# Geospatial Clustering and Forecasting for Global Hotspots

Project Workbook

By
Harshada Baswaraj Jivane
Lakshmi Vihita Kesiraju
Sum Mohan Reddy Mallannagari
Vamsi Krishna Chakravartula

September 2020

**Advisor:** Prof. Dan Harkey

# Chapter 1. Literature Search, State of the Art

## Literature Search

COVID-19 has accounted for a massive global disruption. There have been numerous reports that the impact of the ongoing COVID-19 pandemic has disproportionately impacted the global communities .The goal of the project is to examine the socio-cultural and topographic nature of spatial hot spots of SARS-CoV-2 rates across the United States. In a non-Pharmaceutical way of intervention control policies such as physical distancing or social distancing, self-isolation play a key role. The existing health disparities are likely to be rapidly magnified in the context of COVID-19, and potentially extend well beyond the lifespan of the pandemic.[1]

Incorporating geographic information science and technology (GIS&T) into COVID-19 pandemic surveillance, modeling, and response enhances understanding and control of the disease. An application of the GIS&T is integrating geographic data in COVID-19 modeling and communicating the status of the disease or status of facilities for efficient functioning. Locations and availability of personal protective equipment, ventilators, hospital beds, and other items can be optimized with the use of GIS&T. [2] Geospatial clustering and predictive analysis techniques are used to achieve these objectives.

Classical clustering algorithms (e.g k-mean) are not designed for grouping spatial data. Spatial clustering is one of task mining for grouping objects spatial and extend of classical clustering. In the research, spatial algorithm used is CLARANS (Clustering Large Applications based on Randomized Search), with the reasoning that *k* partitioned clusters are better than other approaches and CLARANS was designed for big data with efficiency (computation complexity or time) and effectiveness (average distortion over the distances)[3].

Different clustering techniques for spatial data mining have certain advantages and disadvantages [4]. Depending on the need of the application, a specific technique can be selected. Another correlation has been established[5], drawing a relation between Spatial Clusters Susceptible and Epidemic Outbreaks due to Undervaccination. The method described for finding critical sets, applied to detailed population and contact network models, provides an operational tool for public health agencies to prioritize limited surveillance and outreach resources towards the most critical clusters. A Visual Multi-Scale Spatial Clustering approach is using Graph theory tools that have natural features proved to be suitable for defining the spatial structure, especially spatial neighborhood relationships. [6]

Normalized mean squared error(NMSE) and mean absolute percentage error(MAPE) are utilized to assess the accuracies of Autoregressive Integrated Moving Average Model (ARIMA), Long Short-Term Memory model (LSTM) and Random

Forests (RF). The findings indicate that RF, the only multivariate model among the three models, performs best, and the results can be used to aid in strategic decision-making on inpatient beds resource planning in response to predictable discharges[7]. Automatic predictive analytics framework (PRAF) for geospatial human geographic data consists of a feature selection procedure and a predictor based on a neural network, handling problems that fit into a Big Data environment[8].

Spatial-Temporal Density-Based Spatial Clustering of Applications with Noise (ST-DBSCAN) is used to spatially-temporally cluster the tweets, summarize word frequencies for each cluster and model the potential topics by the Latent Dirichlet Allocation (LDA) algorithm.[9].

**State-of-the-Art Summary**

Unsupervised machine learning can be very powerful in its own right, and clustering is by far the most common expression of this group of problems. Python language has evolved to support machine learning applications. Various python libraries are available to use as per the need of the application such as sklearn. Algorithms such as CLARANS, DBSCAN and BIRCH have been widely used to address the problems of geo-spatial clustering [10].

With the clusters as a base, predictive analytic models forecast metric values, estimating a numeric value for new data based on learnings from historical data. A cutting edge forecasting tool by Facebook, Prophet produces high-quality forecasts using an additive regression model that is robust to missing data and shifts in the trend, and typically handles outliers[11][12].

Airflow[13] is an end-to-end platform for data science and machine learning where you can build and deploy models quickly and manage your ML workflows at scale. AWS[14], Microsoft Azure[15], and Google Cloud Platform[16] offer many options for implementing machine learning applications on the cloud with the best in class features available to the users. Optimized storage solutions that can scale up with the vast datasets come along with the computational counterparts of the cloud providers.

