Project IX
Big Data Meets Medical Physics Dosimetry

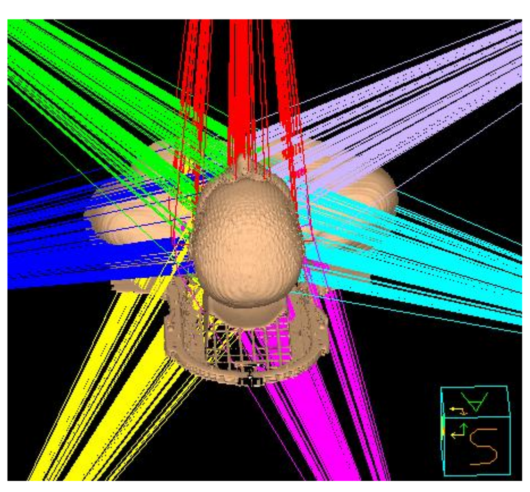
Computer Integrated Surgery II

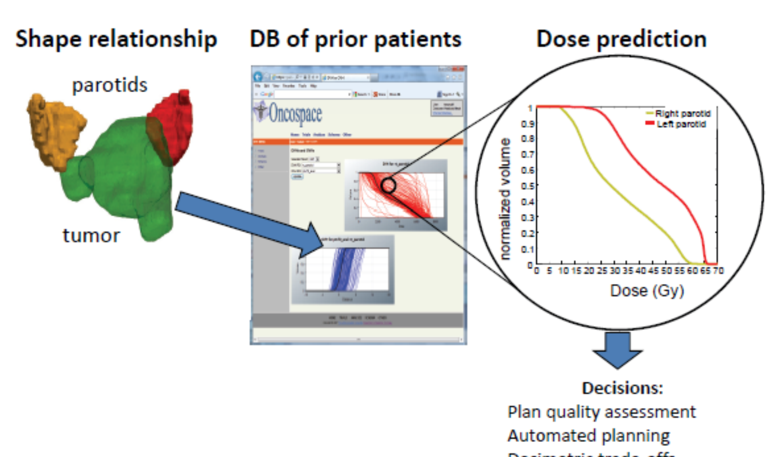
Spring, 2014

Fumbeya Marungo, Hilary Paisley, John Rhee

Introduction

- **General problem**: The goal in radiotherapy is to maximize dose to a region of interest while minimizing the risk of complications. The question that naturally arises is can data collected in the clinic inform future treatment plans.
- **Deliverables**:
  - A general pipeline which processes data and a normal tissue complication probability (NTCP).
  - Apply pipeline to analyze 3D treatment data for the parotid glands.
  - Generalization of the data mining approach to a second organ (the larynx).
- **Big picture**: A multi-factor method of calculating complication risk can allow for personalized treatment plans, better understanding of dosage placement, and automated tools.



