<table>
<thead>
<tr>
<th>Team Name:</th>
<th>7a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team Members and</td>
<td>Kate Strombom (<a href="mailto:kstrombom@ku.edu">kstrombom@ku.edu</a>)</td>
</tr>
<tr>
<td>email addresses:</td>
<td>Julia Drahozal (<a href="mailto:j557d600@ku.edu">j557d600@ku.edu</a>)</td>
</tr>
<tr>
<td></td>
<td>Andrew Megaris (<a href="mailto:andrew.megaris@gmail.com">andrew.megaris@gmail.com</a>)</td>
</tr>
<tr>
<td></td>
<td>Adam Heald (<a href="mailto:adamheald13@gmail.com">adamheald13@gmail.com</a>)</td>
</tr>
<tr>
<td></td>
<td>Quincy Wofford (<a href="mailto:quincy.wofford@gmail.com">quincy.wofford@gmail.com</a>)</td>
</tr>
<tr>
<td></td>
<td>Morgan Metzger (<a href="mailto:the12stringer@gmail.com">the12stringer@gmail.com</a>)</td>
</tr>
<tr>
<td>Team Meeting time:</td>
<td>Wed 4pm</td>
</tr>
<tr>
<td>Lab Meeting time:</td>
<td>Wed 4:15 or Wed 2:30p</td>
</tr>
<tr>
<td>Contact:</td>
<td>Quincy Wofford</td>
</tr>
<tr>
<td>Project Sponsor:</td>
<td>Black and Veach</td>
</tr>
</tbody>
</table>
| Project Description (150-250 words) | DataDriven is a data driven business intelligence tool served from a cloud based web app. Our Minimum Viable Product (MVP) is focused to take input data from our sponsors at Black and Veach (BV), and create a classification machine learning model that predicts capital expenses across various BV markets.

The web application will built using the Ruby language, the Rails framework, and the Trailblazer architecture. Users will log in via the web interface. Users may optionally choose to connect their database with our web app, or upload data as a plaintext csv, or Microsoft xlsx.

The web app will serve descriptive statistics about the input data, and allow users to view capex statistics through the lens of any arbitrary input variable (visuals much like Tableau).

In the predictive analysis section of the web app, time series regressions and classification decision trees will be graphically depicted. Interactive charts will inform the user about various models which are available to predict the target metric, and output predictions of each model.
- Why is the project being undertaken? Describe an opportunity or problem that the project is to address.

In the short term, the project will offer customers insightful, visual, descriptive statistics and predictive analytics for their data.

In the long term, the project will securely aggregate data across industries in order to improve the predictive power of all models implemented within the DataDriven community.

- What will be the end result of the project?

A web app with "Tableau"-like data visualization technology. A database management system that integrates with various client databases and data types. A prediction pipeline that trains models and offers new insights to customers daily. Each day, DataDriven creates new transformations on previous predictions to inform our customers, who are decision makers and business intelligence leaders.

Eventually, analytics services will wane, and subscription services to aggregated and anonymized data sets will be the primary revenue driver. Companies can train models based on data from their industry vertical and receive prediction results from the web interface. Subscribers may query data and implement models, but they will not be able to download data from the web app. The subscription service is required for continued benefit.

Project Milestones

- 3-5 specific and measurable objectives per semester for first & second semester

1.) Website (Ruby + Rails + Trailblazer), with a login page, registration form, user database (SQLite/MySQL), and data visualization dashboard (time series stacked bar charts, correlation heatmaps, histograms, etc.)

Semester one: Draft for BV evaluation and feedback.
Semester two: Final draft for BV evaluation and feedback.

2.) Dynamically rendered SVG data visualizations. Take the outputs of prediction pipeline scripts in JSON format, and display them using JQuery, Javascript, and D3.

Semester one: Use PNG exported graphics for data viz needs
Semester two: Use JSON output to generate SVG graphics that mimic their PNG equivalents.

