Software Engineer Intern - Interactive Data Preparation

PARIS, FRANCE . ENGINEERING – INTERNSHIP

Dataiku's mission is big: to enable all people throughout companies around the world to use data by removing friction surrounding data access, cleaning, modeling, deployment, and more. But it's not just about technology and processes; at Dataiku, we also believe that people (including our people!) are a critical piece of the equation.

Today Dataiku Data Science Studio allows users to analyze & transform large data sets using a visual preparation tool. Users can operate on a column and apply a large set of rules (extract dates, fill empty cells with value, round numbers, find & replace, tokenize text, parse URLs, …)

We are looking for a **software engineer intern** for a 4 to 6 months internship in our Paris office. The objective of the internship is to improve this preparation tool to generate new rules based on user interaction.

For example, if users interactively select the following text:

Name	Birth Date	Comment
Marie Curie	1867	Born in Poland
Stephen Hawking	1942	Born in Oxford, United Kingdom
Rachel Carson	1907	Born in Springdale, USA
Nikola Tesla	1856	Born in Smiljan, Croatia

The improved preparation tool would then propose to users possible actions. For example deleting/keeping all rows similar to the selected ones, … or extracting from all rows the words matching the selection into a new column as shown below:

Name	Birth Date	Birth City	Comment
Marie Curie	1867		Born in Poland
Stephen Hawking	1942	Oxford	Born in Oxford, United Kingdom
Rachel Carson	1907	Springdale	Born in Springdale, USA
Nikola Tesla	1856	Smiljan	Born in Smiljan, Croatia

Note that the tool would be smart enough to figure out that Poland is not a City, while Smiljan is.

Of course, this is just an example. What could be achieved during the internship will only be limited by your imagination!

During this internship, you will:
- **Get familiar with Java, Javascript and Angular** if you don't know them yet.
- **Study research papers** on how to generate regular expressions and other ways to match and extract data from text. Here are some references:
  - Interactive Visual Specification of Data Transformation Scripts
  - HoloClean: Holistic Data Repairs with Probabilistic Inference
  - Auto-Data Cleaning (Columbia Research Group)
- **Implement a working prototype** for a subset of the problem, including user interaction
- **Celebrate and party** because now Data Analysts can now prepare data much faster!

You are the ideal recruit if you:
- are **autonomous**, to drive your subject.
- have **no fear** to dive deep into source code.
- are **eager to learn** new things.
- have a **good knowledge of a programming language** (Python, Java, C#, Javascript, R, Ruby, <insert your favorite language here>).
- have a basic knowledge of **Web development**

Benefits:

As an intern, you'll join the **Engineering team of a startup**, composed of 40+ engineers passionate about software development and developing an amazing product. Yes, we love our product. No, we are not biased ;-).

We have **great office space in Paris** (Gare de Lyon) with free breakfast, coffee, video games, yoga and much more.

The monthly remuneration will be:

- 1000 euros for students in 4th year of university / engineering school.
- 1400 euros for students in 5th years of university / engineering school.

Remuneration additionally includes 50% reimbursement of public transit pass.

A full-time contract position might be offered at the end of the internship.