Big Data Meets Medical Physics Dosimetry
Advanced Computer Integrated Surgery
Johns Hopkins University
Proposal for Project IX

Fumbeya Marungo
Team Member

Hilary Paisley
Team Member

John Rhee
Team Member

Todd McNutt, PhD
Mentor

Scott Robertson, PhD
Mentor

February 20, 2014

1 Recent Events

The following events occurred after the drafting of this proposal and prior to submission:

- On February 20, the project team members and mentors met and discussed the plan in detail. The plan received a highly positive reception from the mentors.
- With mentor approval, the Team Lead — Fumbeya Marungo — will present the project as an example of Hopkins research on the PhD Applicant Visit Weekend.
- Tasks 1, 3, 4, 5, 6, 8, and 11 are complete. Task 2 — maintaining the Wiki — is semester long task. Tasks 7, 9, and 10 are in process.

2 Topic

Medical physicists in oncology dosimetry design and assess treatment plans for radiation therapy. By planning the location and intensity of radiation doses, the dosimetrist's seeks to destroy malignant cells while minimizing the risk of toxicities (side effects) from damage to healthy tissue.

The standard approach to treatment planning uses a dose volume histogram (DVH). The left chart in Figure 1 displays a collection of DVH plans. Each curve is a different plan. Each point on the curve represents the percentage of a given organ's volume that received at least a given amount of ionizing radiation. Curves that are towards the right have larger percentages that receive higher total doses. The right chart in Figure 1 presents the risk of a given toxicity derived from the DVH outcomes.

Kutcher et al. (1991) presents estimates of toxicity from using DVH based on survey data from Emami et al. (1991). The model underlying Kutcher et al. (1991); Emami et al. (1991) have a number of simplifying assumptions, however. Each organ or volume of interest is

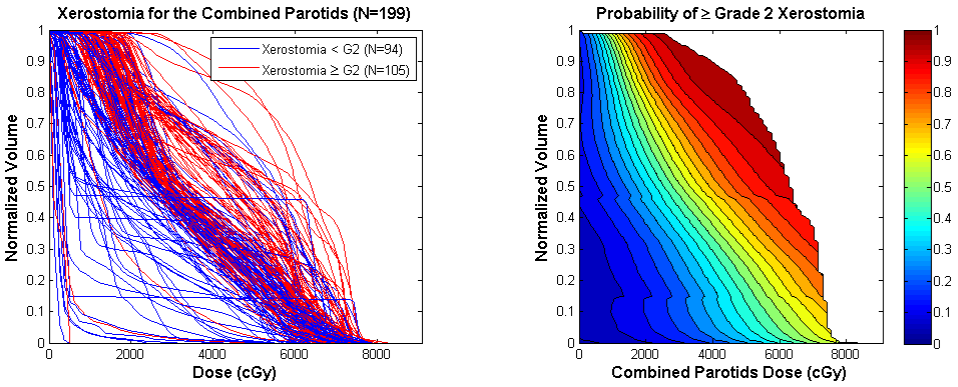


Figure 1: Individual dose exposure profiles and aggregated xerostomia risk (courtesy of Todd McNutt).

considered uniform; the internal structures are not a factor in risk assessment. In addition no information that may be available in the patient's health record — such as family history, alcohol use, previous surgery, etc. — is included. While the limitations underlying the DVH approach are well recognized, they are difficult to address. Moreover, as more patients are surviving treatment, the need to address toxicity risks becomes more acute Bentzen et al. (2010).

3 Goal

This project's goal is to apply "Big Data" analytic techniques to create a toxicity risk model(s). As we state in Section 2, oncological radiotherapy planning accounts for neither the internal structures within organs, nor the other data that is available — such as knowledge from experience with other patients or the current patient's health records.

The project entails applying data analytics to Oncospace — a database developed by Johns Hopkins Hospitals' Radiation Oncology Department. Oncospace has a diverse set of clinical data (Figure 2); by applying "Big Data" machine learning techniques to Oncospace we hope to develop a data-driven model for assessing toxicity risk.

4 Importance and Relevance

There are a number of toxicities associated with associated with radiation therapy. By mining Oncospace, knowledge learned from previous patient outcomes contributes to creating more effective and safer treatment plans (see Figure 3).

In the case of xerostomia, for example, irradiation of the parotid gland leads to severe dry mouth. Worse still xerstomia does not tend to resolve. Figure 4 illustrates that the parotid gland is highly complex. However, as noted in Section 2 the gland is modeled as a single volume. This leads to the simplified, 2-D risk assessment in Figure 1.