3.) External database connection and foreign tables support. Customers may choose to upload data in plaintext and xlsx formats, or they may choose to connect their database so that our systems are continuously updated with their systems.
Semester one: Plain format data upload
Semester two: Foreign database connection

- Estimated completion date for each milestone (Draft -> Final Draft)
  1.) Dec 15, 2016 -> April 15, 2017
  2.) Nov. 15, 2016 -> April 15, 2017
  3.) Nov 15, 2016 -> March 15, 2017

- Both implementation and documentation milestones

Documentation for data and web teams will be developed in markdown syntax on the GitLab wiki page. The implementation milestones listed above will also serve as documentation milestones. Eventually, the wiki will be published to serve as the web app’s Technical Support documentation on the public website.

Burnup Chart (Pivotal Tracker)

![Burnup Chart](image)

Epic Milestone Progress

<table>
<thead>
<tr>
<th>Recent Epics</th>
<th>Total</th>
<th>Accepted</th>
<th>In-progress</th>
<th>Unstarted</th>
<th>Iceboxed</th>
<th>Completion</th>
</tr>
</thead>
</table>

We are using an agile development methodology. The method is organized, prescriptive, and functional. It is not easily converted into a Gannt chart, and conversion to a Gannt chart is redundant. I understand that we need to communicate task ownership and timelines. This is readily available in the GUI at pivotaltracker.com. You may also view a chart that I created, which contains the same information as a Gannt chart, but it is not
visually arranged by date. Please view the chart here. Unfortunately, we have too much data about our progress to encode in this document. We are simply moving too fast. If the chart is not readily digestible, please visit our pivotaltracker profile at https://www.pivotaltracker.com/reports/v2/projects/1863553. All project supervisors should have received invitations to join this group. PivotalTracker is developed to facilitate these types of project management tasks, and it is specifically built for web apps such as ours. Setting it up and learning how to use it was not at all trivial. I hope this will be considered during grading.

Project Budget

- Hardware, software, and/or computing resources

All hardware required for this project will be repurposed from personal resources within the group. All software required for this project will have an open source license and will not require a budget.

Personal expense to upgrade server 6x 4GB DDR3 DRAM Memory = $114
Ruby on Rails book = $80

- Estimated cost

$80 (KU Cost)

- Vendor

Newegg, LeanPub

- Special training (e.g., VR)

None

- When they will be required?

Not required

Work Plan

Julia - Rails integration with Postgres prediction database, and front end design
Adam - Front end design, HTML/CSS, D3 widgets, JQuery widgets
Quincy - Front end design, Back end design, System administration, DataViz design
Morgan - Web server security and maintenance
Andrew - Rails integration with SQLite/MySQL user tables
Kate - Front end implementation, Back end implementation

Github link https://datamunger.ddns.net
Preliminary Project Design

• How the software works

The user is greeted with a video explaining our web app services. During playback of the video, a small chat window with a picture of a customer services representative asks the user: “What are your data analytics needs today?” The answering service is handled by Amazon’s Mechanical Turk. Below the video is a login button. The login send new users to registration form

The user registers with the service. An email activation link is sent to the provided email address to activate and confirm a valid user. Once registered, the user logs into a dashboard. They are immediately greeted with a prompt to specify a prediction task they wish for us to solve. They are instructed to provide us with credentials to a database of their choice with read only access. If they do not wish to specify this, they may upload csv, xlsx format data. Alternatively, they may elect to send us a hard drive at a mailing address to be announced at a later date. They submit this form and a new story is created in our PivotalTracker workflow, using the PivotalTracker API. Both csv and xlsx format files are converted used the read_csv or read_xlsx functions from the Python PANDAS API, and saved to our postgres database for analytics tasks.

On registration, the user decides how much data they are willing to share, and a subscription fee is calculated based on the bulk of their sharing. The fee is waved for the duration of a trial prediction task. The user’s actions are now complete, they wait for a response from the analysis team.

