1Product Requirements and Design - Version 2 (PRDv2)

Constructing Intelligence

Procore Technologies, Inc.

The Team
Team Lead: Vivek Patel
Team Scribe: Nicole Moghaddas
Deanna Pham
Charlie Getzen
Lauren Dumapias

1. Introduction

Define problem, innovation, science, core technical advance (2-3 pages)
Define project specifics, team goals/objectives, background, and assumptions

The problem and importance of this project stems from unsafety in the workplace, specifically on construction sites. A significant amount of on the job accidents are construction related; understanding, identifying, and resolving the underlying causes will be extremely beneficial to increasing the safety of these projects. The goal in this project is to use predictive analytics to identify unsafe conditions before they actually cause accidents on the job. The goal is to create a predictive metric for the construction industry, that will be useful and critical for safety measure on projects.

Furthermore, extending beyond project safety, it would be ideal to also predict metrics on project performance and quality. A reach goal of this project is to predict additional project components, such as schedule, budget, and quality, answering questions as to whether the project will be on schedule or whether the project will be over or under budget. This would be a very helpful tool for companies to be able to identify certain problem aspects in their projects, and take action before any damage is done. A continuously updating prediction of these project characteristics would be very advantageous to the overall project performance.

The science behind this project involves big data and machine learning. Procore currently provides historical analytics in retrospect about each construction project, and would like to determine the feasibility of training machine learning models on its data set so that they can also provide some predictive metrics. This project will demonstrate the feasibility of doing so, as well as identifying additional opportunities for use of machine

learning. The project will delve into applying machine learning with data science to Procore's large database of project logs and project data, and explore what can be predicted from this information, what project attributes can be predicted from previous project data.

This is a critical advance in project planning and management tools, such as Procore's project tool for the construction industry. Instead of solely being able to use the product for documentation and storing project aspects, such as daily logs, project schedules, and project financials, this project will introduce novel and innovative perspectives for companies to analyze their current construction projects. In addition to being able to manage the cumulative progress and assess the current state of projects based on past data, this project will redefine how construction companies can manage and assess their projects. If they not only have information about the current state of their project, but also predictive qualities about project characteristics, they can use this information to make safer, smarter, and more efficient choices in regard to project management.

In this project, the goal is to create a predictive metric for the construction industry. The analysis will focus on Procore's quality and safety tools in order to foresee future hazards on project sites. The most important goal will be to provide a "Risk Score" in order to identify high risk projects and to help prevent accidents (perhaps sending alerts if the score goes above a certain level), allowing construction companies to take immediate action. The "Risk Score" would entail something such as a percentage that describes the safety level of the project. The reach goal would be to predict additional predictive metrics, such as actual likely budget and schedules of projects, and the overall quality of projects, updating continuously as the project advances, or perhaps just whether or not the project will be over budget, under budget, on or off schedule. This will be accomplished by using the extensive data that Procore has from past projects to build machine learning models to predict these characteristics for new projects.

The project data we will be given will include written documentation, such as project logs, project conditions, project specifications, project teams, etc., and project images, such as blueprints, floor plans, etc. We will analyze this data, and extract what features will be useful in predicting the properties we want to forecast. Furthermore, we will need to determine what it is feasible to predict, what is possible to determine from the data. We will then build models, obtain parameters, and use these models for future data, to predict project attributes. Understanding what features will be useful and obtaining the correct models for the different kinds of data will be the challenge of this project.

2. System architecture overview

See attached diagrams for High Level Diagrams.

Adding a whole user interface interaction and design was not part of the project specification from Procore, as it was meant to be more of a back end project; however, we are working with an iPython Notebook for our user interaction and visuals of the data and the outputs of our predictive models. The notebooks help us graph and visualize our input data and results so that have a thorough understanding of the predictive analysis we are doing.

3. Requirements (functional and non-functional)

Use cases:

1) Use case: Predicting Schedule Timeline
Actors: Product manager, Procore application
Precondition: Project has been going on for an x amount of time.
Flow of Events:
Basic Path:

1. Product manager checks the application to get up-to-date information about the project's schedule
2. System will give the product manager an accurate representation of how far along the project is and if the project itself is delayed or not.

Alternative Paths:

If the product manager doesn't check the application, they will not have a good idea on how far along the project is and that may cause them to make an uninformed decision.