An successful risk model offers several benefits, it can: provide guidance for dosimetrists in assessing plans; serve as a component within an automated dosage tool; and provide

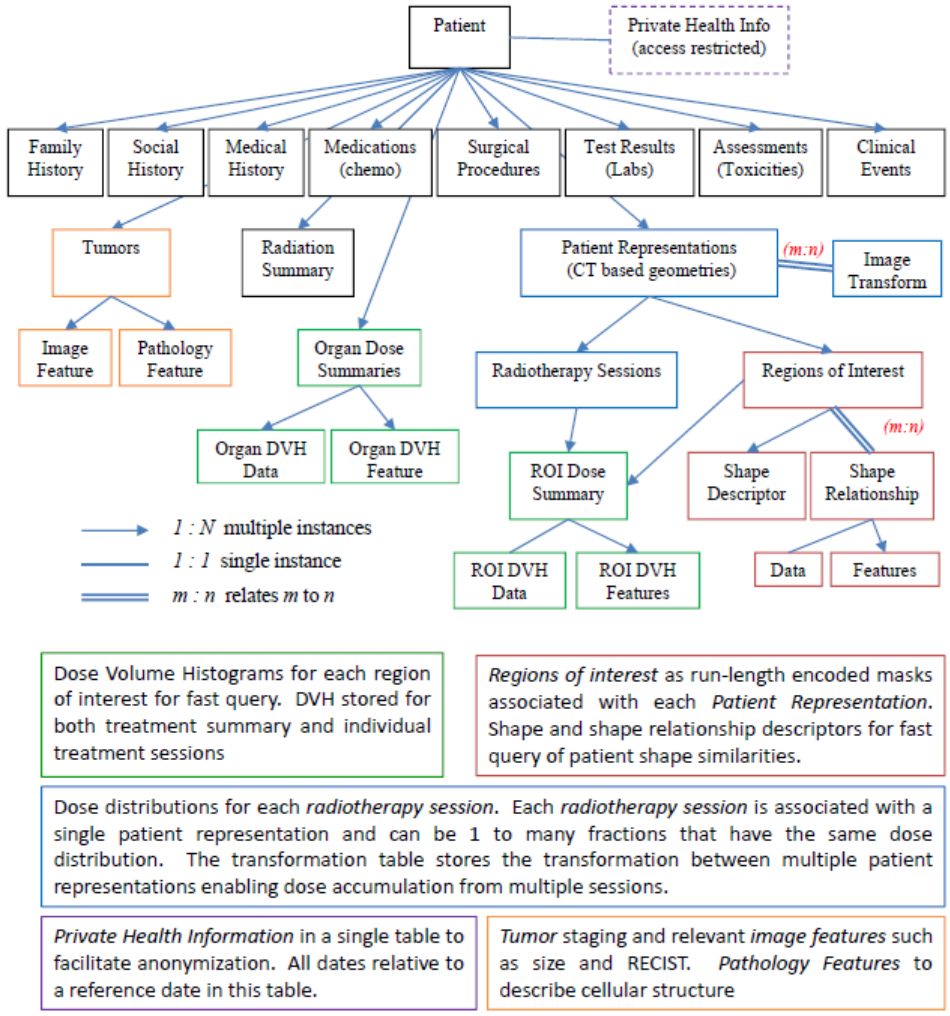


Figure 2: Oncospace (courtesy of Todd McNutt)

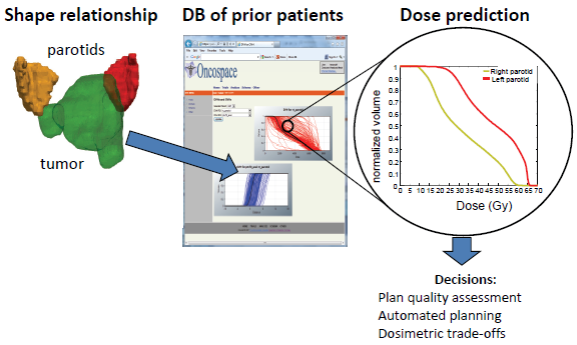


Figure 3: Using data on the clinical outcomes of previous patients provides insights into current patient treatment planning and assessment (courtesy of Todd McNutt)

greater insight into the sensitivity of different regions of healthy tissue to irradiation. Ultimately the work can provide patients with safer effective plans.

5 Technical Approach

Fayyad et al. (1996) presents data mining and knowledge discovery as a nine-step process:

1. Understanding the application domain.
2. Creating a target data set.
3. Data cleansing and preprocessing
4. Data reduction and transformation.
5. Choosing a data mining task (i.e. clustering, classification, or regression).
6. Choosing an algorithm.
7. Data mining using the algorithm.
8. Evaluating the results of the data mining step (e.g. visualization).

