# Logos/cis-logo-exports-final-8-2015-3/Web/2_Department_Lockups/InformationScience/cis-infosci-2-color.pngMaster of Professional Studies in Information Science

# Sponsored Project Proposal Form – Spring 2018

Please complete the following project proposal form to sponsor an MPS Project. This form will be used to determine if your project is appropriate for MPS students and whether it is of sufficient scope for a semester long project (~400-500 person-hours). We will assign teams with complementary skills based on the skills and experience you list in this form. We will also share most of this form with the students to help them make their top project choices before we assign the projects.

Please direct any questions to the MPS Project Coordinator: [is-mps-projects@cornell.edu](mailto:is-mps-projects@cornell.edu)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Sponsor Name | | Credit Suisse | | | | | | Date | 12/4/17 |
| Contact Name(s) | | Wendy Walasek  Jim Tanner | | | Email(s) | [Wendy.Walasek@credit-suisse.com](mailto:Wendy.Walasek@credit-suisse.com)  [Jim.Tanner@credit-suisse.com](mailto:Jim.Tanner@credit-suisse.com) | | Phone | +1 212 325 4987  +44 20 7883 5166 |
| Description of the Sponsor | | | | | | | | | |
| As one of the world's leading banks, Credit Suisse provides its clients with investment banking, private banking and asset management services worldwide. Founded in 1856, Credit Suisse has a long tradition of meeting the complex financial needs of a wide range of clients, offering advisory services, comprehensive solutions and innovative products to companies, institutional clients and high-net-worth private clients globally, as well as retail clients in Switzerland. Credit Suisse is active in over 50 countries and employs approximately 45,000 people. Further information can be found at www.credit-suisse.com.  Cultural diversity is essential to our success. As such, we employ people from more than 100 countries. Credit Suisse empowers employees to work openly and respectfully with each other and with clients, ultimately striving to deliver superior results while offering initiatives and programs to assist employees achieve a healthy work-life balance. | | | | | | | | | |
| Please indicate which academic year and semester you would like to propose your project. | | | | | | | | | |
| Year | 2018 | | Semester | Fall | | Spring |  | | |
| Project Title | | | | | | | | | |
| Data Quality: Rules vs Machine Learning Systems | | | | | | | | | |
| Project Goal or Description | | | | | | | | | |
| Rationalize the use of rule based data quality verification systems verses machine learning type systems. We would like the project team to explore benefits and shortfalls of each system type. This will give students a wide exposure to standard industry tools used for verifying and reporting data quality.  Credit Suisse’s Chief Data Office has supported the purchase of data quality verification systems to support the goals of having high quality data for our operational, risk management and regulatory reporting. We are often asked why we are not supporting products that leverage machine learning. We have periodically looked at machine learning algorithms and have to date chosen products that are more mature and have a larger user base. We believe that the machine learning tooling space is advancing and it is worth looking again at how the data quality tools are maturing especially in relation to leveraging machine learning to identify data issues.  The student team will analyze data from both rule-based and machine learning based approaches to understand the types of data quality issues each tool is best at identifying leveraging sample data and the data quality types informed by Dr. Richard Y. Wang’s work on [Data Quality](http://web.mit.edu/tdqm/www/publications.shtml). The team should look to quantify the differences through testing with sample data that demonstrates the strengths and weaknesses of each method. | | | | | | | | | |
| What activities are necessary to achieve the project goal? | | | | | | | | | |
| * Investigate rules based and Machine learning data quality tools to understand how they are bested used and their pros and cons * Install potential systems with temporary licenses and understand how they are used and for what kinds of data and data issues: structure numeric, structured categorical, continuous, text blobs, etc. * Investigate whether or not Ontologies can help improve the rules and create a new rule type for verifying data quality * Consider the full lifecycle to evaluate if machine learning and ontologies reduce the work required to identify types of data quality issues * Write up findings and present recommendations against use cases explored | | | | | | | | | |

|  |
| --- |
| What outcome would determine that the project is a success? Do you expect specific deliverables? |
| Presentation of systems investigated, pros and cons of rules based systems verses ML systems. Recommendation of data uses cases that best leverage each system type. Systems that should be minimally considered are Informatica Data Quality, Python rules (similar to Palantir data quality checker), plain SQL. Other systems to consider are Taland which offers an ML system. Adding to the system option list would be part of the project to understand the market place. |
| What are the skills and experience must the students already know to start work on the project?  Please be specific and keep in mind that students will be building their skills during the duration of the project. |
| Analytic skills, communication skills, logical reasoning, SQL, structured and unstructured data sets understanding and basic understanding of data quality issues. The team may find Dr. Richard Y. Wang’s work on [Data Quality](http://web.mit.edu/tdqm/www/publications.shtml) enlightening to the subject area and their efforts on this project. |
| What are the skills and experience required to complete the project that the students may learn while completing the project? |
| The project team would expect to learn about data quality and how data quality is evaluated to ensure that data used in analytics and machine learning projects have the quality required to build models and use for other risk and business management decisions. |
| The project representative must be available 30 minutes per week for status reports, the interim report, and the final presentation. As the project sponsor, are you able to make this time commitment?  Yes. Please elaborate. |
| We will reserve time daily to engage for the project during the project timeframe. |
| Some sponsors may choose to spend additional time with the student teams, e.g. phone contacts for monthly status discussions, reviewing research results, providing midpoint project feedback, and offering input to the final deliverables in advance of its completion. As the project sponsor, are you available to participate in these or any additional activities?  Yes. Please elaborate. |
| Yes, I would expect the team to meet for weekly meetings while they are working on the project and we would be available for questions daily. |
| The project representative needs to facilitate access to company resources as needed and approve expenses. As the project sponsor, are you able to facilitate access to such resources, should the need come up?  Yes. Please elaborate. |
| We are going to grant the student team individual access to the firm’s systems and give them entitlement to run systems on firms development hardware. |
| Please consider other contributions listed below. Are you willing to make these contributions? (check all that apply)  Provide existing industry and company data as background at the beginning of the project.  Pay one or more team members to travel to your location for initial briefing / work session / final presentation.  Please elaborate. |
| My understanding is the teams are in NY. If they would find it helpful we can meet at our offices in Manhattan or I could go to the school to meet in person. Otherwise we will be able to meet via Skype calls and share materials as necessary. |
| Please send your completed project proposal to the MPS Project Coordinator: [is-mps-projects@cornell.edu](mailto:is-mps-projects@cornell.edu) |