Postcondition: Product manager is now able to make informed decisions based on the data that was presented

2) Use case: Employee on site
Actors: Employee
Precondition: Employee is unaware of project conditions
Flow of Events:
Basic Path:

1. Employee arrives on site, not up-to-date since the x amount of time he has been off site
2. System will send a daily alert of an overview of the project as of that moment
3. Employee will be able to prioritize certain things that need to be done and if it is safe to be on site

Alternative Paths:

If the employee was not informed about the project, they would not be able to make the best decisions for the project from that point forward or

Postcondition: The correct decisions were made in scheduling, budgeting, and other important aspects of the project.

3) Use case: Instant inspections update
Actors: Inspector
Precondition: Unsafe working conditions not updated in daily inspection log
Flow of Events:
Basic Path:

1. Inspector inspects project site and notices many hazardous details
2. Inspector checks application and notices it is not accurately portrayed
3. Inspector is able to quickly add data
4. Immediately, system is able to accurately update the application

Alternative Paths:

If new data isn't added regularly, the application predictions could become inaccurate.

Postcondition: Everyone has access to a more updated application

4) Use case: General contractor supplies equipment
Actors: General contractor
Precondition: Project has been purchasing equipment without noticing the budget
Flow of Events:
Basic Path:

1. Employees under the general contractor have been making purchases as told without consulting the application
2. General contractor analyzes data given from application and realize the project has gone over budget

Alternative Paths:

General contractor never consults the application and the project becomes more expensive than the quota given at the beginning of the project

Postcondition: General contractor and their employees are now more aware of this when they purchase necessary materials

5) Use case: Updating data with daily alerts
Actors: Employees
Precondition: Data is not as accurate as it could be
Flow of Events:
Basic Path:

1. System alerts employees to update daily logs
2. Employee fills out daily logs and other features that will make predicting more precise daily
3. Application has the most up to date data to run on

Alternative Paths:

System does not alert employees and the project analysis is not accurate.

Postcondition: All employees have access to the most recent analyzation and will be able to make better decisions based on that

User stories (with links) in order of priority:

1. As a construction company, I would like to be able to know if my project is likely to be on schedule or not.
https://github.com/procore/capstone2016/blob/bebb96212222512540545b34f81cfa7b38119149/SubmittalsNotebook.ipynb
Story points: 10
Acceptance Test: The model will run on unseen data for that company's project and should output if the project is late or not. If the output is present, this is a passing test (aiming for an accurate prediction).

2. As a developer, I want to add attributes for project submittals so that I can use them to predict whether a project is likely to be on time or not
https://github.com/procore/capstone2016/blob/125c7e182ec8ecd4caf25befcfa871f22f524284/Titanic.ipynb
Story points: 5
Acceptance Test: The model input dataframe should have additional features (attributes/columns) that will be used in the model to make the target prediction.

3. As a developer, I would like to know what features in the data will map to useful outputs to predict whether a project will be late or not
https://github.com/procore/capstone2016/blob/2188d1fe37566caf89902a48d0f543fbdf19a1e0/Titanic.ipynb
Story points: 5
Acceptance Test: Graphs, charts, visuals, analysis should show which features (columns) in the data will be used towards a model.

4. As a developer, I would like to be able to create and edit data in tables (ie pandas dataframes) easily so that I can analyze the data and create predictive models
https://github.com/procore/capstone2016/blob/master/data.py
Story points: 3
Acceptance Test: The methods that edit the dataframes pass their tests. For example, the createTable() method should successfully produce a dataframe with the requested data, the remove_nan() (not a number) operation should remove the necessary values from the dataframe, etc.

5. As a developer, I would like to plot features to see if they contribute to a project being late or not so I can know which features to use to create ML models
https://github.com/procore/capstone2016/commit/1c9bf9c1cc382a0736c6207e5713649b7bd1f640
Story points: 3

Acceptance Test: Graphs, plots, charts, and such visuals are created that demonstrate the importance and weights of different features in the data to predicting our models.

6. As a developer, I would like to create a class to make adding columns easier.
https://github.com/procore/capstone2016/commit/a03ab4e843414593647c25f7e19f6023cadeabf5
Story points: 2
Acceptance Test: The Table class successfully can add columns to a dataframe with the correct information; the addColumn() method should pass its test.