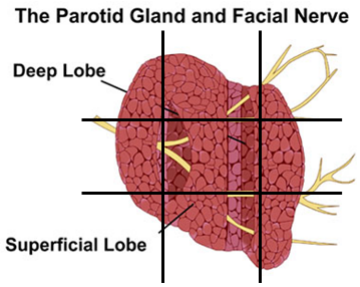


Figure 4: Oncospace has 3-D dosage data (courtesy of Todd McNutt)

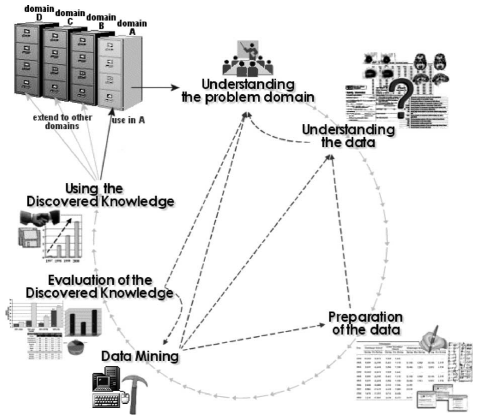


Figure 5: Data mining and knowledge discovery (copied from Cios et al. (2002))

9. Consolidating the results of the discovered knowledge.

We intend to apply this approach, with the toxicity risk model representing the final step of consolidating the results. While the Fayyad et al. (1996) appears linear, it is best visualized as an interactive process where there is a central activity stream with the process also frequently revisiting previous steps (see Figure 5).

The project plan anticipates that early and late stages will tend to require more input from the domain experts (the team's mentors in this project). The work will tend to move between the first four steps. As comfort with the domain increase, work normally begins to cycle between data preparation and mining as the models are refined.

6 Deliverables

6.1 Overview

In order to conduct this project, significant work must be done to construct a data mining pipeline. Constructing the pipeline is our minimum target deliverable. The project expects the pipeline to produce a toxicity model that equals or exceeds a DVH based model. Ideally the project can then apply the model. The block diagram in Figure 6 depicts the relationship of the various deliverables.

Data transfer between the components uses SQL queries, or .arff format text files. Both of these methods support automated lookup of the metadata for interface specification. The components are as follows:

- Data Cleaning:
  - Description:
    - Selects target data for examination from Oncospace.
    - Performs preprocessing steps, such as removing any nulls or other spurious data.
  - Deliverable:

    - Software that is responsible for cleansing the target data in Oncospace and transporting the results to the Analytic Sandbox.
    - Documentation of the cleansing process operations.

- Analytic Sandbox:
  - Description:
    - The analytic sandbox is a workspace were the project team can experiment with the data without impacting the Oncospace data set Schmarzo (2013). The sandbox will be a combination of a file server, to store .arff data text files[1], and a SQL Server database store. Data within the sandbox's SQL Server component will be indexed for faster query performance.
  - Deliverable:
    - Software that can construct the Analytic Sandbox SQL Server database.
    - A populated database and set of data files.
    - Documentation of the database and data files.

- Data Preparation:
  - Description:
    - Transforms the results of the Data Cleaning process that are stored in the Analytic Sandbox into features for Data Mining Algorithms.
    - Features are based on data from treatment plans, 3-D dose data and other clinical data available within Oncospace.
    - Determine ground truth feature.
  - Deliverable:
    - Software to perform data transformations.
    - Documentation of the stored features.

[1] see http://weka.wikispaces.com/ARFF+(stable+version)