The analysis team creates a descriptive statistic python notebook, accessible via their DataDriven Dashboard. A simple machine learning model is trained, and accessible via the predictive analytics tab of the dashboard. The predictive analytics tab describes which model was chosen by the analysis team, and why it is an appropriate model for the given task. All source is open and viewable as a Jupyter notebook in html format. This transparency is key, because DataDriven is ultimately a subscription to an anonymized database to an industry vertical. For this minimum viable product (MVP) the industry is production engineering tasks for power industry.

Machine learning predictions improve with the addition of new data. First, we train the model on data provided by the customer. Accuracy, precision, and recall for the target metric are quantified and displayed on screen. In an adjacent panel, the same model is trained on data from the entire industry vertical data set. Accuracy precision and recall are listed for this prediction task. This motivates the value of a subscription to DataDriven, and it provides source code to enable the enterprise customer to begin building their own data science team, or to begin re-training business intelligence personnel on data science programming. Alternatively, we may offer in-house data science experts to operate on an hourly basis.

The user’s experience is now locked in. They will decide to continue a subscription, and optionally hire a data scientist to perform additional tasks. Subscription fees are
determined by the bulk of data shared with the DataDriven network. All data will be transformed and anonymous. This is the key value of data munger. Enterprise customers may benefit from shared data, without giving away sensitive information about with competitors. Everyone benefits from competitors model training, and DataDriven allows them to do so without risk.

Additional project description, regarding how the analysis team will train models is specified in the "how_to_train_a_model" link below. Summarily, we will use feature engineering, data cleaning, model prediction, and model testing with cross validation.

- 3-5 graphs, charts, or illustrations
  http://datamunger.ddns.net/how_to_train_a_model.html
  http://datamunger.ddns.net/geopandas_exercise.html
  http://datamunger.ddns.net/dataviz.html (dynamically generated SVG plots with sample data)

Ethical and Intellectual Property Issues

- Ethical issues
  Our project's ethical issues are ultimately tied to data sharing. Privacy and anonymity are the key value components of the DataDriven community, and steps are taken to ensure this. We are using an SSL secured server. The SSL certificate is provided by LetsEncrypt.org. The functionality of this certificate may be viewed by visiting https://datamunger.ddns.net. The core ethical concept for DataDriven is to bring the most good to the most people. A subscription costs more for companies who do not share data with us, to encourage what may be considered ethical data sharing. Our responsibilities as the DataDriven team is to be absolutely certain these data are not traceable back to the parent company. Anonymous ID's are used, and transformations on data provide a homogeneous scheme. This task is not trivial, and this is where DataDriven creates value. Once the data are homogenenized and anonymized, they are randomly shuffled so that it is impossible to tell which data belongs to a particular company.

- IP Issues
  Our sponsors at BV have been generous to willingly give us all IP. We are using open source methodologies, and so we do NOT take credit for the algorithms we provide. We educate our users on open machine learning toolkits so they may learn how to perform analysis within their own teams. It is not our intention to offer analytics long term, and we do not retain IP rights to the prediction algorithms. This is because we believe in the open source ethos. If everyone has access to machine learning modeling source code, then the softwares will improve as a collaborative effort. Our IP, and our value, is twofold: we will homogenizing the data provided by individual companies so that each customer can develop their own models by querying the relevant columns. Second, we will ensure anonymity. For companies that allow us a persistent database, we will only store metadata that aggregates data from the source tables. This way we avoid storing data on our servers directly, and instead store the functions that transform the data. The customer's databases will be queried and transformed to comply with our industry
vertical shema on demand. This allows us to provide the experience of a single database, without having to store everything in house. This is also a key IP for our project.

Change Log

- Explain what parts of the Initial Project Description were changed and why

On the data team, we have a new member, Ryan Feehan, who will be working on genetic algorithms.

On the web team, we have decided to use an architecture that allows our web app to scale. Trailblazer is the architecture of choice.

Our project tagline is changed. “Data driven business intelligence” is our new tagline. This is more descriptive than “Data driven merit quantification”. We are solving business intelligence needs. We seek to describe some key concepts in data science and connect these ideas with the business intelligence teams that will most benefit from the insights provided by machine learning, applied statistics, and data visualizations.