Dennis Sun **DATA 401:** Data Science Alexander Dekhtyar

Requirements Elicitation

Requirements Elicitation is the first step of any Data Science acitvity.

In this respect, Data Science is similar to other types of software development. In general, the same rules that guide requirements elicitation in Software Engineering apply for Data Science.

However, there are specific things one has to pay attention to, when performing requirements elicitation for Data Science projects.

Requirements Elicitation for Data Science

Requirements Elicitation requires a data scientist to communicate with a customer.

A typical result of a requirements elicitation process in Software Engineering is an SRS (System Requirements Specification) document.

For Data Science projects, there may not be an explicit need for a formal written SRS.

However, it may be benefitial to record observations made during the process, if without the level of formality a typical SRS commands[1].

Customer

A **customer** for a Data Science project is a person, or a group of people who initiate the project and either directly or indirectly sponsor it. The **customer** may own the data that needs to be analyzed, but it is also possible that the customer may not actually have any data prior to the start of the project.

There are different categories of customers. Typical customers are business people, scientists, journalists, small business owners, administrative personnel.

Occasionally, the customers are computer scientists or other professionals with siginifacnt computing experience.

Starting the Requirements Elicitation Process

There are two possible starting points for the requirements elicitation process:

[1] The question of what documentation shall be kept during the data science projects is an important one, and we will discuss it later in this course.

1. Customer has a formal/informal Requirements Specification prepared. In this case, the requirements elicitation process consists of
    - Study of the prepared requirements specification.
    - Clarification of any underspecified information.
    - Extension of the requirements specification where it is incomplete.
2. Customer does not have a prepared Requirements Specification. In this case, the data science team must engage the customer in a full requirements elicitation process.

Elicitation Process

During the requirements elicitation process, the data scientists must determine answers to the three main sets of questions:

1. **Customer and domain questions:** Who is the customer, what is the underlying domain of study, and how much information about the underlying domain the team must learn/understand?
2. **Data Questions:** What is the data and who has it?
3. **Analysis Questions:** What are the analytical questions the customer wants answered, what insight the customer expects to see from the data, and how this insight needs to be conveyed back to the customer?

Customer and Domain Questions

Here are some of the things you need to establish about the customer and the domain in which the customer operates.

- Who is the customer?
- What is the customer's domain?
- Is the customer an expert in the domain in question? If not, who is a domain expert the data science team can talk to?
- How much domain expertise knowledge is the data science team expected to have/acquire?

Data Questions

Here are some of the questions related to the data that needs to be used for analysis.

- What data needs to be studied?
- How is this data available?
- Is all data available from the customer?
- Is there some data that needs to be collected as part of the analysis *that the customer knows about*?
- Are there any "unknown unknowns"? (i.e., incompleteness in the data that the customer doess not immeidately know how to compensate for).
- Who knows the structure of customer's data?
- How can information about customer's data be obtained?

- What information is important? What information is not important?
- What is the expected total amount of data that needs to be analyzed?
- Is the data static or is it dynamic/streaming?
- How is the available information stored?
- What are the sources of supplemental information?

Analysis Questions

In Software Engineering, these are the actual functional requirements for the software. To elicit these, the data science team must find out answers to the following questions:

- What are the key questions the customer wants answered?
- What is the key insight the customer wants obtained?
- What types of actions the customer wants to perform based on the obtained insight?
- Who are the recepients of the analytical results?
- In what form are the results expected to be presented?
- Who are the final decision-makers?
- How important is predictive accuracy? Is there a lower bound on predictive accuracy that the customer considers to be sufficient?
- How important is system performance? Is there a system performance lower bound that the customer considers to be sufficient/acceptable?

Eliciting Requirements

There are a few ways in which information about the customer requirements can be conveyed between the customer and the data science team. The core ways are:

1. **Special Purpose Documentation.** Any formal or informal documents customer has prepared *specially* for the data science team prior to engagement, or in the very early stages of engagement.
2. **Existing Documentation.** Any existing documentation describing the customer's data, process, needs, etc, that has been created prior to the customer's engagement with the data science team, and which had an original purpose for creation unrelated to the proposed data science activity (i.e., documentation generated as a byproduct of the customer's usual activities).
3. **Customer Interviews.** Conversations the data science team has with the customer.
4. **Written correspondence with the customer.** Typically the combined collection of email exchanges (also any other written discussion transcripts).
5. **Notes.** Informal documents developed by the data science team (and also possibly the customer) as the result of interactions.
6. **Formal specifications.** Formal documents certifying the data science team's understanding of the tasks at hand.

7. **Thrid-party documentation.** Existing information about data sources (and other possible issues) that does not come directly from the customer.

8. **Third-party interviews/correspondence.** Any information exchanges between the data science team and any third parties (e.g., owners of external data sources) related to the proposed analytical tasks.