Compositional security and privacy for biomedical analyses using shared genetic data

Mario Südholt

(joint work with Fatima-zahra Boujdad)

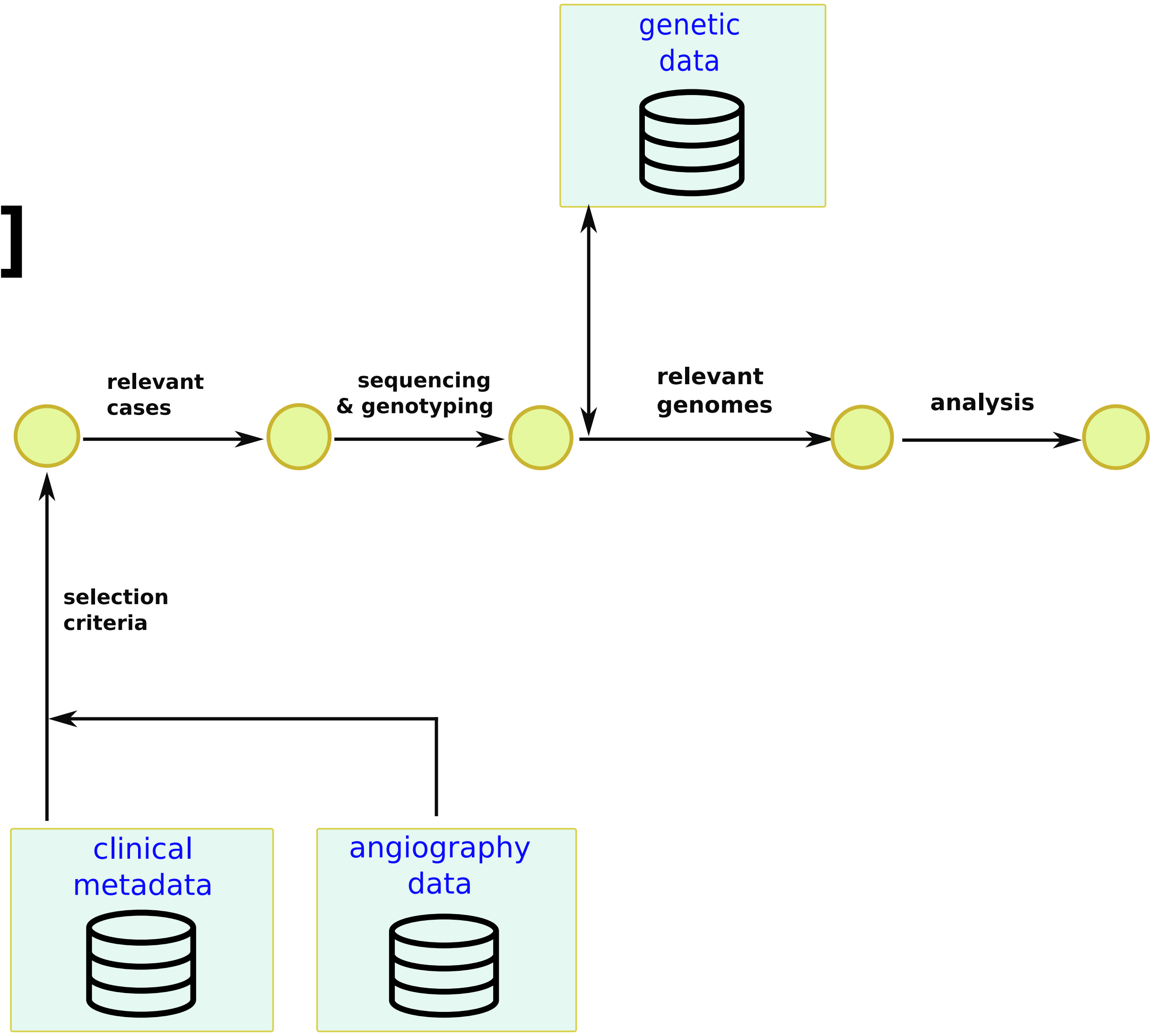
Meeting Biosphère 7, 24 Jan. 2017

IMT Atlantique

Analyses and shared data

Ex.: Recent result on
intracranial aneurysm [BOU18]