PySpark is widely used for performing exploratory data analysis at scale, building machine learning pipelines, and creating ETLs for a data platform.It is a convenient language to learn especially for someone who already is familiar with Python and libraries such as Pandas.[17]

**References (Add description)**


1. Maroko, A.R., Nash, D. & Pavilonis, B.T. COVID-19 and Inequity: a Comparative Spatial Analysis of New York City and Chicago Hot Spots. *J Urban Health* 97, 461–470 (2020). https://doi-org.libaccess.sjlibrary.org/10.1007/s11524-020-00468-0
   Description: The goal of this ecological cross-sectional study is to examine the demographic and economic nature of spatial hot and cold spots of SARS-CoV-2 rates in New York City and Chicago

2. Smith CD, Mennis J. Incorporating Geographic Information Science and Technology in Response to the COVID-19 Pandemic. Prev Chronic Dis 2020;17:200246. DOI: http://dx.doi.org/10.5888/pcd17.200246
   Description: The paper talk about applications of GIS&T include developing spatial data infrastructures for surveillance and data sharing, incorporating mobility data in infectious disease forecasting, using geospatial technologies

3. I. M. K. Karo, K. MaulanaAdhinugraha and A. F. Huda, "A cluster validity for spatial clustering based on davies bouldin index and Polygon Dissimilarity function," 2017 Second International Conference on Informatics and Computing (ICIC), Jayapura, 2017, pp. 1-6, doi: 10.1109/IAC.2017.8280572.
   Description: Main subject of this paper is a cluster validity for spatial region clustering by using modified of Davies Bouldin index with Polygon Dissimilarity function

4. Chetashri Bhadane and Ketan Shah. 2020. Clustering Algorithms for Spatial Data Mining. In Proceedings of the 2020 3rd International Conference on Geoinformatics and Data Analysis (ICGDA 2020). Association for Computing Machinery, New York, NY, USA, 5–9. DOI:https://doi.org/10.1145/3397056.3397068
   *Description*: This paper gives an introduction to the clustering techniques used in the GPS based mobility datasets to find clusters.

5. Jose Cadena, Achla Marathe, and Anil Vullikanti. 2020. Finding Spatial Clusters Susceptible to Epidemic Outbreaks due to Undervaccination. In Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1786–1788.
   *Description:* This paper aims to find the possible geographical hotspots in case of an epidemic based on the prior vaccination records.

6. Jiandong Tu, Chongcheng Chen, Hongyu Huang, and Xiaozhu Wu. Proceedings. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005.

IGARSS '05. Vol. 2. 2005. 2005 IEEE International Geoscience and Remote Sensing Symposium, 2005. IGARSS '05. Web.
*Description:* This paper attempts to provide a demonstration of visual hierarchical clustering based on the graph partitioning algorithm VSG-CLUST.

7. L. Luo, X. Xu, J. Li and W. Shen, "Short-term forecasting of hospital discharge volume based on time series analysis," 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom), Dalian, 2017, pp. 1-6, doi: 10.1109/HealthCom.2017.8210801.
*Description:* This paper discusses the machine learning models and techniques for predicting the bed resources in order to support the management of the hospital for decision making on admits.

8. J. M. Keller, A. R. Buck, A. Zare and M. Popescu, "A human geospatial predictive analytics framework with application to finding medically underserved areas," 2014 IEEE Symposium on Computational Intelligence in Big Data (CIBD), Orlando, FL, 2014, pp. 1-6, doi: 10.1109/CIBD.2014.7011525.
*Description:* This paper is a study about geographic data across various location to improve medical conditions in underserved areas.

9. G. Bordogna, L. Frigerio, A. Cuzzocrea and G. Psaila, "Clustering Geo-tagged Tweets for Advanced Big Data Analytics," 2016 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, 2016, pp. 42-51, doi: 10.1109/BigDataCongress.2016.78.
*Description:* This paper is about performing analysis on locations tagged to the tweets and to suggest trips accordingly.

