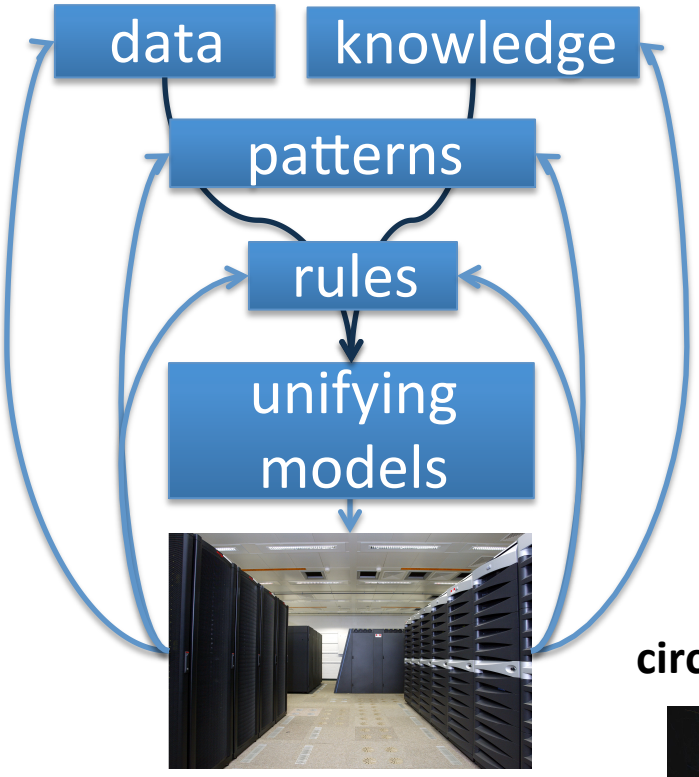
Extracting value out of medical data

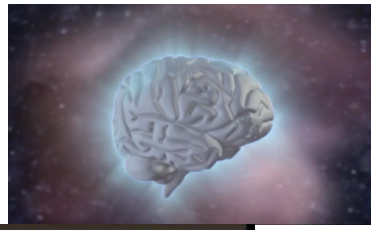
Anastasia Ailamaki

EPFL and RAW Labs

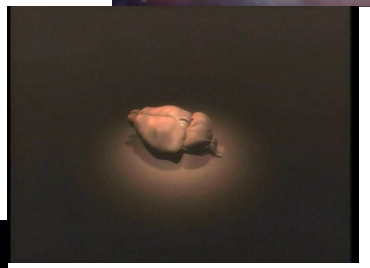
human brain project

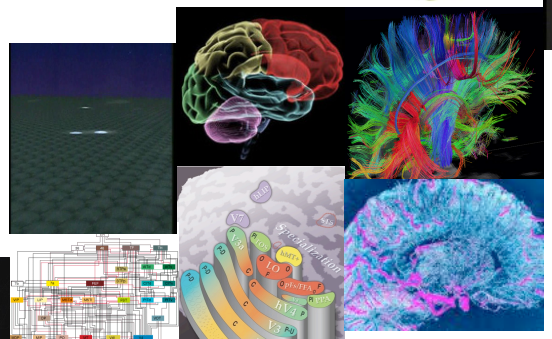


cognition

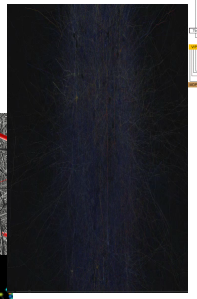


whole brain

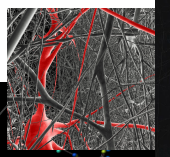




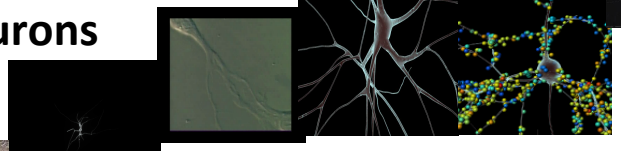
circuits



synapses

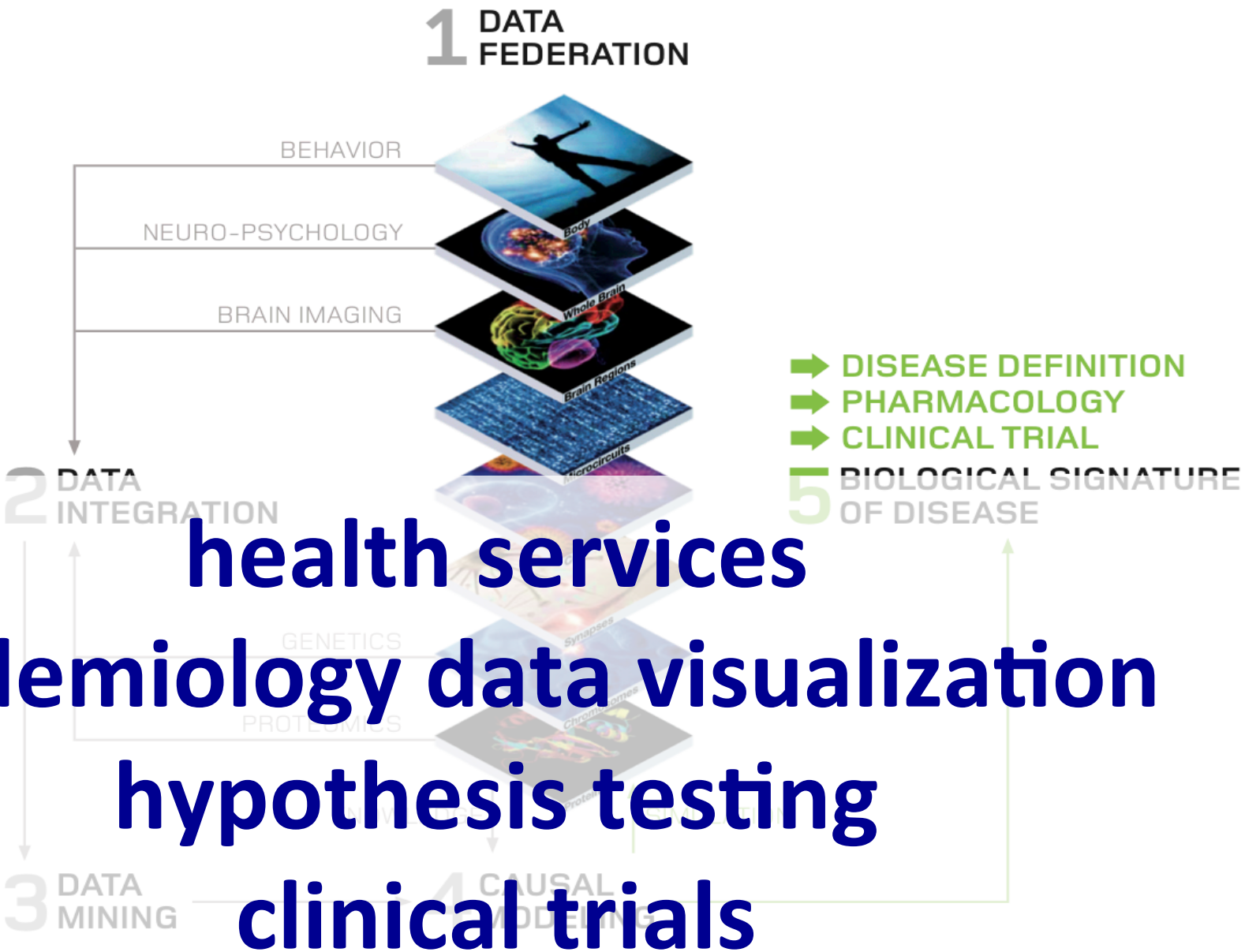


neurons



integrate clinical and simulation data

medical informatics



health services

epidemiology data visualization

hypothesis testing

clinical trials

biological disease signatures

the coupling of
clinical measurements with
validated biomarkers

Example: Alzheimer's disease

Clinical - Phenotype	Proteomic Biomarkers	Genomic Biomarkers	Structural Biomarkers
Cognition: memory	**CSF protein**: beta amyloid	**Mutation**: APP, PSEN1, PSEN2	**Volumetric change**: hippocampus, inferior temporal cortex…
Functional capacity		**Common genetic variant**: APOE e4e4 …	**Beta amyloid imaging**
General			
Quality of life			