- **Manual introspection** of multiple databases
- **Manual selection** of subjects for genome sequencing and analysis



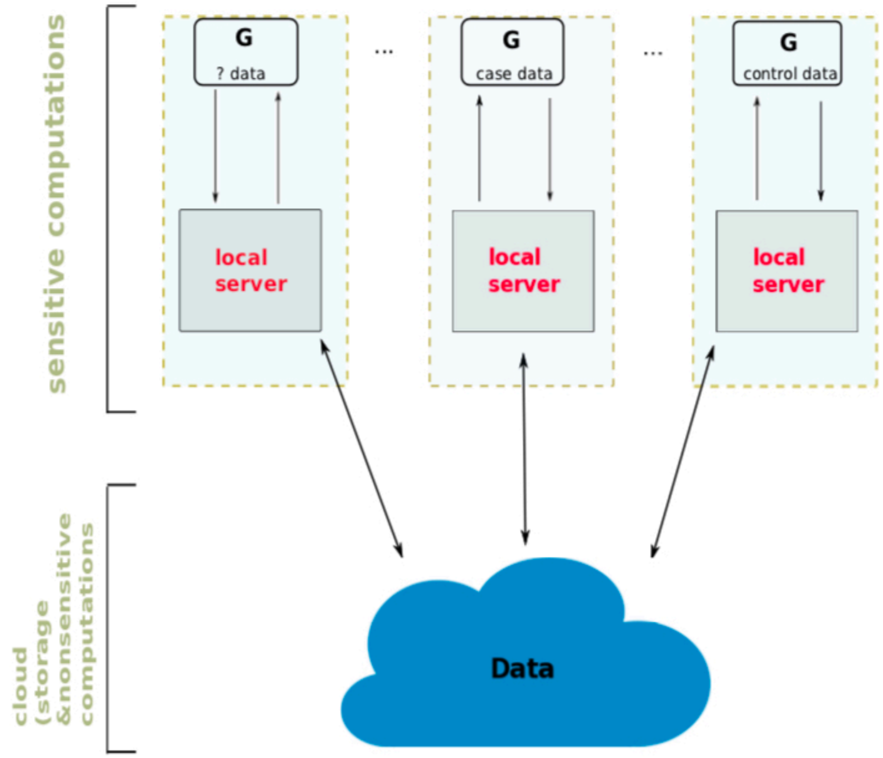
[Bou18] Bourcier R, Le Scouarnec S, Bonnaud S, Karakachoff M, Bourcereau E, Heurtebise-Chrétien S, Menguy C, Dina C, Simonet F, Moles A, Lenoble C, Lindenbaum P, Chatel S, Isidor B, Génin E, Deleuze JF, Schott JJ, Le Marec H; ICAN Study Group, Loirand G, Desal H, Redon R. **Rare Coding Variants in ANGPTL6 Are Associated with Familial Forms of Intracranial Aneurysm.** Am J Hum Genet. 2018 Jan 4;102(1):133-141. doi: 10.1016/j.ajhg.2017.12.006.

Data sharing: potential benefits

- **Share clinical and research data** in hospitals

- **Co-locate analyses with data**

- **Facilitate access** using Cloud storage and computations



Data sharing: issues

- **Socio-economic issues**
  - Data is **valuable**: potential losses through unrestricted sharing
  - **Transfer** of large data may be time consuming or costly
- **Technical issues**
  - Guarantee **privacy** properties: no data divulgation, no data re-identification
  - Preserve **ownership** information
  - Ensure data **integrity**

Sharing requirements

- **Protect data** from unauthorized access
- Support **de-identification** of data
- **Move analyses to data**
- Mark data with **ownership information**
- Support **traceability** of data