10. https://medium.com/predict/three-popular-clustering-methods-and-when-to-use-each-4227c80ba2b6
11. https://towardsdatascience.com/a-quick-start-of-time-series-forecasting-with-a-practical-example-using-fb-prophet-31c4447a2274
12. Taylor SJ, Letham B. 2017. Forecasting at scale. PeerJ Preprints 5:e3190v2 https://doi.org/10.7287/peerj.preprints.3190v2
13. https://airflow.apache.org/
14. https://aws.amazon.com/
15. https://azure.microsoft.com/en-us/
16. https://cloud.google.com/deep-learning-vm
17. https://spark.apache.org/docs/latest/api/python/index.html

## Chapter 2. Project Justification

Advancements in the technologies and analytical methods are key aspects in acknowledging the spatial spread of COVID-19 disease pandemic that include Interactive web-based maps and dashboards for quick understanding of reasons for illness. Lack of access to better resources in few areas has been a major drawback in identifying the cases to provide treatment on time. According to the census provided by Johns Hopkins University along with ESRI, Redlands, California, with details on the number of cases, deaths, and recoveries was useful to track the cases globally, allocate resources, and to figure out the possible preventions.

This web-based application would monitor the patient inflow, availability in nearest hospitals to visualize the information of resources in order to provide improved services to the COVID patients. The inadequate knowledge on the virus is one of the key reasons for uncontrollable spread of the virus. Suggesting the nearest hospital with the equipment information could be helpful in admitting the patients in the right time for proper assistance. Multiple media sources have influenced the spread of rumours about the disease. Accuracy in the news plays a vital role in social, economic and health consequences.

# Chapter 3. Project Requirements

## Functional Requirements

- The web application should be able to list the hotspots in nearby regions.
- The web application should be able to display the USA map locating the hotspots geographically.
- The web application should be able to locate and provide the details of nearest hospitals along with availability of beds and other resources.
- The web application should be able to direct to the nearest hospital using GPRS features.
- The web application should be able to perform sentiment analysis based on tweets in context to COVID.

## User stories for the Project

- Governing Body : As a governing body, I should be able to get a list of hotspots similar to my region. This would help me formulate control policies.
- Hospital - As a hospital admin, I should be able to see a prediction about patient inflow. This would help in allocating and mobilizing hospital resources efficiently.
- Patient - As a patient, I should be able to know the nearest hospital for treatment. This would help me get timely treatment by reducing any physical search for hospitals.
- Internet User - As an internet user, I should be able to view the trending COVID-19 tweets and correlate them to the information available on healthcare websites.

## Data requirement.

### A. Clustering

The following are the attributes for clustering hotspot and prediction process.

- Province_State - The name of the state within the USA.
- Country_Region - The name of the County (US).
- Last_Update - The most recent date the file was changed.
- Lat - Latitude.
- Long - Longitude.
- Confirmed - Aggregate case count for the state.
- Deaths - Aggregate death toll for the state.
- Recovered - Aggregate recovered case count for the state.
- Active - Aggregated confirmed cases that have not been resolved. ( Active cases = Total cases - total recovered - total deaths).

- Incident_Rate - cases per 100,000 persons.
- People_Tested - Total number of people who have been tested.
- People_Hospitalized - Total number of people hospitalized.
- Mortality_Rate - Recorded deaths *100 / confirmed cases.
- UID - Unique Identifier for each row entry.
- ISO3 - Officially assigned county code identifiers.
- Testing_Rate - Total test results per 100,000 persons. The " Total test results" are equal to " Total test results (Positive + Negative)"
- Hospitalization_Rate - US Hospitalization Rate (%) = Total number Hospitalized / Number cases. The "Total number Hospitalized" is the "Hospitalized - Cumulative" count from COVID Tracking Project. The "Hospitalized rate" and "total number hospitalized" is only presented for those states which provide cumulative hospital data.