memory loss/tau level/APOE e4e4/hippocampus atrophy/amyloid

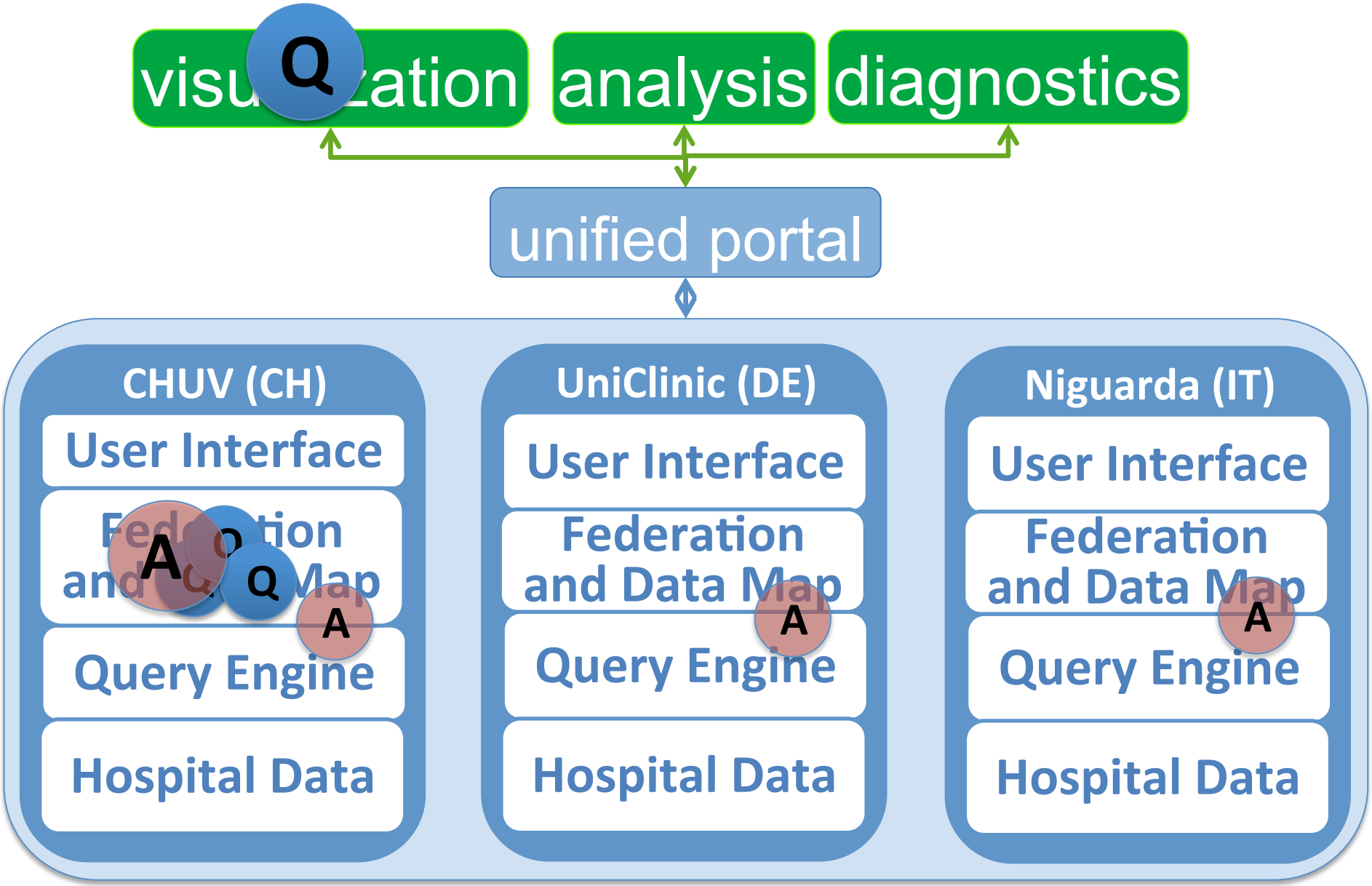
what's in this talk

Medical informatics platform

Local hospital data stores

Ad-hoc queries with RAWSQL

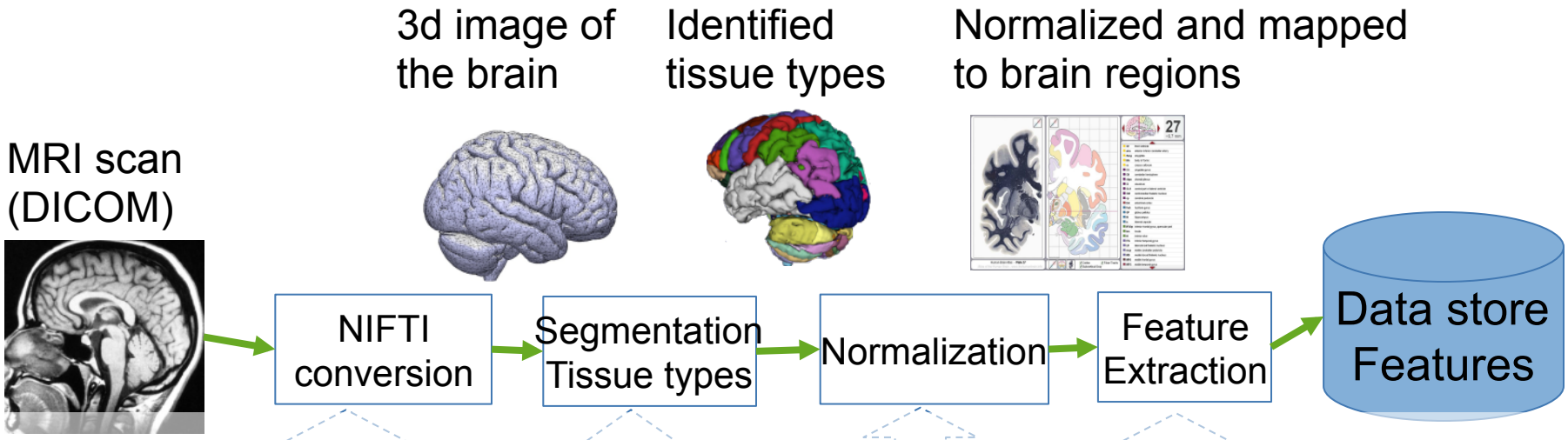
Medical Informatics Platform



Recruiting hospitals

- visited NHS headquarters in London, Niguarda in Milan, and Uniklinik in Freiburg (more to come)
- vast difference in requirements (even privacy)
  - invariant: bureaucracy
- use cases include support for
  - medical research
  - efficient patient treatment
- like the MIP, but want to get to know their own data

