

**Drought Risk Prediction in the Continental U.S.**

**Statement of Work Document**

| **Prepared by**  **(alphabetical last name)** | Yujie Cai <[yujiecai@fas.harvard.edu](mailto:yujiecai@fas.harvard.edu)>  Michael Fein <[mfein@college.harvard.edu](mailto:mfein@college.harvard.edu)>  Thee Ngamsangrat <[tngamsangrat@g.harvard.edu](mailto:tngamsangrat@g.harvard.edu)>  Jim Zhang <[jzhang4@g.harvard.edu](mailto:jzhang4@g.harvard.edu)> |
| --- | --- |
| **Prepared for**  **(alphabetical last name)** | Christopher McGrillen <[Christopher.McGrillen@quantumblack.com](mailto:Christopher.McGrillen@quantumblack.com)>  Diana Murgulet <[Diana.Murgulet@quantumblack.com](mailto:Diana.Murgulet@quantumblack.com)>  Vidur Nayyar <[Vidur\_Nayyar@mckinsey.com](mailto:Vidur_Nayyar@mckinsey.com)>  Francesca Sogaro <[Francesca.Sogaro@quantumblack.com](mailto:Francesca.Sogaro@quantumblack.com)>  Rob Wavra <[Rob\_Wavra@mckinsey.com](mailto:Rob_Wavra@mckinsey.com)> |

**Background**

QuantumBlack is a London-based worldwide data science consulting firm that is part of the management consulting firm McKinsey & Company. QuantumBlack advises clients with its expertise at the intersection of strategy, technology & design, and is proud to be one of the few leading companies in the industry applying advanced data science beyond business problems -- it publishes peer-reviewed research papers, creates open-source data science tools, and contributes to social good by providing data-driven & AI-powered advice to the public.

QuantumBlack is particularly interested in AI for social good; one of its primary goals is to optimize the usage of existing data with cutting-edge machine learning. Currently, QuantumBlack is focusing on realizing a cost-effective drought prediction project based on existing satellite images and open-source climate data.

Drought and its secondary hazards have unwittingly become the leading cause of humanitarian disasters on a global scale. The United Nations World Food Programme calls for urgent action to tackle the unprecedented hunger crises following the droughts around the globe. Being able to predict the onset of drought is particularly important, as post-drought aid is time-consuming and labor-intensive. Advance drought preparedness based on AI is effectively related to saving human lives and is worth the utmost attention from the data science and machine learning research community.

**Problem Statement**

Our goal is to build a machine learning model to predict the drought risk in the continental U.S. (CONUS). Specifically, we aim for a time-series type of prediction, incorporating the past/current states of drought risk with meteorological data to predict the drought risk for future time stamps. We plan to usse models such as autoregressions and neural networks. As we iterate further, we will add more complexity to the model architecture. Moreover, if feasible, we will add satellite images as a set of predictor into our model, and thus the project requires some computer vision techniques. Since this is a prediction project, we consider using standard metrics such as mean-squared error loss, and prediction accuracy to perform model evaluation and selection.

Besides modeling, an exciting extension of our project will be to estimate the drought impact in the CONUS region. For example, we can use the data of crop prices and the drought risk in certain areas to conjecture the economic impact caused by drought. We hope to achieve this extension because understanding the impact of drought is significant, and it perfectly matches our institution and our partner corporation's goal: AI for social good.

We foresee a few challenges in our project. First, our data are from multiple sources. We must find a way to extract the most helpful information and find a way to concatenate datasets into a favorable shape for modeling. Second, the model architecture will be considered complex if we make it to the end. It requires us to actively manage our progress and decide how much complexity we want to achieve. The complexity might also influence the prediction accuracy.