7. As a developer, I would like to create a class to make training a model easier.
https://github.com/procore/capstone2016/commit/134286b14c3ee8c30947bf13f03dd81a2a4810f0
Story points: 2
Acceptance Test: The Table class's train() method should successfully train a machine learning model that can be used to predict new target variables on new unseen data.

8. As a developer, I would like to add features that relate to possibly predicting a time-frame in which a project is most likely late.
https://github.com/procore/capstone2016/blob/2188d1fe37566caf89902a48d0f543fbdf19a1e0/Titanic.ipynb
https://github.com/procore/capstone2016/blob/12415c5f1203dd9abb6efc3cb918265d84e64325/Titanic.ipynb
Story points: 3
Acceptance Test: Features are analyzed and the ones that aid towards predicting the target variable are used in the model.

9. As a developer, I would like to use k-fold methods to get a better understanding of which training model to use.
https://github.com/procore/capstone2016/blob/c59ce8933c880e0e9271ae7cf4fbe8060385b4b6/kfold/cross-validation.ipynb
Story points: 4
Acceptance Test: K-fold cross validation is run, and the output should chose a model that most accurately predicts the target variable on test data (the best model with the best parameters).

10. As a developer, I want to organize late submittals into bins so that I can obtain a better estimate of a good feature/attribute.
https://github.com/procore/capstone2016/blob/5ea7036c8c918be7d20d859856a6ee9d6eb5ebf7/SubmittalsNotebook.ipynb
Story points: 2
Acceptance Test: Ensure bins were created and data is spread out evenly by displaying bar graph.

11. As a developer, I want to add columns in Budget to get attributes/features to train models.
https://github.com/procore/capstone2016/blob/b4477841dce028514ddff5b8c143b839f8097c48/BudgetNotebook.ipynb
Story points: 2
Acceptance Test: Ensure features are added by checking if columns of information was added in dataframe.

12. As a developer, I want to be able to visualize my data.
*add link once we add some graphs
Story points: 3
Acceptance Test: Ensure visualizations output correct data.

13. As a developer, I want quantize my data.
*add link to quantization
https://github.com/procore/capstone2016/blob/bebb96212222512540545b34f81cfa7b38119149/SubmittalsNotebook.ipynb
Story points: 2
Acceptance Test: Ensure the data is quantized correctly by trial and error.

14. As a developer, I want to make sure I'm not leaking data for security purposes.
https://github.com/procore/capstone2016/commit/5e45b21d188f7d2c86beaef1def5db3aa3c1cb3d
Story points: 2
Acceptance Test: Ensure data is not leaked by throwing exceptions.

15. As a developer, I want to process images to create heat maps so that I can categorize images.
https://github.com/procore/capstone2016/commit/d6ce3994335457d76754263d4b969c420cba25b6
Story points: 7
Acceptance Test: Ensure that the categorization of the image is correct with its given title.

16. As a developer, I want to associate budget with its corresponding cost code so that I can train a model on that specific cost code.
https://github.com/procore/capstone2016/blob/019a61bb2e4fe092755074b413cbce70e45eb09a/BudgetNotebook.ipynb
Story points: 5
Acceptance Test: Ensure associated cost codes are correct with supervised learning (we have the answers already).

17. As a developer, I want to be able to query data to test on.
https://github.com/procore/capstone2016/blob/ecc525193836824f57594c57af8ad40d6b8c877f/mortenson-query.sql
Story points: 1
Acceptance Test: Ensure data is available to use.

18. As a developer, I want to train on multiple models to see which works best.
*add
Story points: 5
Acceptance Test: Ensure the model chosen is the most precise.

19. As a developer, I want to create a simple table to hold data.
https://github.com/procore/capstone2016/commit/a03ab4e843414593647c25f7e19f6023cadeabf5
Story points: 1
Acceptance Test: Ensure data is correctly placed in table by running class.

20. As a developer, I want to know how accurate and precise our prediction is.
Story points: 2
Acceptance Test: Ensure prediction of schedule and budget is accurate using supervised learning.