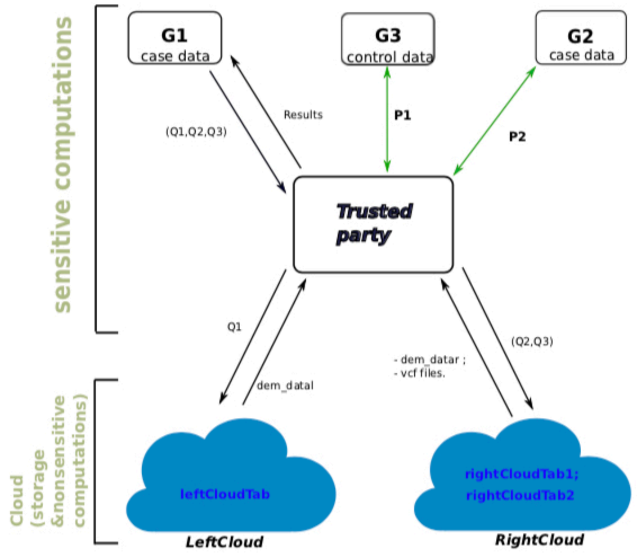
The COSHED approach

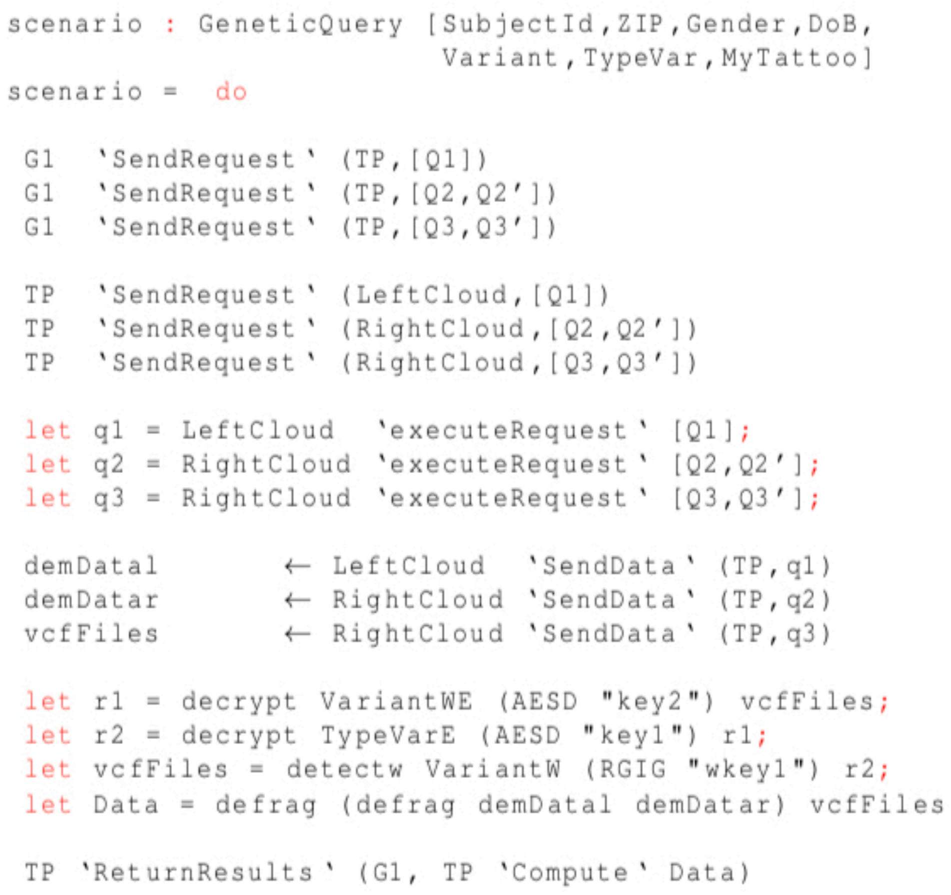
Enforcement mechanisms

- **Encryption**: protect agains unauthorized access
- **Fragmentation**: support de-identification of plain data
- **Localized certified computations**: co-localize trusted analyses with data
- **Watermarking**: support for ownership, traceability and integrity