Image preprocessing @CHUV



intermediate data also available

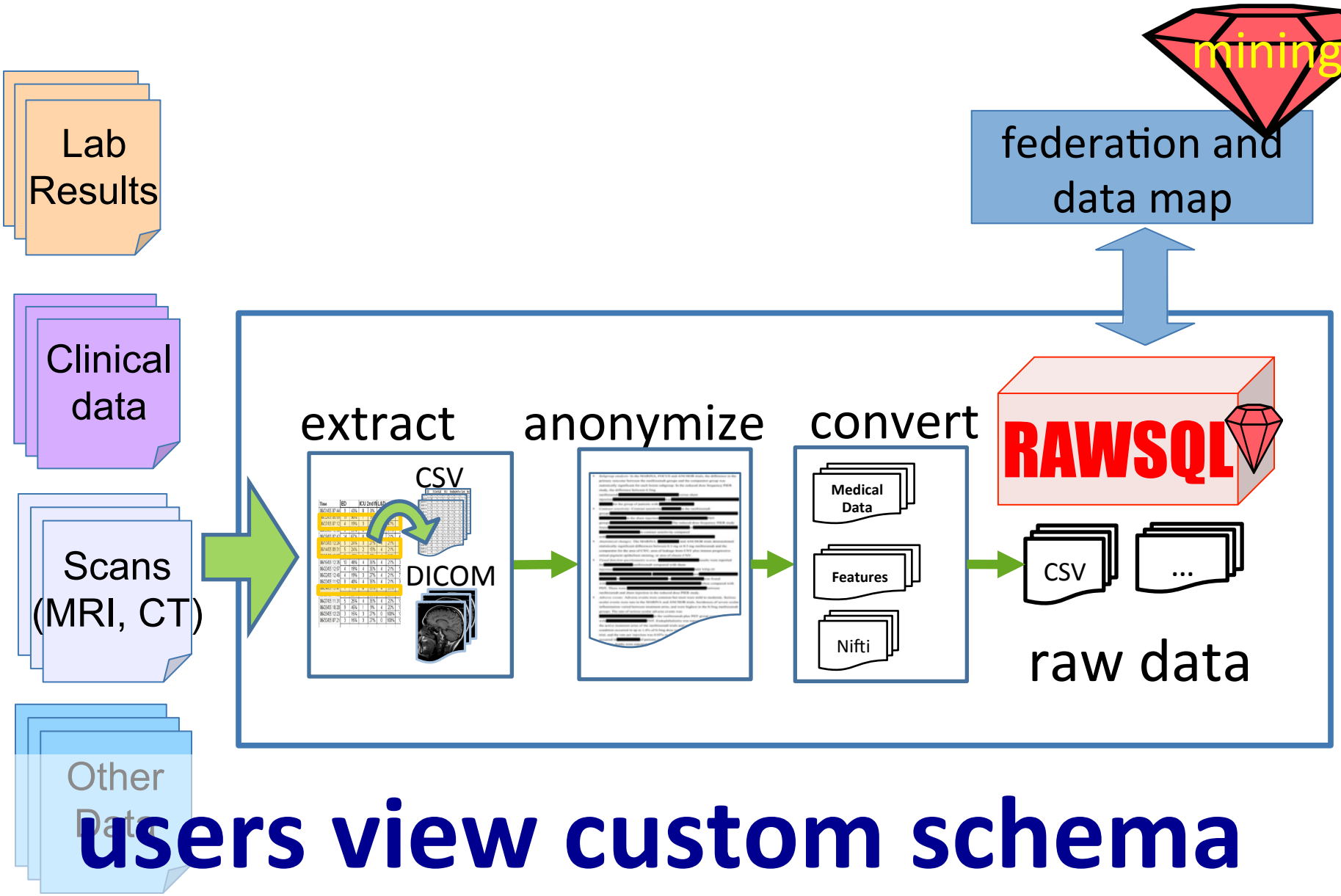
script converts DICOM to NIFTI format (3d volume)

Image processing to discover