**Resources Available**

For our research we seek to combine predictor and indicator variables across a variety of data sets. Using a tool like AWS we can combine and store very large datasets remotely. Some data sets and resources we have sourced thus far include

* AWS
  + GPU or EC3 instances for model training
  + S3 Buckets for large data storage
* [FAOSTAT - Food and Agriculture data](https://www.fao.org/faostat/en/#data)
  + Production indices
  + Cop and livestock products
  + Annual data on crop yields, prices, trade, balances, etc.
* [USDA Geospatial Data Discovery on ArcGIS Open Data 1900-2020](https://www.fs.usda.gov/treesearch/pubs/43361)
  + Drought and Moisture Surplus for the Conterminous United States
  + API data service available
* [US Drought Monitor](https://droughtmonitor.unl.edu/DmData/DataDownload.aspx)
  + Week by week by location split down to US counties
  + Drought indices and drought thresholds
* [USGS Earth Explorer](https://earthexplorer.usgs.gov/)
  + Satellite images from many different satellites
  + Historical data from satellites
* [SMAP - Soil Moisture Active Passive](https://smap.jpl.nasa.gov/data/)
  + Soil moisture data based on land radar and microwave data
  + Includes historical data

**High-Level Project Stage**

We envision that our work will consist of four main phases:

1. *Data processing.* Pre-process and aggregate datasets, aligning their geographical locations, spatial resolutions, and times. Store them on AWS S3 and conduct the exploratory data analysis (EDA).

2. *Modeling.* Build a machine learning baseline model to forecast the drought risk of any geographical location ahead of time. Then, improve the model architecture from existing literature and utilize additional features.

3. *Analysis and extension.*  Evaluate the models and identify feature importance in drought prediction. Measure the correlation between drought risk and economic, agriculture impacts.

4. *Visualization.* Develop a dashboard that generates the drought risk forecast map in a certain region of a developing country. Alternatively, write a research paper to conclude our findings.

Please see the project timeline (tentative) for a detailed division of project sub-phases.

**Project Timeline (Tentative)**

| **Time of Group meeting (Tues)** | **Milestones or goals\*** |
| --- | --- |
| 02/08 | * Project set-up   + Narrow down a topic   + Identify datasets and assign datasets and literature to people   + Set up a weekly meeting time with partners * Statement of work **(Due Feb 9)** * Set up Github repo |
| 02/15 | * Discuss concerns/questions with partners * Familarize with literature and data   + Annotated bibliographies   + Access to all data * Perform EDA and data management   + Concatenate all useful data * Ignite Talk slide **(Due Feb 16)** |
| 02/22 | * EDA and data management   + Concatenate all useful data**\*\*** * Baseline model |
| 03/01 | * Baseline models * Milestone #1 presentation **(Due Mar 1)** * Summary Report **(Due Mar 2)** * Technical Attachment **(Due Mar 2)** |
| 03/08 | * Further modeling   + Possibilities of more modeling direction?   + Build up baseline architecture |
| 03/15 | * Spring break * Further modeling   + Build up the architecture |
| 03/22 | * Milestone #2 presentation **(Due Mar 24)** * Further modeling and evaluation   + Build up the architecture   + Assess accuracy |
| 03/29 | * Summary Report **(Due Mar 30)** * Technical Attachment **(Due Mar 30)** * Model selection and discussion   + Assess accuracy   + Feature importance |
| 04/05 | * Analysis and discussion   + Measuring correlation on impact |
| 04/12 | * Analysis and discussion   + Measuring correlation on impact |
| 04/19 | * Iteration of analysis, discussion * Generate pedagogical examples, i.e. use case |
| Remaining weeks | * Prepare for final deliverables |

**\*** We complete this document before we meet with our partners to discuss our project goals in great details. So, the milestones/goals are subject to change if we identify something to be unrealistic to achieve.

**\*\***Due to the data-intensive nature of our project, we might need to use the additional time to figure out the data aggregation. But ideally, we shall begin our modeling no later than the first week of March.

**Project Deliverables**

| Deliverable 1 | Machine Learning model(s) for drought risk prediction, in the form of a Github repo (not necessarily a great API but will make sure to include pedagogical examples). |
| --- | --- |
| Deliverable 2\* | A written report in the form of a research paper or a blog post. |
| Deliverable 3\* | Dashboard that generates the drought risk forecast map in a certain region of a developing country. |

**\*** Will discuss the options as we progress forward.