User stories (still need to be implemented):

1) As a developer, I want to be able to visualize the data features to see what features correlate together so that I can accurately base emphasis on needed areas of data.
2) As a developer, I need good real or simulated data that maps correctly so that I can accurately predict possible outcomes based on features and the map.
3) As a developer, I can predict whether the project is hazardous to the employees or not.
4) As a developer, I can predict on a scale from 1-10 whether the project is safe for a specific employee.
5) As a developer, I can go through millions of data sets and learn with each set to predict whether the project is on schedule, on budget, and safe so that construction companies can be more efficient.

Prototyping Code, Tests, Metrics (User Stories)

1. As a developer, I want to establish a model overview for project budget to outline how predictive analytics can be run on project budget.
   https://github.com/procore/capstone2016/blob/b4477841dce028514ddff5b8c143b839f8097c48/BudgetNotebook.ipynb
   Story points: 4

2. As a developer, I want to establish a model overview for project schedule to outline how predictive analytics can be run on project schedule.
   https://github.com/procore/capstone2016/blob/5ea7036c8c918be7d20d859856a6ee9d6eb5ebf7/SubmittalsNotebook.ipynb
   Story points: 4

3. As a developer, I want to query for project data to train and test our models on.
   https://github.com/procore/capstone2016/commit/a4d74377671a97e50b8d5a0df5f046dfd8d0722e
   Story points: 2

4. As a developer, I want to measure how much data each project/company can contribute to find the most data-rich to test our models on.
   https://github.com/procore/capstone2016/commit/a4d74377671a97e50b8d5a0df5f046dfd8d0722e
   Story points: 1

5. As a product manager, I want to translate the client's necessities into questions to allow developers to know what to provide predictive analytics on.

   Story points: 3

6. As a developer, I want to test that the data has been massaged so the schedule model will be receptive to it.
   https://github.com/procore/capstone2016/commit/584108f8681883bf44357a9da5ff6d6ce8163143
   Story points: 1

7. As a developer, I want to test that the data has been massaged so the budget model will be receptive to it.
   https://github.com/procore/capstone2016/commit/584108f8681883bf44357a9da5ff6d6ce8163143
   Story points: 1

8. As a company user, I want to visualize my data to see which data points I want to have a predictive analysis on.

   Story points: 1

9. As a developer, I want to measure the accuracy of different estimators to select the best one in the given context.

   Story points: 2

10. As a developer, I want to test that my estimator accuracy measurement is correct to ensure that predictions on yet-unseen data will be as accurate as possible.

    Story points: 2

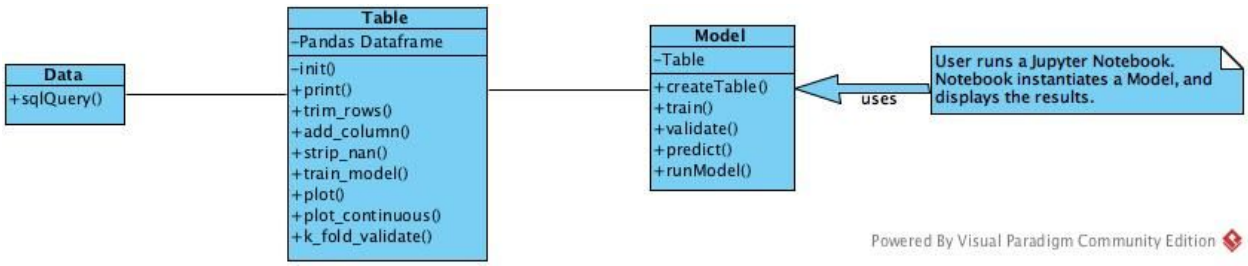
4. System Models

Contexts, sequences, behavioral/UML, state

- *Contexts, interactions, structural, behavioral (UML)*
- *Use cases, sequencing, event response, system state, classes/objects*

The main interaction with the project is the iPython Notebook which will be used to show the predictions/visualizations/outcomes. The user will run the Notebook, with the desired data in the /data folder. The Notebook then runs and outputs the results. The sequence diagram is seen in Figure 1.

