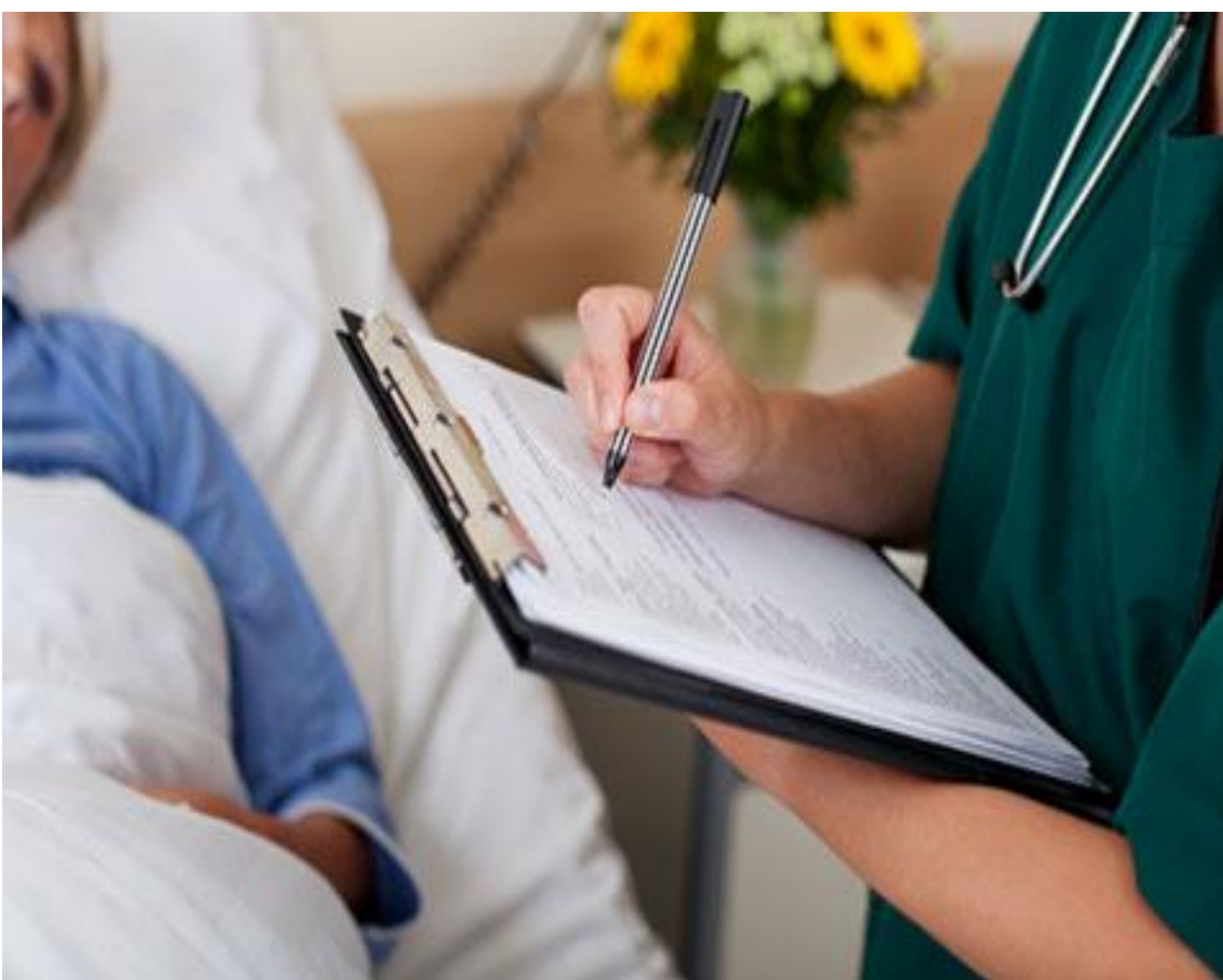


## Do they look familiar to you?

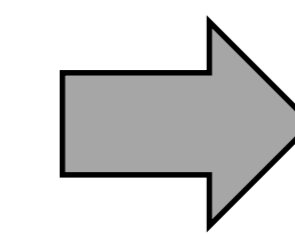
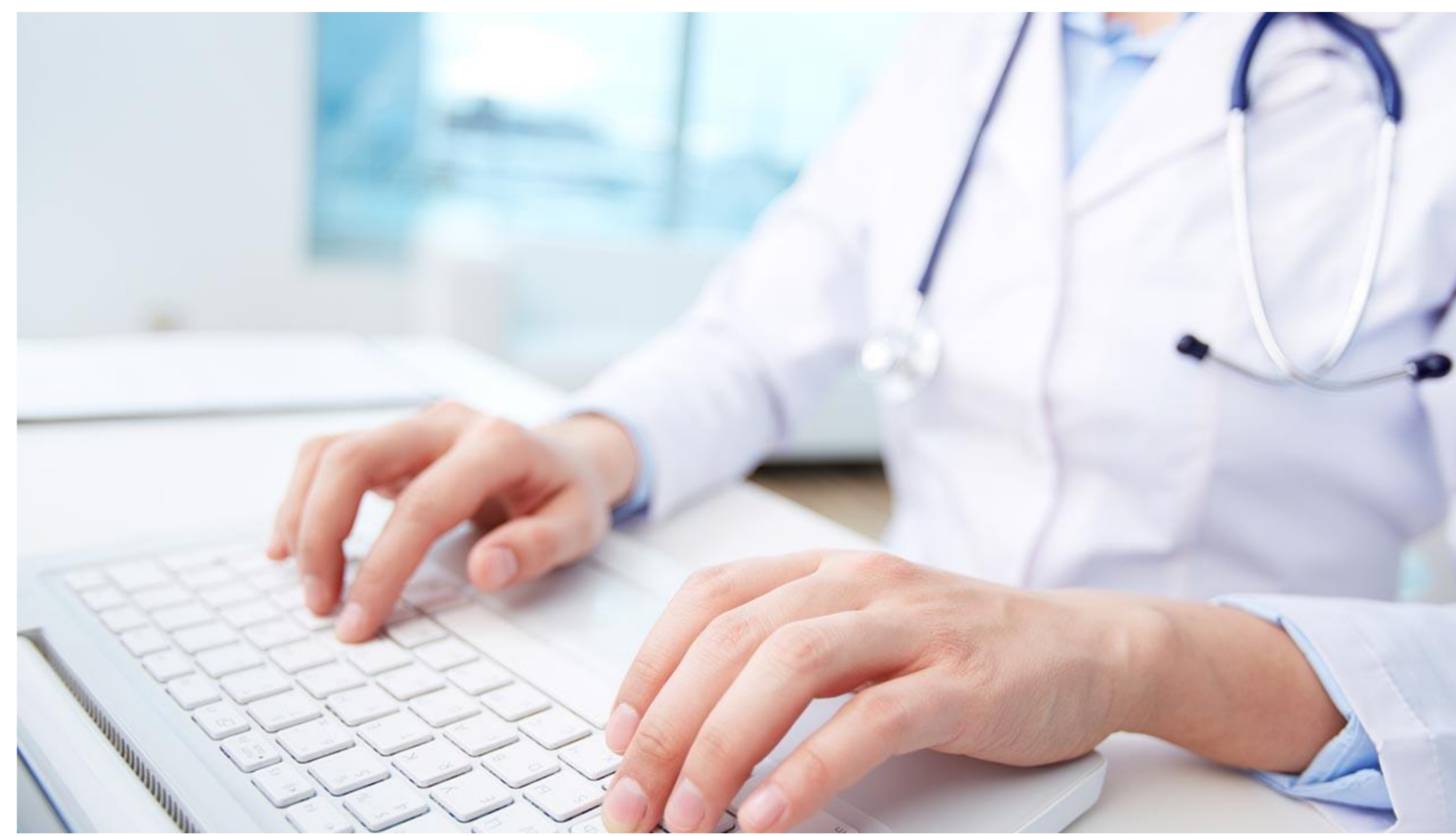


Every time you “Like”, “Share” or “Comment” a fake post you help a fraud to get wealthy! What if we could know about the trustworthiness of the post?

### 1 Recording medical history



### 2 Inserting medical history into the hospital database



**1 million medical errors/year** because of wrong data inputs due to human factor. What if we could know the reliability of people?

## Why do we need “provenance”?

Nowadays, knowing from where and how data has come from is of crucial importance. Recording the *provenance* of data, i.e. their history, allows us to support applications such as:

- ✓ Data Quality
- ✓ Reliability
- ✓ Trustworthiness
- ✓ Copyrights
- ✓ Access Control
- ✓ Accountability

## Our Work

Until now a great number of provenance models have been proposed. Thus, we are able to know from *where* and *how* each piece of information was generated. In our work, we introduce a query language (*ProvQL*) that is suitable for seeking information related to data provenance. ProvQL can answer queries like the following:

1. Which data records or sources contributed in deriving a data record?
2. Identify all data records whose provenance includes a specific data source (or data item)
3. Identify all sources referring to “Donald\_Trump” that originate from a specific source
4. Assessing the trustworthiness of a data item

## ProvQL Examples

- ✓ `SELECT PROV(?id) WHERE QUADS(?id) = (?a, ?b, “Donald Trump”, ?n) (3)`
- ✓ `SELECT QUADS(?id) WHERE PROV(?id) CONTAINS ?v AND QUADS(?v) = (<a>, <b>, <c>, <d>)(2)`

This work is funded and supported by

