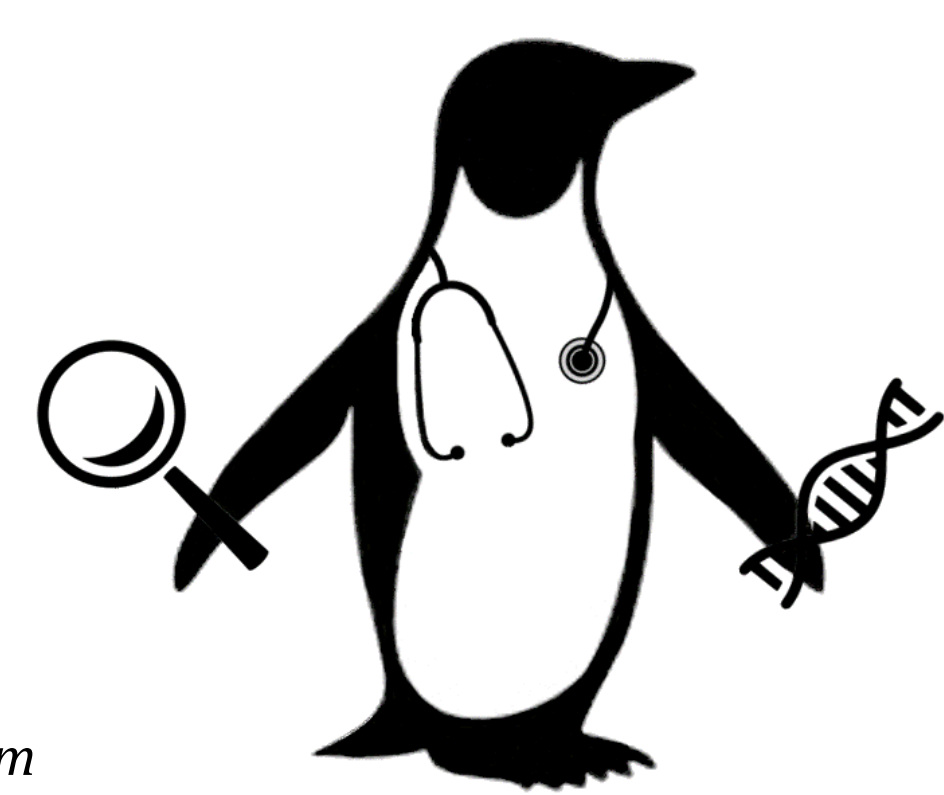# PENGQUIN:
PErsoNal Genome QUery IN Healthcare and Clinical Practice

Elias Crum[1,2]

[1]IDLab, Department of Electronics and Information Systems, Ghent University – imec, Belgium
[2]Flemish institute for Technological Research (VITO) Mol, Belgium