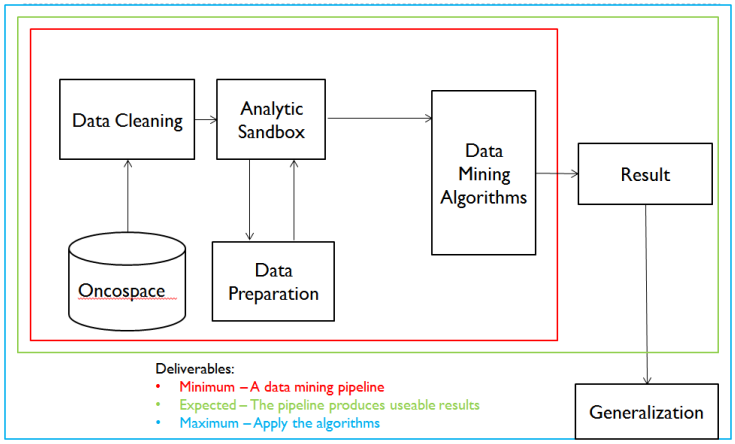


Figure 6: Deliverables

- Data Mining Algorithms:
  - Description:
    * A toxicity risk model, that uses the results of the Data Preparation process that are stored in the Analytic Sandbox.
  - Deliverable:
    * Software implemented in Java and Weka Hall et al. (2009).
    * Performance analysis of the models — e.g. using ROC analysis.
    * Documentation of the algorithm.
- Result:
  - Description:
    * The Algorithm(s) that perform well relative to the base case of a traditional DVH approach are integrated into Matlab.
  - Deliverable:
    * Matlab Integration
    * Documentation for using the algorithm within Matlab.
- Generalization:
  - Description:
    * Combining the risk model with hypothetical treatment plans to refine 3-D volume models.
    * Applying pipeline framework to another toxicity.
    * Testing additional algorithms.
  - Deliverable:
    * Documentation of results.

6.2 Technology

The project will use the following technologies:

- Platform:
  - Microsoft SQL Server. MS SQL Server database is the engine for Oncospace and the Analytic Sandbox.
  - Weka Hall et al. (2009). Weka is a mature open-source data mining platform.
  - Java 7 Standard Edition. Java is a prerequisite for Weka.
  - Matlab.
  - Scripting using JavaScript, Groovy, Jython, and/or Python.
- Software Engineering:
  - LCSR's Git Repository.
  - Maven build automation tool.
- Project Management:
  - ProjectLIBRE open-source project management software.

7 Team Member Responsibilities

The team's roles are divided as follows:

- Fumbeya Marungo, Team Lead
- Hilary Paisley, Project Manager
- John Rhee, Software Engineer

Team management uses a scaled down version of the Surgical Team metaphor in Brooks Jr (1995). The team meets two times a week. During the meeting, the team lead reviews the project plan. Work for the following meeting is then distributed based on the requirements to fulfill upcoming and open work threads.

The team member roles are designed to provide a general framework for agreeing to tasks and responsibilities. The meetings maintain communication.

8 Key Dates and Tasks

Table 1 lists the tasks, and critical dependencies for completing the project research, report and, poster by May 9, 2014. The nine Fayyad et al. (1996) steps (see Section 5) tasks are key parts of the plan; steps 5-7 are combined into the single data mining task (Task 20). A calendar of the key dates follows.

February 2014

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
						1
2	3 Intro Discussion M @ 10pm	4	5	6 Report Completed M @ 3pm	7	8
9	10 Presentation Prep M @ 10pm	11 Project Presentation	12	13 Discuss Project MM @ 4pm	14	15
16	17 Setup dev environment M @ 10pm	18	19	20 Get Database Access M @ 3pm	21	22
23	24	25	26	27 Step 1. Read papers M @ 3pm	28	

March 2014

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
						1
2	3 Prep Paper Presentation M @ 10pm	4	5	6 Paper Presentation	7	8
9	10	11	12	13 Steps 2 – 7/8 M @ 3pm	14	15
16	17 Prep Checkpoint M @ 10pm	18 Checkpoint Presentation	19	20	21	22
23	24	25	26	27 Write Initial Paper M @ 3pm	28	29
30	31					

No.	Task	Start	End	Critical Dependencies
1	Select Project	28-Jan-14	30-Jan-14	None
2	Maintain Wiki	28-Jan-14	9-May-14	None
3	Project Planning Presentation	11-Feb-14	11-Feb-14	None
4	Project Planning Report	17-Feb-14	17-Feb-14	None
5	Project Planning	3-Feb-14	17-Feb-14	None
6	Setup Development Environment	6-Feb-14	20-Feb-14	None
7	Literature Review	11-Feb-14	28-Feb-14	Input from mentors
8	IRB	14-Feb-14	19-Feb-14	None
9	Database Access	20-Feb-14	27-Feb-14	Task 8, Mentor action, Support JHH IT
10	Target Database Access	20-Feb-14	20-Feb-14	Task 8, Mentor action, Support JHH IT
11	Meeting with mentors	20-Feb-14	20-Feb-14	
12	Develop Target Database	20-Feb-14	11-Mar-14	Input from mentors
13	Begin Preparing Paper Seminar	20-Feb-14	5-Mar-14	Task 7, Input from mentors
14	Data Cleansing and Preprocessing	24-Feb-14	6-Mar-14	Task 12, Input from mentors
15	Meeting with mentors	27-Feb-14	27-Feb-14	None
16	Paper Presentation	6-Mar-14	6-Mar-14	Task 13
17	Data Reduction and Transformation	6-Mar-14	25-Mar-14	Task 14
18	Meeting with mentors	10-Mar-14	10-Mar-14	None
19	Meeting with mentors	14-Mar-14	14-Mar-14	None