The class diagram for components used:



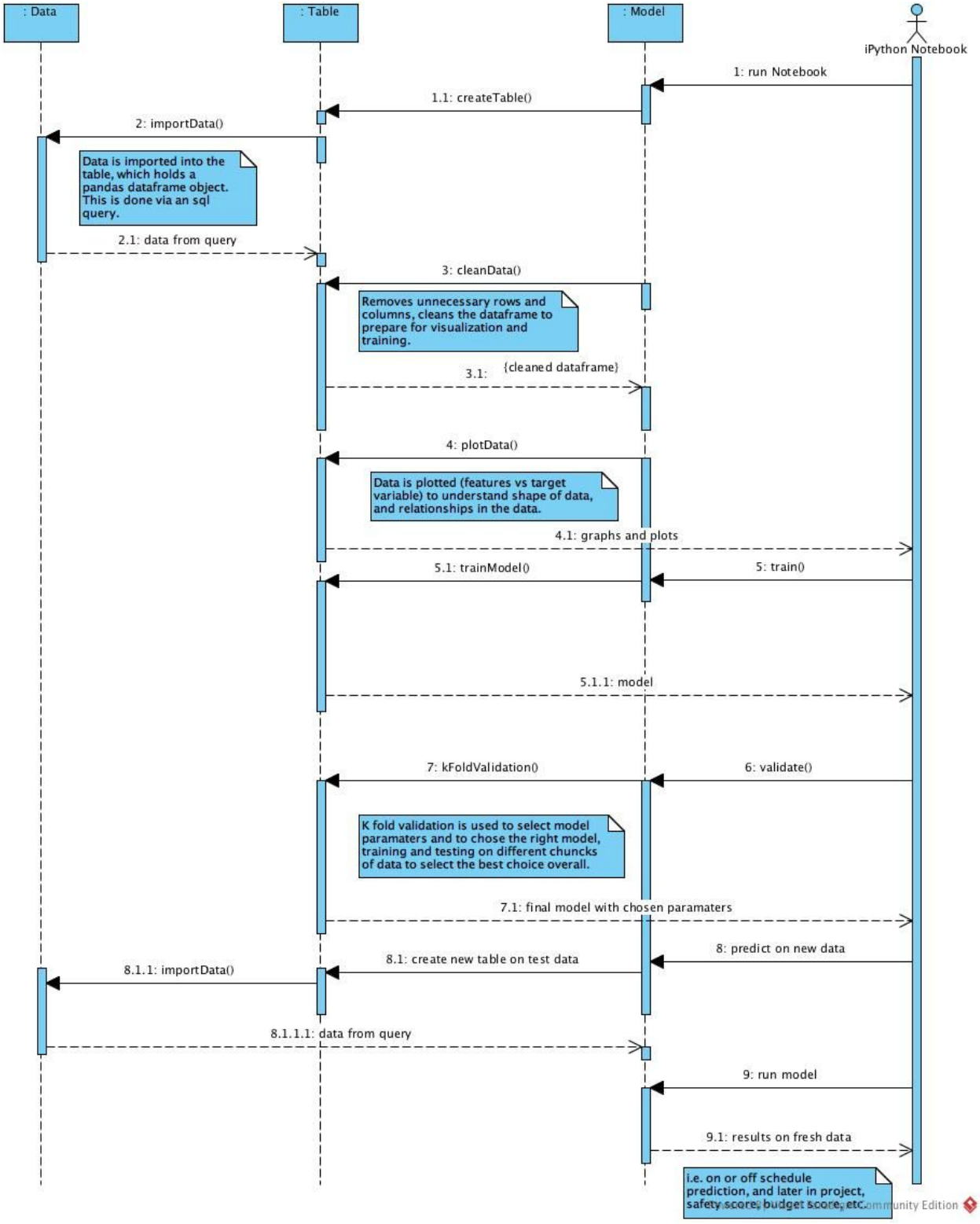


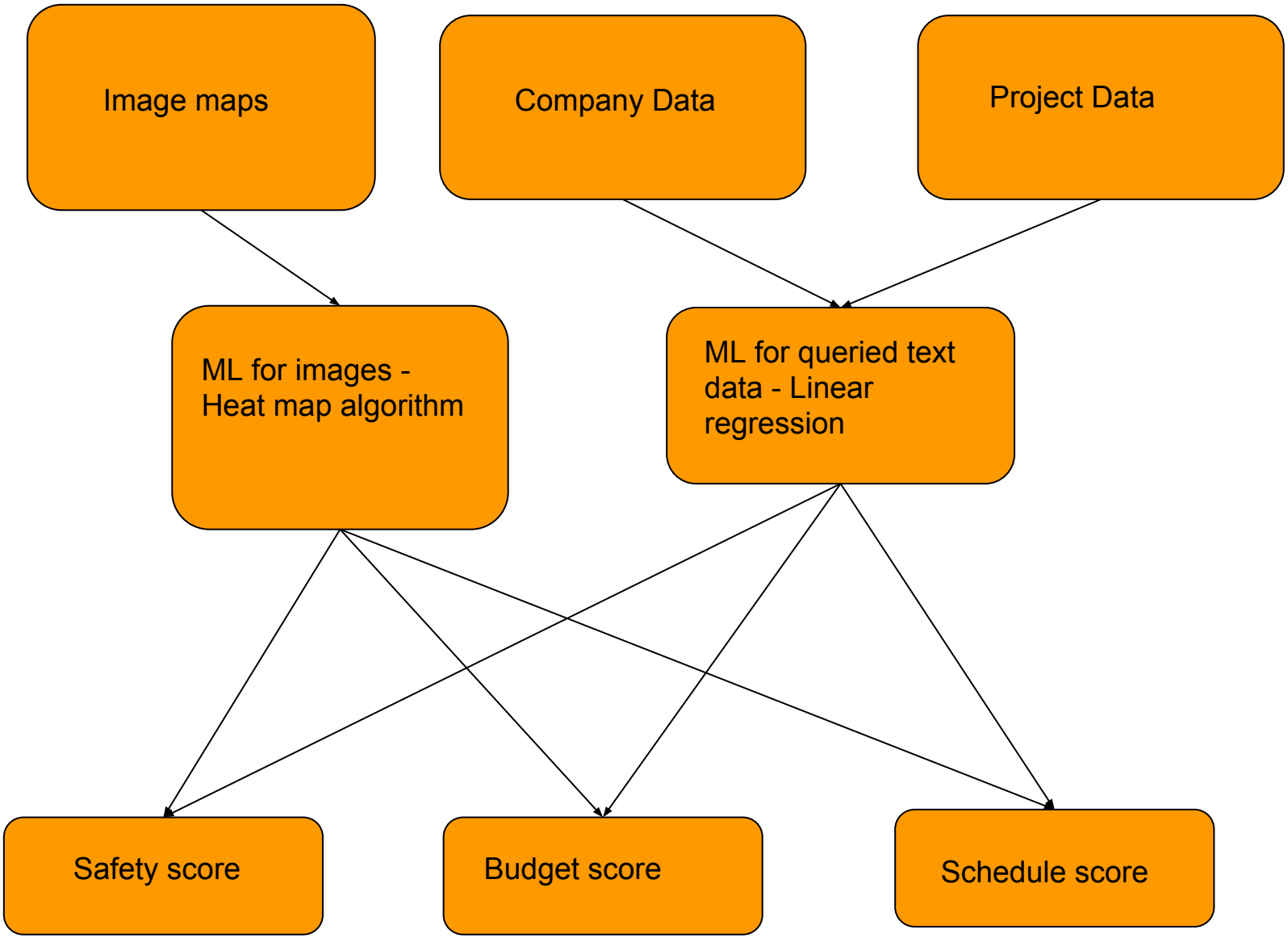
Figure 1: Sequence diagram.

5. Appendices

- Jupyter Notebook: A web application used to create and share “notebooks” containing code that can be compiled and ran by the application.
- Python: Our choice of programming language.
- Trello: A web application used to track our project’s progress with tasks under each ‘stage’ of our progress. Each member owns each card with certain tasks we each need to complete.
- Trello card numbers: Chrome extension that tracks the tasks/issues by number so that we can organize and keep track of the project.
- Burndown chart for Trello: Extension to Trello that takes the cards by number and creates the scrum burndown chart to track the progress of each sprint.
- Google Docs: Web application we use for our documentation.
- Machine Learning: A type of artificial intelligence we are using to train our program to predict many outcomes of a project.
- Python Data Science Libraries: Pandas, SciKit Learn, Numpy. These libraries will allow us to access their functions aiding us in our data analysis.
- Github: Our repository hosting service and our issue tracker.