**B. Predict the increase in patient inflow and Suggest the nearest healthcare facility**

- Total Hospital Beds - Count of beds available for COVID patients in a hospital.
- Total ICU Beds - Count of ICU beds available for COVID patients in a hospital.
- Available Hospital Beds - Count of currently available beds in a hospital.
- Available ICU Beds - Count of currently available ICU beds in a hospital.
- Adult Population - Count of number of adult citizens in a selected region.
- Population 65+ - Count of senior citizens in a selected region.
- Projected Infected Individuals - Count of COVID Infected patients
- Projected Hospitalized Individuals - Count of hospitalized patients
- Projected Individuals Needing ICU care - Count of patients who needs to be admitted to hospital
- Hospital beds needed, six months - Estimation of beds required in six months
- Percentage of Available Beds Needed, Six Months - Percentage of estimated beds required in six months
- Percentage of Total Beds Needed, Twelve Months - Percentage of estimated beds required in twelve months
- Hospital Beds Needed, Eighteen Months - Estimation of beds required in Eighteen months
- Percentage of Available Beds Needed, Eighteen Months - Percentage of estimated beds required in Eighteen months
- ICU Beds Needed, Six Months - Estimation of ICU beds required in six months
- Percentage of ICU Beds Needed, Six Months - Percentage of estimated ICU beds required in Six months
- Percentage of ICU Beds Needed, Twelve Months - Percentage of estimated ICU beds required in Twelve months
- ICU Hospital Beds Needed, Eighteen Months - Estimation of ICU beds required in Eighteen months
- Percentage of ICU Beds Needed, Eighteen Months -Percentage of estimated ICU beds required in Eighteen months

- Patient Lat - Latitude of the patient's location.
- Long - Longitude of the patient's location.
- Hospital Lat - Latitude of the hospital location.
- Hospital Long - Longitude of the hospital location

**C. Correlate COVID-19 information shared by healthcare**

- TweetID- ID of the tweet
- Tweertag- Name of the tags for each tweet
- Tweet Count - Count of tweets for each tweet
- TweetLike - Count of likes for each tweets
- Guidelines from healthcare institutions

## Techniques for evaluating accuracy

Cluster's internal evaluation -
Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified.
One such commonly used technique is Silhouette coefficient

Predictive model evaluation -
Model Evaluation is an important part of the model development process. It helps to find the best model that represents our data.
One such commonly used technique is Root Mean Square Error

# Chapter 4. Dependencies and Deliverables

## Dependencies

**COVID-19 open-source data:**

Datasets are the key source to analyze the data. All the COVID-19 data is collected from various resources, concentrated from WHO, CDC, and JHU. Pharmaceutical and other data are extracted from other sources which are to be integrated into the geographic data to represent the spatial hotspots.

**Social media data:**

Twitter Data and other social media data that is obtained from various government websites, and used for correlating with the information that is being spread on the Internet.

**Distributed enterprise systems:**

Google Cloud Platform credit to use cloud services

Kafka to publish (write) and subscribe to (read) streams of events

**Machine learning:**

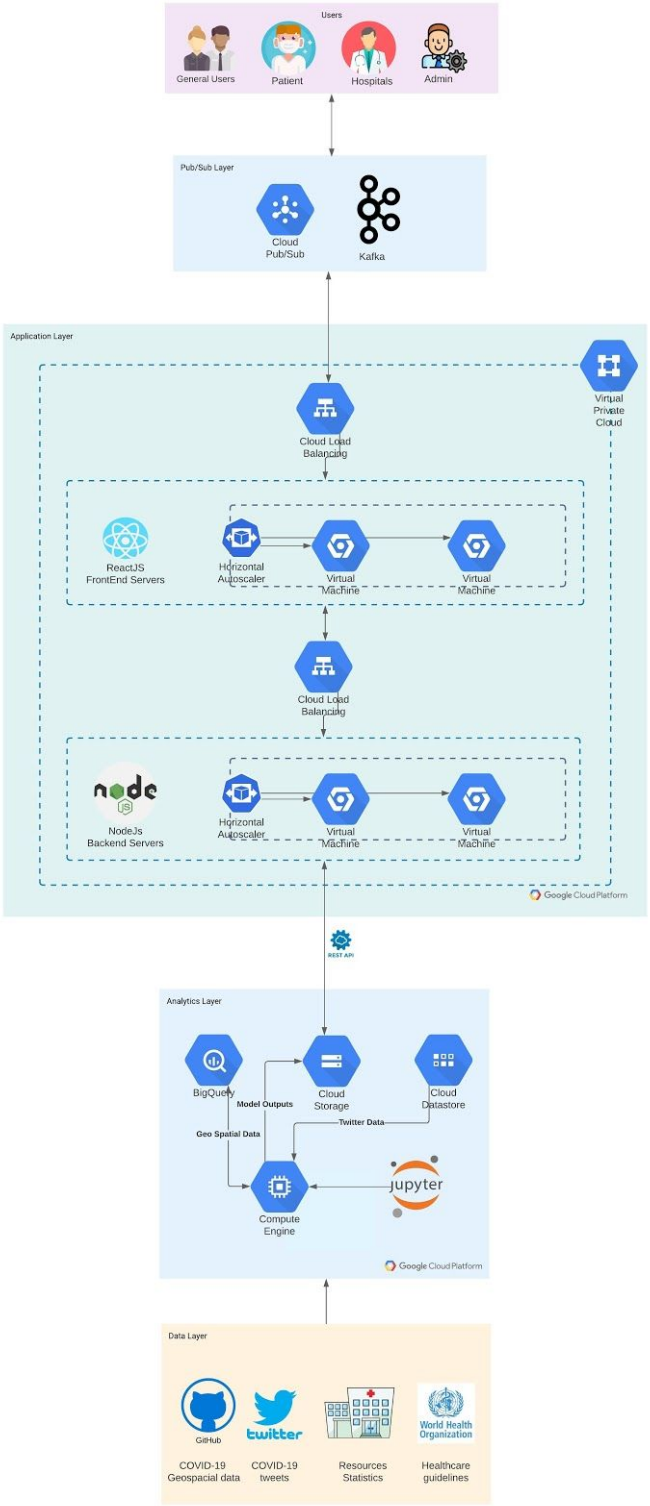
PySpark for performing exploratory data analysis at scale, building machine learning pipelines.