April 2014

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
		1	2	3 Evaluate Other Methods M @ 3pm	4	5
6	7	8	9	10 Write Up Changes M @ 3pm	11	12
13	14	15	16	17	18	19
20	21	22	23	24 Integration M @ 3pm	25	26
27	28	29	30			

May 2014

Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
				1	2	3
4	5	6	7 Finish Poster M	8 Solidify Presentation M	9 Poster Session	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

9 Management Plan and Dependencies

As noted in Table 1, there are a number of critical dependencies that are necessary for the project's progress and completion. These dependencies include:

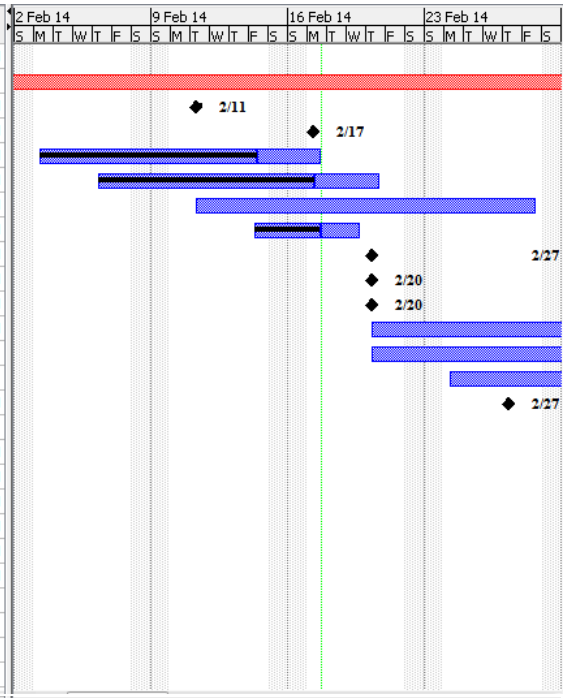
- Remote read access to the Oncospace SQL database.
- Remote read and write access to a separate SQL Server database and remote file directory for the Analytic Sandbox (see Section 6.1).
- Access to the project's mentors on a weekly to biweekly basis in person, and routinely via email.

In addition to the critical dependencies, approved funding of up to $750 per team member can serve to increase productivity and mitigate risks. For example, if the team recognizes the need for certain books, or software licenses then the members can quickly make purchases rather than waiting for approval. Student licenses for software are often inexpensive; the low costs can lead to benefits in terms of faster results and/or further progress.

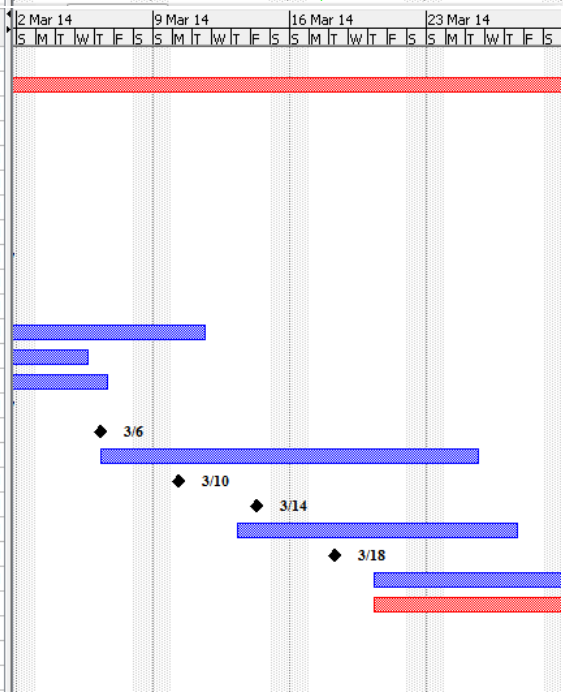
In order to avoid failure due to critical dependencies, the team maintains a project plan using ProjectLIBRE[2] open-source software. During meetings, the Gantt diagrams below, provide visual guidance to the project's current condition.