the type of tissue each voxel corresponds to

data normalized to same standards, format

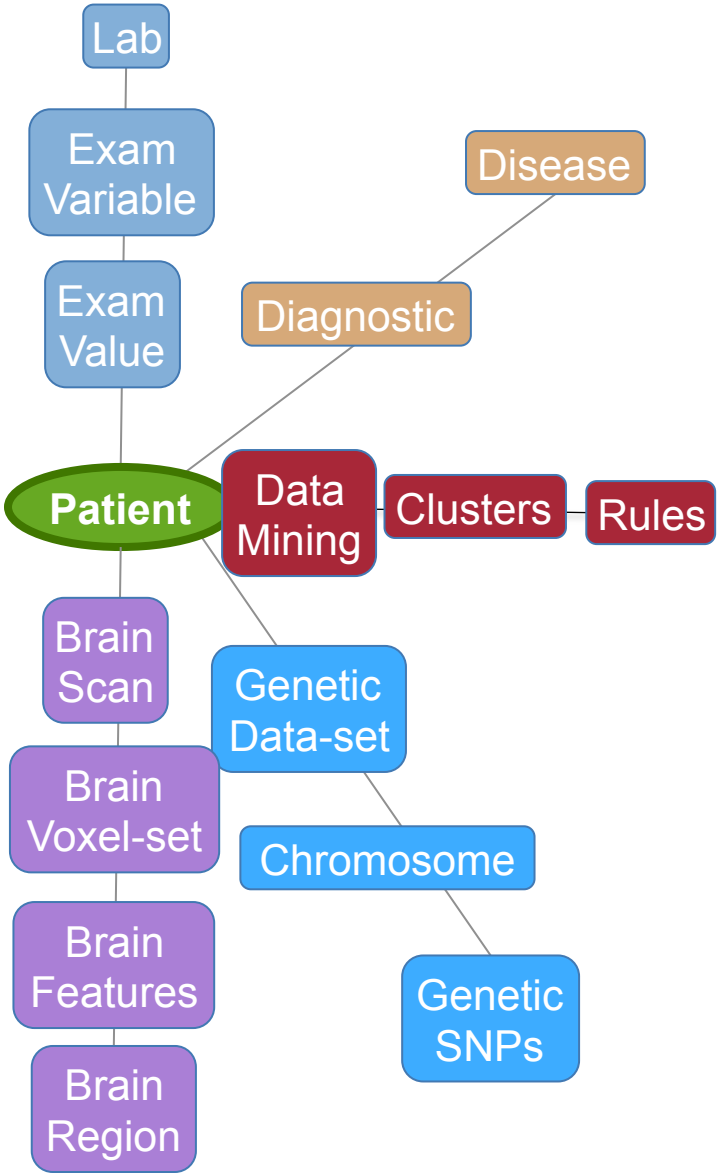
extract interesting features from normalized data