Program secure workflows

- Declare **database fragmentation**
- **Encrypt** data
- Apply **watermarks**
- **Execute certified analyses** locally or remotely





```
scenario : GeneticQuery [SubjectId,ZIP,Gender,DoB,
                          Variant,TypeVar,MyTattoo]
scenario =  do

 G1  `SendRequest` (TP,[Q1])
 G1  `SendRequest` (TP,[Q2,Q2'])
 G1  `SendRequest` (TP,[Q3,Q3'])

 TP  `SendRequest` (LeftCloud,[Q1])
 TP  `SendRequest` (RightCloud,[Q2,Q2'])
 TP  `SendRequest` (RightCloud,[Q3,Q3'])

 let q1 = LeftCloud  `executeRequest` [Q1];
 let q2 = RightCloud `executeRequest` [Q2,Q2'];
 let q3 = RightCloud `executeRequest` [Q3,Q3'];

 demDatal      ← LeftCloud  `SendData` (TP,q1)
 demDatar      ← RightCloud `SendData` (TP,q2)
 vcfFiles      ← RightCloud `SendData` (TP,q3)

 let r1 = decrypt VariantWE (AESD "key2") vcfFiles;
 let r2 = decrypt TypeVarE (AESD "key1") r1;
 let vcfFiles = detectw VariantW (RGIG "wkey1") r2;
 let Data = defrag (defrag demDatal demDatar) vcfFiles

 TP `ReturnResults` (G1, TP `Compute` Data)
```

Ex.: database def.

```
Database:
  Subject    (SubjectId,ZIP,DoB,Gender,CaseCtrl)
  SubjectVcf (recordId,Variant,TypeVariant
              position,SubjectId)
```

- Fragmentation for confidentiality
  triplet `(zip,gender,DoB)` forms quasi-identifier: store pair `(zip,gender)` and `DoB` in different Clouds.

- Encryption for confidentiality
  vcf file is symmetrically encrypted

- Watermarking: ownership/integrity protection of genomes

- Client-side computations used for TP computations.

Database def. 2

Resulting relational database :

```
leftCloudTab   (SubjectId, ZIP, Gender)

rightCloudTab1 (RecordId, VariantWE, TypeVarE,
                position, SubjectId)

rightCloudTab2 (SubjectId, DoB, CaseCtrl)
```

Security/privacy props.

Prove properties using **composition algebra**

Laws for watermarking

$$decrypt_{(s,a)} \circ crypt_{(s,a)} \circ detectw_a \circ wat_a \equiv$$
$$detectw_a \circ decrypt_{(s,a)} \circ crypt_{(s,a)} \circ wat_a$$

$$\pi_a \circ detectw_a \equiv detectw_a \circ \pi_a$$

$$detectw_a \circ \sigma_p = \sigma_p \circ detectw_a \qquad if\, dom(p) \cap a = \emptyset$$

Derivation of distributed query

$$\pi_{(variant,typeVar)} \circ$$
$$\sigma_{((subjectId \in mdd) \wedge (position=i, position=j,..))}$$

(a) local query

$$\pi_{(variant,typeVar)} \circ$$
$$\sigma_{((subjectId \in mdd) \wedge (position=i, position=j,..))} \circ$$
$$decrypt_{variant,typeVar} \circ crypt_{variant,typeVar} \circ$$
$$detectw_{variant} \circ wat_{variant}$$

$laws\ 3,4,6,7,8 \quad \downarrow$

$$detectw_{variant} \circ decrypt_{variant,typeVar} \circ$$
$$\pi_{(variant,typeVar)} \circ$$
$$\sigma_{((subjectId \in mdd) \wedge (position=i, position=j,..))} \circ$$
$$crypt_{variant,typeVar} \circ wat_{variant}$$

(b) distributed query

Conclusion

- **Requirements** for distributed analyses over shared genetic data
- **COSHED** approach
  - **Secure complex workflows** of biomedical analyses using multiple security/privacy enforcement mechanisms
- Future work
  - **Java libraries** for shared genetic data and distributed analyses
  - **Automatic property verification** using ProVerif