[2] http://www.projectlibre.org/

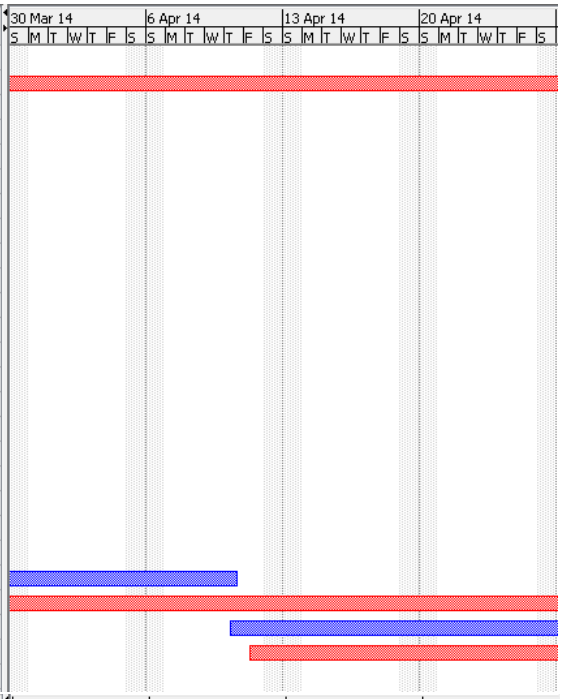
	Name	Start	Finish
1	Select Project	1/28/14 8:00 AM	1/30/14 5:00 PM
2	Maintain Wiki	1/28/14 8:00 AM	5/9/14 5:00 PM
3	Project Planning Presentation	2/11/14 8:00 AM	2/11/14 5:00 PM
4	Project Planning Report	2/17/14 8:00 AM	2/17/14 5:00 PM
5	Project Planning	2/3/14 8:00 AM	2/17/14 5:00 PM
6	Setup Development Environment	2/6/14 8:00 AM	2/20/14 5:00 PM
7	Literature Review	2/11/14 8:00 AM	2/28/14 5:00 PM
8	IRB	2/14/14 8:00 AM	2/19/14 5:00 PM
9	Database Access	2/20/14 8:00 AM	2/27/14 5:00 PM
10	Target Database Access	2/20/14 8:00 AM	2/20/14 5:00 PM
11	Meeting with Mentors	2/20/14 8:00 AM	2/20/14 5:00 PM
12	Develop Target Database	2/20/14 8:00 AM	3/11/14 5:00 PM
13	Begin Preparing Paper Seminar	2/20/14 8:00 AM	3/5/14 5:00 PM
14	Data Clensing and Preprocessing	2/24/14 8:00 AM	3/6/14 5:00 PM
15	Meeting with Mentors	2/27/14 8:00 AM	2/27/14 5:00 PM
16	Paper Presentation	3/6/14 8:00 AM	3/6/14 5:00 PM
17	Data Reduction and Transformation	3/6/14 8:00 AM	3/25/14 5:00 PM
18	Meeting with Mentors	3/10/14 8:00 AM	3/10/14 5:00 PM
19	Meeting with Mentors	3/14/14 8:00 AM	3/14/14 5:00 PM
20	Data Mining	3/13/14 7:00 AM	3/27/14 5:00 PM
21	Check Point Presentation	3/18/14 8:00 AM	3/18/14 5:00 PM
22	Assess Models	3/20/14 8:00 AM	4/10/14 5:00 PM
23	Writing Report	3/20/14 7:00 AM	5/9/14 5:00 PM
24	Integrate Software	4/10/14 7:00 AM	5/2/14 5:00 PM
25	Work on Poster	4/11/14 7:00 AM	5/9/14 5:00 PM
26	Poster Day	5/9/14 8:00 AM	5/9/14 5:00 PM