What is defined in this doc (from slides):
1. Define project specifics 2. Team goals and objectives 3. Background and strategic fit 4. Assumptions 5. User Stories or Use Cases 6. User Interaction and Design 7. Questions 8. What we’re NOT Doing • Evolve the document over time, concurrently with development

2. System Architecture Overview



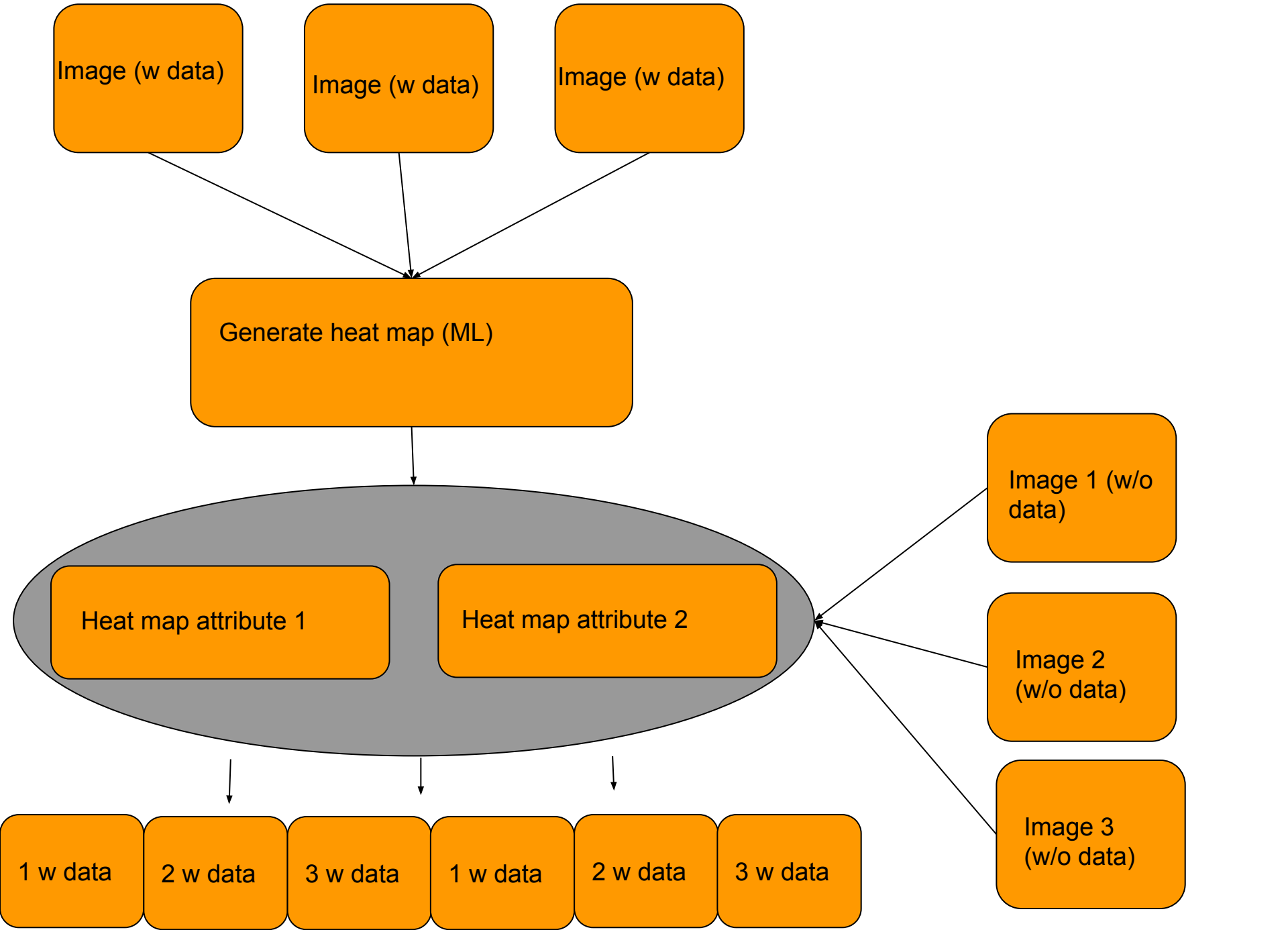


Image (w data)
Image (w data)
Image (w data)
Generate heat map (ML)
Heat map attribute 1
Heat map attribute 2
Image 1 (w/o data)
Image 2 (w/o data)
Image 3 (w/o data)
1 w data
2 w data
3 w data
1 w data
2 w data
3 w data