IDLab INTERNET & DATA LAB
VITO
imec
GHENT UNIVERSITY

Genome Data for Clinical Care?
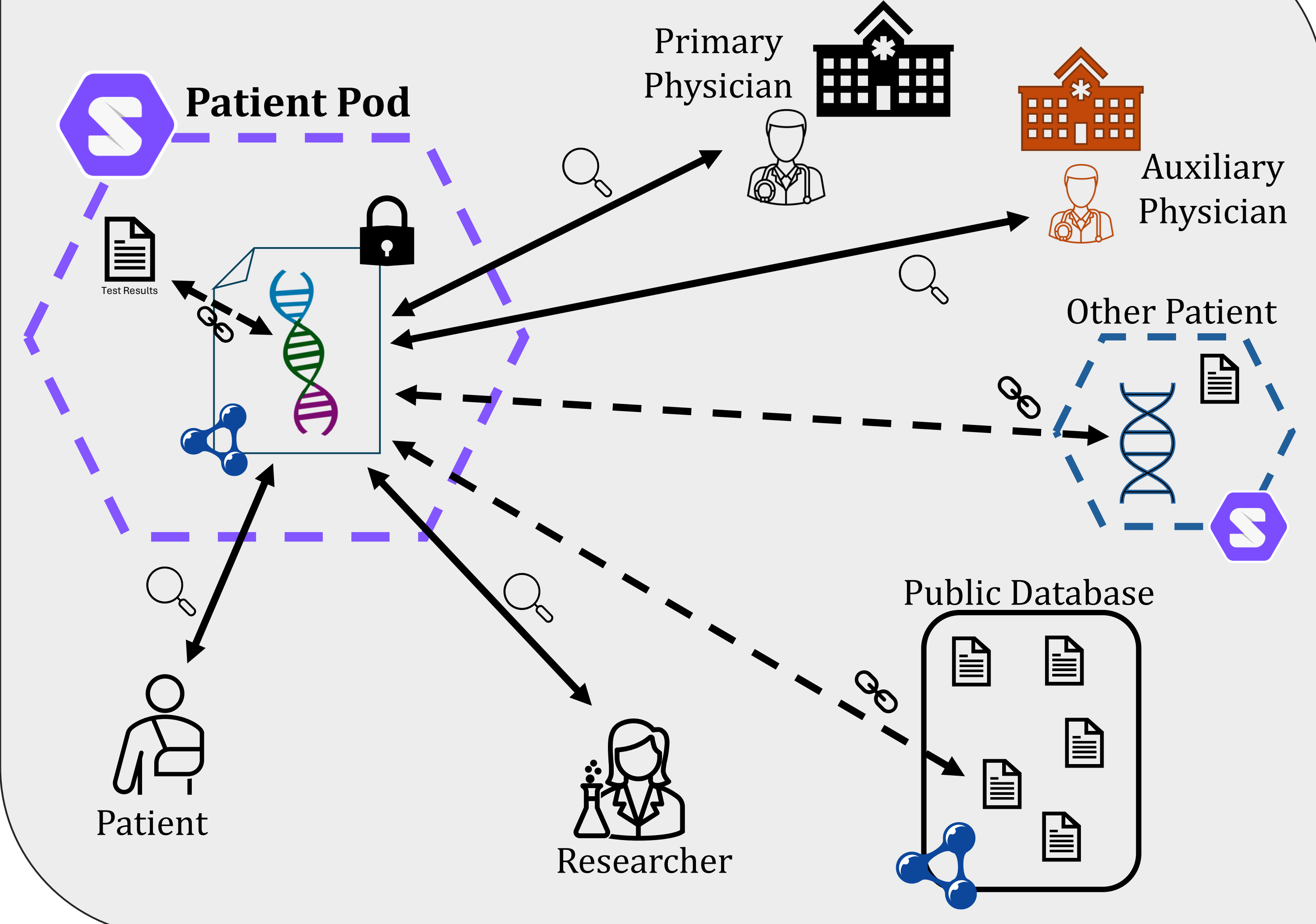
Still Many challenges:
- Storage
- Privacy
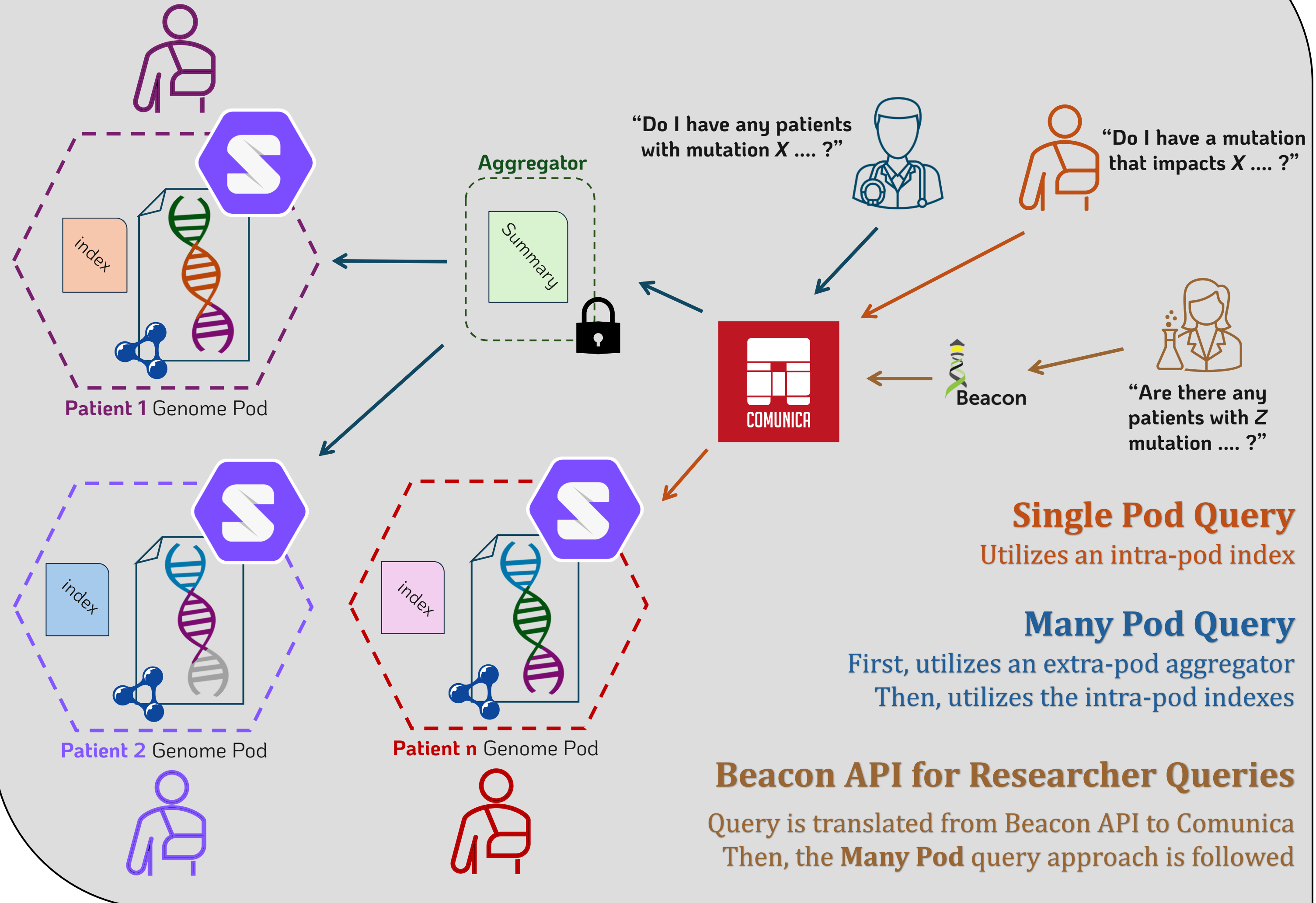- Governance
- Formatting