Scikit-Learn, Keras libraries used for data clustering and data analysis for hotspots detection and multi-scale visualization purposes. Feature selection and scaling for analysis, modeling by various techniques for obtaining the best results can be taxing.

## Deliverables

- The project's product would be an integration of machine learning and full-stack shown in the form of a web application.
- Picking up similar localities based on the previous COVID-19 hotspots and visualizing them for applying control policies.
- Predicting the rise in patient inflow and suggesting the accessible healthcare facility for patients affected.
- Correlating COVID-19 information that is being spread on Twitter and other social media sites with healthcare institutes.

# Chapter 5. Project Architecture



The proposed architecture for this project has a multi layer approach that spans from the data sources to the end users. The divided layers are:

1. End Users :

   The users will be of a varied categories covering the general users, Hospital personnel, patients and application admins.

2. Pub Sub Layer :

   The Pub-Sub layer will be serving as a message broker, improving the request handling capacities while ensuring better fault tolerance.

3. Application Layer:

   The application layer can be subdivided into 2 components that contain Frontend and Backend application servers. Each set of servers are governed by an auto-scaling system that can increase or decrease the number of active servers that are running in accordance with the incoming load. The traffic is directed to these autoscaled servers through a load balancer that sits on top of each set, ensuring a stable load to each server. The backend server picks up data from Cloud Storage, processes it and sends it to the frontend servers. The web pages render the information provided by the backend.

   Cloud providers terminologies:

| | **AWS** | **Azure** | **GCP** |
|---|---|---|---|
| **Compute** | Amazon EC2 | Azure Virtual Machines | Google Compute Engine |
| **File Storage** | Amazon S3 | Azure Blob Storage | Google Storage |
| **NoSQL** | Amazon DynamoDB | Azure DocumentDB | Google Cloud Datastore |
| **Function as a Service** | Amazon Lambda | Azure Functions | Google Cloud Functions |
| **Relational Database** | Amazon RDS | Azure SQL Database | Google Cloud SQL |
| **Container Scheduler** | Amazon EC2 Container Service | Azure Container Service | Google Kubernetes Engine |
| **App Deployment** | Amazon Elastic Beanstalk | Azure Cloud Services | Google App Engine |
| **Data Warehouse** | Amazon Redshift | Azure SQL Data Warehouse | Google BigQuery |

4. Analytics Layer:

   The analytics layer consists of the machine learning core which is the base of the complete application. It consists of multiple technological solutions that serve different purposes. The geospatial data is handled by the BigQuery, while the Cloud datastore handles the Tweets storage. The processing is done in the Compute Engine that uses Jupyter notebook to analyze the data and provide outputs. The results are then sent to the Cloud Storage component.

5. Data Source Layer :

   The data is sourced from multiple sources which include JHU's Github, Twitter feed data from APIs, Policies and Hospital data provided by the official government websites.

