PhD position (M/F) – Thesis offer
(Reference: DS_PP_PhD_InfClean_062019)

**Research topics**  Data management, Data quality

**Department**  Data science

**Publication date**  11 June 2019

**Start date**  September 2019

**Duration**  Duration of the thesis

**Description**  As collecting data becomes easier, the number of data-driven applications increases exponentially. However, in order to unveil useful information, data must be prepared for analysis to avoid errors due to poor data quality.

Data cleaning is the process of identifying and fixing issues in the data, such as errors and duplicates. Unfortunately, cleaning is a human-centric process that hugely impacts the overall cost for data driven applications, taking up to 80% of the analyst's time. One of the main problems in data cleaning is that domain experts must be taught the languages to express their cleaning programs in a top-down fashion, a difficult and tedious user process.

Our first objective is to remove this technological barrier and let users express cleaning programs by data examples. The second objective is to design new algorithms for this user interaction that radically reduce the human effort in preparing data.

This PhD program will address the following topics:

1. Explore the connections between complexity and feedback in the inference of cleaning programs from data annotations. For example, the user annotates values in a noisy dataset with labels such as "valid"/"invalid". A formal framework is needed to identify the complexity of inferring cleaning programs from user annotations. Such a study is key to enable a new bottom-up data cleaning interface based exclusively over the data, which is what the domain experts immediately understand.

2. New algorithms that interactively define cleaning programs with a "data only" interface will reduce the IT barrier, but inevitably lead to a large number of data interactions. The goal is to properly fuse the requirement of using data annotations with several key ideas to reduce the human effort in cleaning.
   1. **Enrichment:** Existing evidence, such as past user updates, anomalies in the data, and correlations with other datasets, should be used to focus the users towards the annotations that are likely to clean large amount of errors.
   2. **Evolution:** Interactive data cleaning inevitably brings the need of re-evaluating previous decisions in the construction of the system. The same principle applies when new data is inserted. The system should evolve until matching the properties of the data and the users' interactions.
   3. **Crowdsourcing:** With the data-centric interface, a larger number of domain experts can work on cleaning. This crowd (internal, or external with services such as Amazon Mechanical Turk) should be leveraged to scale the cleaning of large datasets with guarantees in the trade-off between the quality of the answers and the cost of the interactions.

3. Implement and evaluate a data cleaning tool to support real use cases provided by industrial collaborations.

**Requirements**
- Education Level / Degree: MSc (with excellent grades)
- Field / specialty: Computer Science
- Technologies: Data Management / Excellent programming skills
- Languages : English
- Strong mathematical and algorithmic background
- Solid motivation and personal preparation

EURECOM specifically encourages women to apply with a view towards increasing the proportion of female researchers.
What we offer

- Strong international research environment with supervision from experienced faculty.
- Opportunities to collaborate and for research stays with renowned collaborators worldwide.
- Well-paid PhD position, in an area known for its beauty and the mild weather.
- English used for research and communication but French will be a plus.

Application

The position is available immediately. The cutoff date for the application is August 15, 2019, but the application evaluation will start immediately, so early applications are encouraged. Interested individuals should submit the following documents (in English):

- I-Curriculum Vitae, including your current contact address, transcript of certificates and grades (with a list of university courses taken), and your previous publications (if any)
- II-Motivation letter including research and education perspectives
- III-Contact information for 2 referees at your current and/or previous affiliations

Applications should be submitted by e-mail to secretariat@eurecom.fr with the reference: DS_PP_PhD_InfClean_062019

Postal address
CS 50193 - 06904 Sophia Antipolis, France

Contact
secretariat@eurecom.fr

EURECOM is a French graduate school and a research center in digital sciences based in the international science park of Sophia Antipolis, which brings together renowned universities such as Télécom ParisTech, Aalto University (Helsinki), Politecnico di Torino, Technische Universität München (TUM), Norwegian University of Science and Technology (NTNU), Chalmers University (Sweden) and Czech Technical University in Prague (CTU). The Principality of Monaco is a new institutional member. The Institut Mines-Télécom is EURECOM's founding member.

EURECOM benefits from a strong interaction with the industry through its specific administrative structure: Economic Interest Group (kind of consortium), which brings together international companies such as: Orange, BMW Group Research & Technology, Symantec, Monaco Telecom, SAP, IABG.

EURECOM deploys its expertise around three major fields: Digital Security, Data Science and Communication Systems. EURECOM is particularly active in research in its areas of excellence while also training a large number of doctoral candidates. Its contractual research is recognized across Europe and contributes largely to its budget.

Thanks to its strong ties set up with the industry, EURECOM was awarded the "Institut Carnot" label jointly with the Institut Telecom right from 2006. The Carnot Label was designed to develop and professionalize cooperative research. It encourages the realization of research projects in public research centers that work together with socioeconomic actors, especially companies.