Could Semantic Web technology help?

Personal Genome Sequences

- **VCF** Variant Call Format (VCF)
- Very large (~GBs)
- **GDPR** Strict privacy protections

A Patient-Centric Storage Approach

Patient Pod
Test Results
Primary Physician
Auxiliary Physician
Other Patient
Patient
Researcher
Public Database

An Index-Based Link Traversal Querying Approach

Aggregator
Summary
"Do I have any patients with mutation X …. ?"
"Do I have a mutation that impacts X …. ?"
COMUNICA
Beacon
"Are there any patients with Z mutation …. ?"
index
Patient 1 Genome Pod
index
Patient 2 Genome Pod
index
Patient n Genome Pod

Single Pod Query
Utilizes an intra-pod index

Many Pod Query
First, utilizes an extra-pod aggregator
Then, utilizes the intra-pod indexes

Beacon API for Researcher Queries
Query is translated from Beacon API to Comunica
Then, the **Many Pod** query approach is followed

Research Challenges

Privacy granularity
- Index privacy concerns?
- How granular is adequate?

RDF Conversion + Data Linkages
- VCF to RDF Conversion?
- What data should be linked?
- Automation of data linking?

Querying performance
- Many, large data sources
- Algorithms that use indexes?

Engineering Tasks ⚙️

- Genomic data storage in Solid pods
- Data privacy management
- Data Sharing via the web

Research Pursuits 🔍

- Represent genomic data as RDF?
- Linking genomic and health data?
- Single and Many pod Querying?

Ph.D. Symp. Paper
Contact