IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

Secure distributed workflows for biomedical data analytics

Laboratory: DAPI/LS2N

Start: 1/9/2018

Financing: Half scholarship DAPI/IMT Atlantique

Cofinancing: Half scholarship Colombian School of Engineering Julio Garavito, co-supervision with Prof. Luis Daniel Benavides (in the context of a starting institutional partnership with IMTA at the doctoral level)

Supervision:

- Mario SÜDHOLT, DAPI, IMT Atlantique
- Luis Daniel BENAVIDES, Colombian School of Engineering Julio Garavito

English keywords: Distributed medical workflows, security and privacy, data analytics, medical genetic and imaging data, distributed algorithms

Context

The development of Precision Medicine is strongly supported by the availability of large, comprehensive, population-wide real-life biomedical datasets. However, assembling and sharing such datasets is generally challenging at both the societal and technical levels. Personal data protection policies, such as the EU General Data Protection Regulation (GDPR), provide guarantees from a citizen perspective, but make data exchanges challenging when setting-up multi-disciplinary scientific collaborations (involving clinicians, life and data scientists). Other challenges arise when considering multi-site collaborations. The volume of produced genomic or imaging data makes it hardly relocatable (network cost, duplicated storage), thus requiring the reproduction of data analyses in multiple infrastructures. In addition, researchers or clinicians may not be able to locally access high performance computing infrastructures, thus requiring moving less sensitive data close to computing infrastructures. [1] It becomes crucial to develop novel approaches to share pseudo-individual data on secured distributed computing infrastructures while guaranteeing data usage policies.

The thesis will be performed in cooperation with Mario Südholt's existing projects on the securization of biomedical workflows (PrivGen [2], Oncoshare and Lama projects) with (CS and medical) researchers from PdL and Brittany regions.

Objectives

In the context of this thesis we aim at designing and implementing a secure and privacy-preserving infrastructure for data analytics for large-scale distributed medical analyses.

Concretely, we will develop a secure and privacy-preserving model and distributed infrastructure for biomedical data analyses. This model and corresponding prototype infrastructure will target three main challenges that are crucial for distributed biomedical analyses.

- Protect data throughout complete workflows (storage and computations) through a specific notion of secure and privacy-aware container, *called biomed software containers* that safely encapsulate data and trusted/certified computations on such data. Through these containers data may be manipulated using the encapsulated computations at arbitrary sites under full control of the owner of the data.
- Because of the value of genomic data and data resulting from biomedical analyses, it is of utmost interest to be able to identify the provenance and ownership of such data throughout its lifecycle. We will integrate traceability functionality to the biomed container.

These objectives will be explored based on previous results by the co- directors, notably a compositional method for securing distributed workflows [3], medical information systems [4] and distributed event systems [5].

Required competences

Profound knowledge and proficiency in the practicalities of the development and execution of distributed work flows and data flow is mandatory. Detailed knowledge in Cloud computing, security and privacy issues is also necessary.

Some knowledge and proficiency in the following domains constitute and advantage:

- Biomedical analyses

References

[1] Jake Luo, Min Wu, Deepika Gopukumar, and Yiqing Zhao**: Big Data Application in Biomedical Research and Health Care: A Literature Review.** In: *Biomedical Informatics Insights,* Jan. 2016.
https://doi.org/10.4137/BII.S31559

[2] Fatima-Zahra Boujdad, Mario Südholt: **Constructive Privacy for Shared Genetic Data.** *Proceedings of the 8th International Conference on Cloud Computing and Services Science (CLOSER 2018)*, pp.1-8, Mar 2018. https://hal.archives-ouvertes.fr/hal-01692620

[3] R.-A. Cherrueau, R. Douence, and M. Südholt. **A Language for the Composition of Privacy-Enforcement Techniques**. In *IEEE RATSP 2015, The 2015 IEEE International Symposium on Recent Advances of Trust, Security and Privacy in Computing and Communications*, Helsinki, Finland, August 2015.
https://hal.inria.fr/hal-01168303

[4] Ismael Mejía, Mario Südholt, and Luis Daniel Benavides Navarro. **A study of invasive composition for the evolution of a health information system**. In *Proceedings of the 2nd international workshop on Variability and composition*, VariComp '11, pages 7-11, 2011. ACM. http://doi.acm.org/10.1145/1961359.1961362

[5] Luis Daniel Benavides Navarro, Andrés Barrera, Kiyoshige Garcés, and Hugo Arboleda. **Detecting and coordinating complex patterns of distributed events with Ketal**. In.: *Electronic Notes in Theoretical Computer Science*, 281(0):127 - 141, 2011.
http://dx.doi.org/10.1016/j.entcs.2011.11.030