CHUV data store



users view custom schema

the user's perspective



two-level anonymization

each hospital's responsibility

clinical+genetic+imaging data → signature

Patients (CSV)

id	Protein : AACT	Age	Phenotype	…
1	1.4	45	Trauma	…
2	2	55	Chronic Symptoms	…
3	0.2	56	…	…

Brain_GrayMatter (Binary)

	0	1	…	n
0	0.45	0.75	…	0.1
1	0.33	0.3	…	0.38
…	…	…	…	…
m	0.12	0	…	0.47

signature:

age > 50
AND
amygdala.Vol > 0.3
AND
AACT < 1

BrainRegions (JSON)

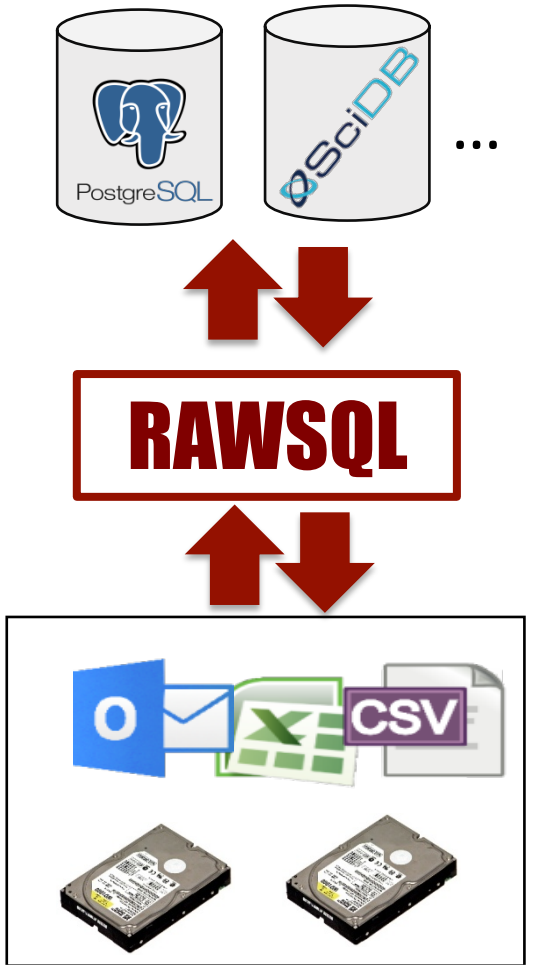
```
[{"id": 1,
 "amygdala": {"X":15,"Y":20, “Vol”: 0.5},
 "hippocampus": {"X":17, "Y":10, “Vol”:0.2}},
{"id": 2, ...},
{"id": 3, ...}]
```

challenge: data integration & ad-hoc queries

queries on heterogeneous data

cannot load into a Database System!