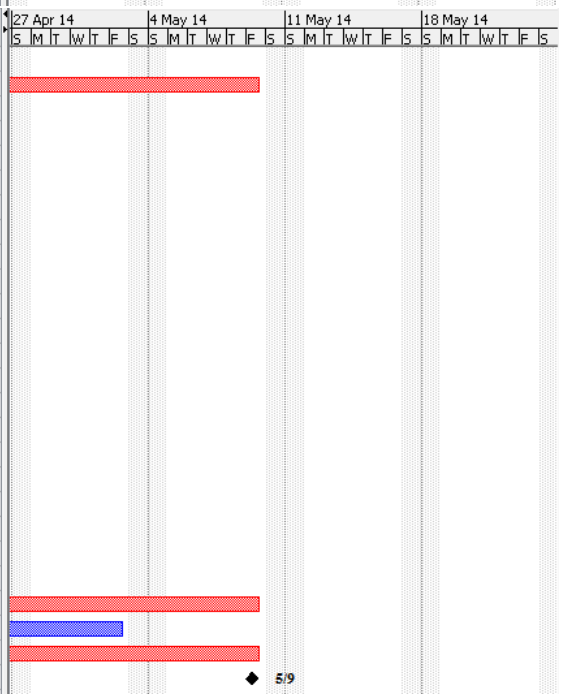
	Name	Start	Finish
1	Select Project	1/28/14 8:00 AM	1/30/14 5:00 PM
2	Maintain Wiki	1/28/14 8:00 AM	5/9/14 5:00 PM
3	Project Planning Presentation	2/11/14 8:00 AM	2/11/14 5:00 PM
4	Project Planning Report	2/17/14 8:00 AM	2/17/14 5:00 PM
5	Project Planning	2/3/14 8:00 AM	2/17/14 5:00 PM
6	Setup Development Environment	2/6/14 8:00 AM	2/20/14 5:00 PM
7	Literature Review	2/11/14 8:00 AM	2/28/14 5:00 PM
8	IRB	2/14/14 8:00 AM	2/19/14 5:00 PM
9	Database Access	2/20/14 8:00 AM	2/27/14 5:00 PM
10	Target Database Access	2/20/14 8:00 AM	2/20/14 5:00 PM
11	Meeting with Mentors	2/20/14 8:00 AM	2/20/14 5:00 PM
12	Develop Target Database	2/20/14 8:00 AM	3/11/14 5:00 PM
13	Begin Preparing Paper Seminar	2/20/14 8:00 AM	3/5/14 5:00 PM
14	Data Clensing and Preprocessing	2/24/14 8:00 AM	3/6/14 5:00 PM
15	Meeting with Mentors	2/27/14 8:00 AM	2/27/14 5:00 PM
16	Paper Presentation	3/6/14 8:00 AM	3/6/14 5:00 PM
17	Data Reduction and Transformation	3/6/14 8:00 AM	3/25/14 5:00 PM
18	Meeting with Mentors	3/10/14 8:00 AM	3/10/14 5:00 PM
19	Meeting with Mentors	3/14/14 8:00 AM	3/14/14 5:00 PM
20	Data Mining	3/13/14 7:00 AM	3/27/14 5:00 PM
21	Check Point Presentation	3/18/14 8:00 AM	3/18/14 5:00 PM
22	Assess Models	3/20/14 8:00 AM	4/10/14 5:00 PM
23	Writing Report	3/20/14 7:00 AM	5/9/14 5:00 PM
24	Integrate Software	4/10/14 7:00 AM	5/2/14 5:00 PM
25	Work on Poster	4/11/14 7:00 AM	5/9/14 5:00 PM
26	Poster Day	5/9/14 8:00 AM	5/9/14 5:00 PM



	Name	Start	Finish
1	Select Project	1/28/14 8:00 AM	1/30/14 5:00 PM
2	Maintain Wiki	1/28/14 8:00 AM	5/9/14 5:00 PM
3	Project Planning Presentation	2/11/14 8:00 AM	2/11/14 5:00 PM
4	Project Planning Report	2/17/14 8:00 AM	2/17/14 5:00 PM
5	Project Planning	2/3/14 8:00 AM	2/17/14 5:00 PM
6	Setup Development Environment	2/6/14 8:00 AM	2/20/14 5:00 PM
7	Literature Review	2/11/14 8:00 AM	2/28/14 5:00 PM
8	IRB	2/14/14 8:00 AM	2/19/14 5:00 PM
9	Database Access	2/20/14 8:00 AM	2/27/14 5:00 PM
10	Target Database Access	2/20/14 8:00 AM	2/20/14 5:00 PM
11	Meeting with Mentors	2/20/14 8:00 AM	2/20/14 5:00 PM
12	Develop Target Database	2/20/14 8:00 AM	3/11/14 5:00 PM
13	Begin Preparing Paper Seminar	2/20/14 8:00 AM	3/5/14 5:00 PM
14	Data Clensing and Preprocessing	2/24/14 8:00 AM	3/6/14 5:00 PM
15	Meeting with Mentors	2/27/14 8:00 AM	2/27/14 5:00 PM
16	Paper Presentation	3/6/14 8:00 AM	3/6/14 5:00 PM
17	Data Reduction and Transformation	3/6/14 8:00 AM	3/25/14 5:00 PM
18	Meeting with Mentors	3/10/14 8:00 AM	3/10/14 5:00 PM
19	Meeting with Mentors	3/14/14 8:00 AM	3/14/14 5:00 PM
20	Data Mining	3/13/14 7:00 AM	3/27/14 5:00 PM
21	Check Point Presentation	3/18/14 8:00 AM	3/18/14 5:00 PM
22	Assess Models	3/20/14 8:00 AM	4/10/14 5:00 PM
23	Writing Report	3/20/14 7:00 AM	5/9/14 5:00 PM
24	Integrate Software	4/10/14 7:00 AM	5/2/14 5:00 PM
25	Work on Poster	4/11/14 7:00 AM	5/9/14 5:00 PM
26	Poster Day	5/9/14 8:00 AM	5/9/14 5:00 PM