# Chapter 6. Project Design

[

Describe your project design. Types of design artifacts that you provide might include:

- UML diagrams (class diagrams, sequence diagrams, etc.)
- UI Mockups
- Database Entity diagrams

Evaluated: Workbook Assignment 2

]

# Chapter 7. QA, Performance, Deployment Plan

[

Describe the testing, performance evaluation, and deployment plan for your project

Evaluated: Workbook Assignment 2

]

# Chapter 8. Implementation Plan and Progress

Software implementations and requirements that are used for the project.

- Cloud Pub/Sub is used for real-time messaging services and independent applications.
- Apache Kafka is used for handling real-time data feeds. It requires the latest version of Scala and Java.
- Virtual private cloud is used for the storage and running of the code.
- Cloud load balancing is used for distributing multiple workloads among various computing platforms.
- The latest version of Kubernetes is used for horizontal autoscaler.
- Virtual machines are used for running various software.
- For the front end servers, here the latest version of ReactJS is used.
- For the Back end servers, here the latest version of NodeJS is used.
- Cloud Storage is used for storing twitter, and other social media data.
- The latest version of Python and Jupyter Notebook is required for data processing and analysis of the data.

The following are the Implementations to be done as part of the project.

- Collecting and storing geospatial data for clustering of COVID-19 hotspots.
- Collecting Hospital data for resource availability and suggesting closest available hospitals to the patient/user.
- Gathering data related to tweets to correlate COVID-19 information with guidelines provided by healthcare institutes (WHO, CDC).
- Data cleaning and preprocessing to model the data in orientation to the project.
- Evaluating and implementing the clustering technique for the hotspot.
- Predictive analysis for patient inflow to update the status of availability.
- Develop a model to suggest the nearest hospital using longitude and latitude of patient and availability in hospitals.
- Develop backend API for the following components:
    - Geoclustering Maps
    - Patient Inflow
    - Nearest Hospital
    - Twitter Correlation.
- Create the Front End with the following Web page component:
    - Web pages for displaying maps with hotspots.
    - Patient inflow would be displayed when clicked on a specific geographic area on the cluster.
        - Nearest hospital would be displayed on the Map based on the user's location.
        - Twitter data displayed as a split screen displaying the healthcare guidelines would be displayed.

## Chapter 9. Project Schedule

| Task | Specifics | Start Date | End Date | Days | Task Owner |
|---|---|---|---|---|---|
| Start of the project | Project Initiation | 08/21/20 | 08/21/20 | 1d | Harshada,Summohan, Vamsi, Vihita |
| | Research | 08/22/20 | 08/25/20 | 4d | Harshada,Summohan, Vamsi, Vihita |
| Planning | Scope | 08/26/20 | 08/27/20 | 2d | Harshada,Summohan, Vamsi, Vihita |
| | Goal Setting | 08/28/20 | 08/28/20 | 1d | Harshada,Summohan, Vamsi, Vihita |
| | Communication Plan | 08/29/20 | 08/29/20 | 1d | Harshada,Summohan, Vamsi, Vihita |
| Requirements | Data Sets | 08/30/20 | 09/02/20 | 3d | Harshada,Summohan, Vamsi, Vihita |
| | Explore Machine Learning Models | 09/03/20 | 09/05/20 | 3d | Harshada,Summohan, Vamsi, Vihita |
| | Technology Stacks | 09/06/20 | 09/08/20 | 3d | Harshada,Summohan, Vamsi, Vihita |
| | User story development | 09/09/20 | 09/09/20 | 1d | Harshada,Summohan, Vamsi, Vihita |
| Design | Architecture | 09/10/20 | 09/11/20 | 2d | Harshada, Vamsi |
| | Database Schema | 09/12/20 | 09/13/20 | 2d | Vihita, Summohan |
| | UI mockups | | | | |
| | UML diagrams | | | | |
| | Sequence | | | | |

| | | | | | |
|---|---|---|---|---|---|
| | Diagrams | | | | |
| | Component Design | | | | |
| Implementation | Environment Setup | | | | |
| | Data Preprocessing | | | | |
| | Algorithm Design | | | | |
| | Model Training | | | | |
| | Model Tuning | | | | |
| | API design | | | | |
| | Backend resourcing | | | | |
| | User interfaces | | | | |
| | Integration | | | | |
| | Deployment | | | | |
| QA | Functional Testing | | | | |
| | Unit Testing | | | | |
| | Performance Testing | | | | |
| | Regression Testing | | | | |
| Report | | | | | |
| End | | | | | |