- heterogeneous formats
- legacy software
- privacy limitations
- data "owned" by one database



RAWSQL: interface to raw data

SQL, SCALA, iPython notebooks

code-generated engine

key: data virtualization

Experimental Setup

- Intel(R) Xeon(R) CPU E5-2660 @ 2.20GHz
- 128 GB RAM
- 7500 RPM SATA

Relation name	Tuples	Attributes	Size	Type
Patients	41718	156	29 MB	CSV
Genetics	51858	17832	1.8 GB	CSV
BrainRegions	17000	20446	5.3 GB	JSON

```
SELECT val1, ..., valN
FROM Patients p
JOIN Genetics g ON (p.id = g.id)
JOIN BrainRegions b ON (g.id=b.id)
WHERE pred1 AND ... AND predN
```

```
for {  p <- Patients,
       g <- Genetics,
       b <- BrainRegions,
       p.id=g.id, g.id=b.id,
       pred1, ..., predN
    } yield val1,…,valN
```

RAWSQL vs State-of-the-art

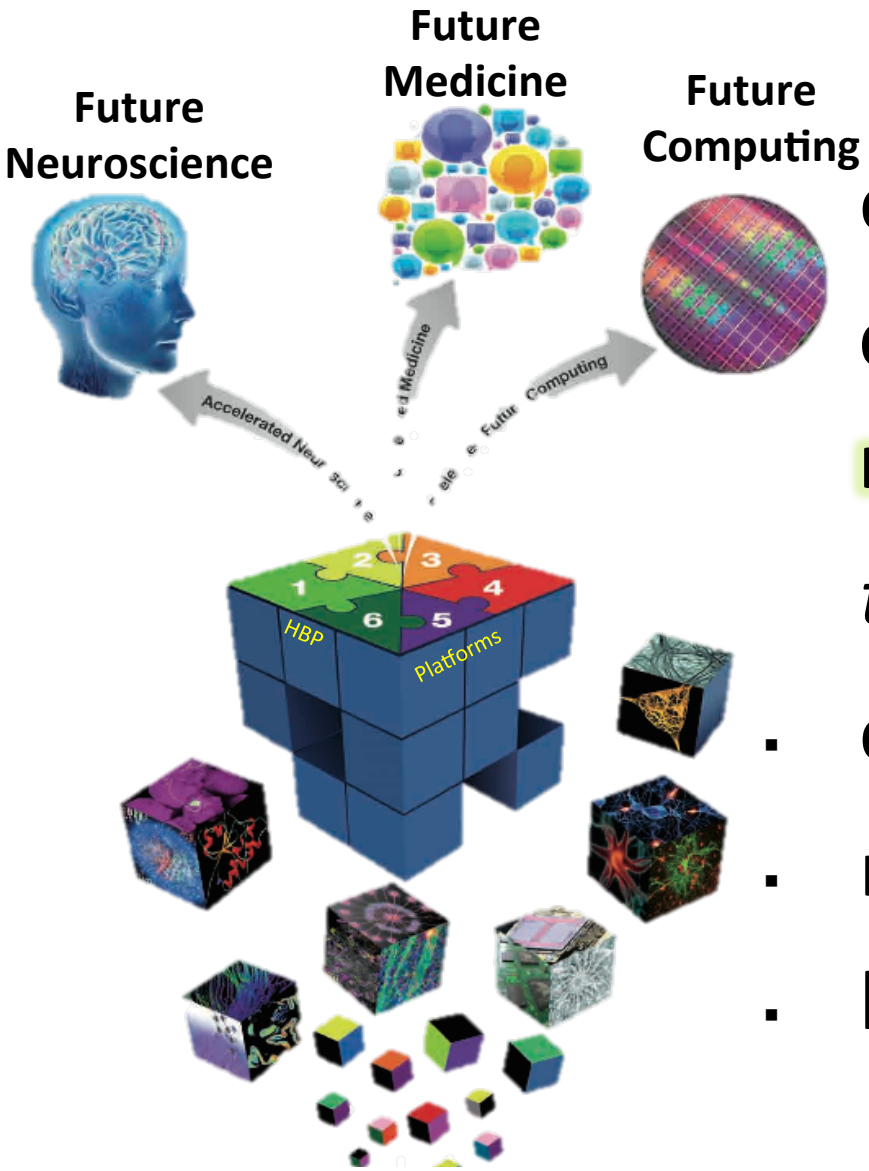
150 analytics queries on CSV & JSON data

fully support heterogeneous data

runs standalone or on Spark

richer analytical operators

summary



clinical data federation with

query technologies that respect anonymity

to support

- complex dataflow processing
- rule-based clustering
- biological disease signatures