	Name	Start	Finish
1	Select Project	1/28/14 8:00 AM	1/30/14 5:00 PM
2	Maintain Wiki	1/28/14 8:00 AM	5/9/14 5:00 PM
3	Project Planning Presentation	2/11/14 8:00 AM	2/11/14 5:00 PM
4	Project Planning Report	2/17/14 8:00 AM	2/17/14 5:00 PM
5	Project Planning	2/3/14 8:00 AM	2/17/14 5:00 PM
6	Setup Development Environment	2/6/14 8:00 AM	2/20/14 5:00 PM
7	Literature Review	2/11/14 8:00 AM	2/28/14 5:00 PM
8	IRB	2/14/14 8:00 AM	2/19/14 5:00 PM
9	Database Access	2/20/14 8:00 AM	2/27/14 5:00 PM
10	Target Database Access	2/20/14 8:00 AM	2/20/14 5:00 PM
11	Meeting with Mentors	2/20/14 8:00 AM	2/20/14 5:00 PM
12	Develop Target Database	2/20/14 8:00 AM	3/11/14 5:00 PM
13	Begin Preparing Paper Seminar	2/20/14 8:00 AM	3/5/14 5:00 PM
14	Data Clensing and Preprocessing	2/24/14 8:00 AM	3/6/14 5:00 PM
15	Meeting with Mentors	2/27/14 8:00 AM	2/27/14 5:00 PM
16	Paper Presentation	3/6/14 8:00 AM	3/6/14 5:00 PM
17	Data Reduction and Transformation	3/6/14 8:00 AM	3/25/14 5:00 PM
18	Meeting with Mentors	3/10/14 8:00 AM	3/10/14 5:00 PM
19	Meeting with Mentors	3/14/14 8:00 AM	3/14/14 5:00 PM
20	Data Mining	3/13/14 7:00 AM	3/27/14 5:00 PM
21	Check Point Presentation	3/18/14 8:00 AM	3/18/14 5:00 PM
22	Assess Models	3/20/14 8:00 AM	4/10/14 5:00 PM
23	Writing Report	3/20/14 7:00 AM	5/9/14 5:00 PM
24	Integrate Software	4/10/14 7:00 AM	5/2/14 5:00 PM
25	Work on Poster	4/11/14 7:00 AM	5/9/14 5:00 PM
26	Poster Day	5/9/14 8:00 AM	5/9/14 5:00 PM



10 Reading List

Background:

- Emami et al. (1991)
- Kutcher et al. (1991)
- Burman et al. (1991)
- Bentzen et al. (2010)

Data Mining and Knowledge Discovery:

- Fayyad et al. (1996)
- Cios et al. (2002)

State of the Art:

- Kazhdan et al. (2009)

References

Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbrunch, A., Jackson, A., Marks, L. B., Haken, R. K. T., & Yorke, E. D. (2010). Quantitative analysis of normal tissue effects in the clinic (QUANTEC): An introduction to the scientific issues. *International Journal of Radiation Oncology Biology Physics*, *76*(3), S3–S9.

Brooks Jr, F. P. (1995). *The Mythical Man-Month, Anniversary Edition: Essays on Software Engineering*. Pearson Education.

Burman, C., Kutcher, G., Emami, B., & Goitein, M. (1991). Fitting of normal tissue tolerance data to an analytic function. *International Journal of Radiation Oncology* Biology* Physics*, *21*(1), 123–135.

Cios, K. J., Kagadis, G. C., & Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial Intelligenc in Medicine*, *26*, 1–24.

Emami, B., Lyman, J., Brown, A., Cola, L., Goitein, M., Munzenrider, J., Shank, B., Solin, L., & Wesson, M. (1991). Tolerance of normal tissue to therapeutic irradiation. *International Journal of Radiation Oncology* Biology* Physics*, *21*(1), 109–122.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10–18.

Kazhdan, M., Simari, P., McNutt, T., Wu, B., Jacques, R., Chuang, M., & Taylor, R. (2009). A shape relationship descriptor for radiation therapy planning. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2009*, (pp. 100–108). Springer.

Kutcher, G., Burman, C., Brewster, L., Goitein, M., & Mohan, R. (1991). Histogram reduction method for calculating complication probabilities for three-dimensional treatment planning evaluations. *International Journal of Radiation Oncology* Biology* Physics*, *21*(1), 137–146.

Schmarzo, B. (2013). *Big Data: Understanding how Data Powers Big Business*. John Wiley & Sons.