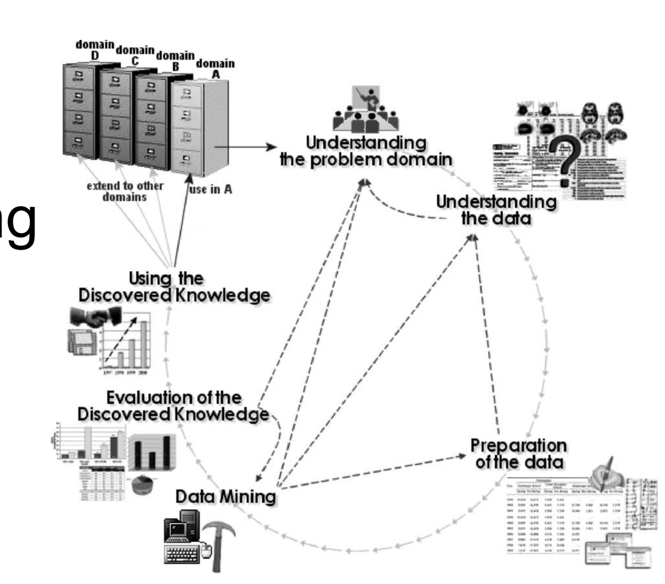


The Problem

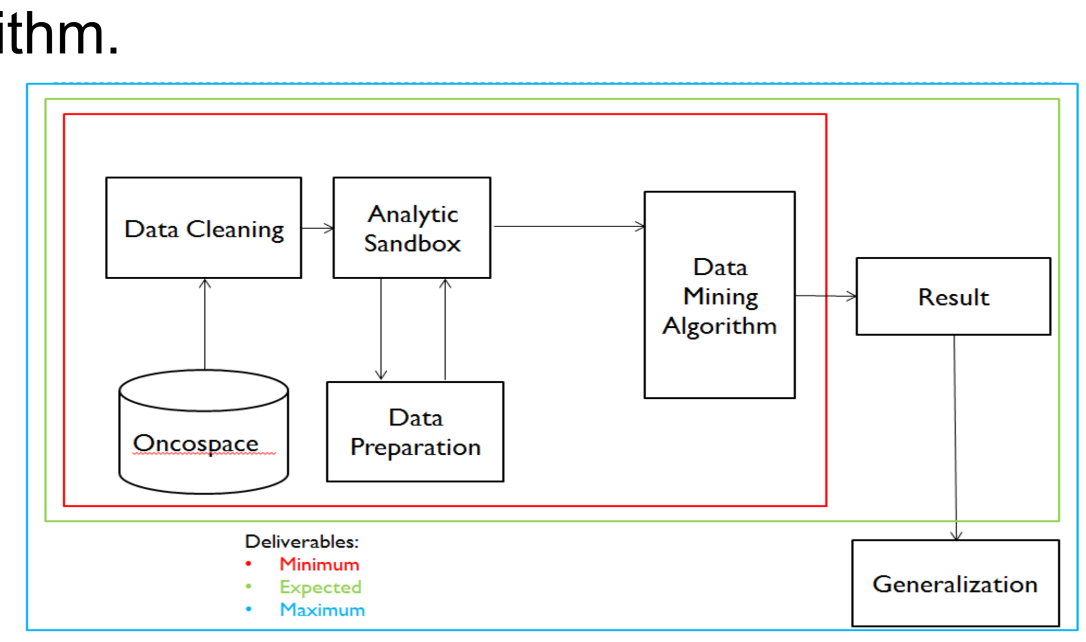
- Underlying medical problem: Currently, clinicians receive little guidance in determining the dose levels and placement.
- Current state: The LKB model is of limited use in the clinic. It assumes an equivalence between large doses over small volume and small doses over larger volumes. No other factors are considered.
- Need for solution: Highly personalized risk assessments may arise from using data collected in the clinic.

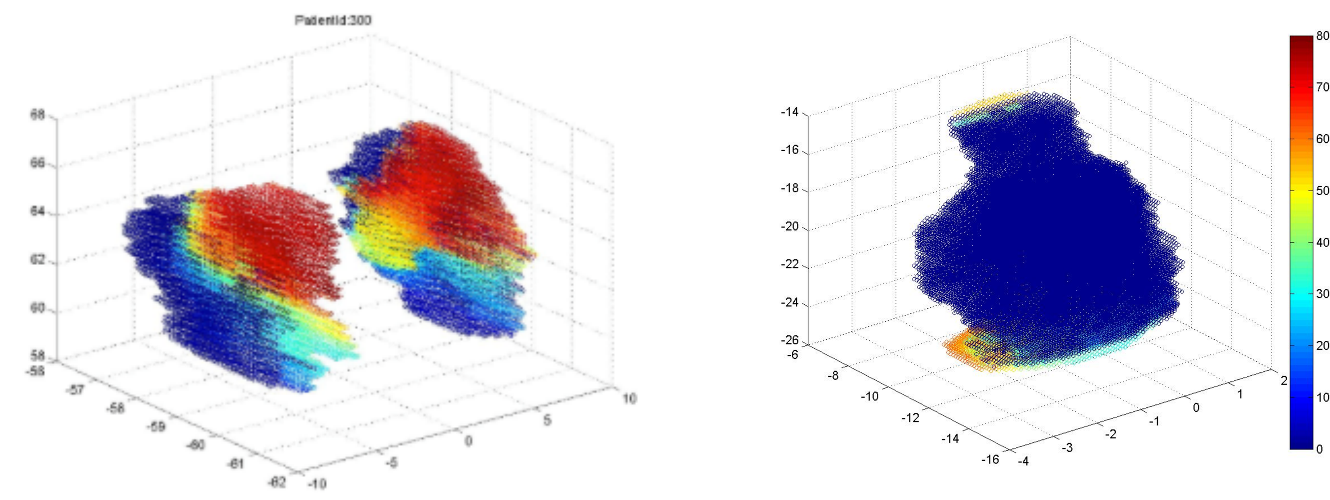
The Solution

- We have utilized the 9 steps of Knowledge Discovery in Databases (Fayyad et al.).
  1. Understand Application Domain
  2. Create Target Data Set
  3. Data Cleaning and Preprocessing
  4. Data Reduction and Projection
  5. Find a Data Mining Method
  6. Exploratory Analysis
  7. Data Mining
  8. Interpretation
  9. Implementation



- Pipeline: We have cleansed the data and transformed it into a form where we can visualize the doses of the region of interest. Additionally, we have divided the parotid into sub-regions and run the data mining algorithm.





Data Mining Outcomes and Results

- Classification using Random Forest and Linear Regression.
- Linear Regression performed with and without bagging.
- Calculated ROC AUC using leave-one-out validation.
- Compared to conventional LKB model's ROC AUC.

Xerostomia (Parotid Glands)				Voice (Larynx)			
BLR	LR	LKB	RF	BLR	LR	LKB	RF
0.732	0.730	0.700	0.701	0.916	0.893	0.596	0.870

Future Work

- Deformable registration of organ regions.
- Creating a clinical interpretation of the results.
- Further data collection; potential integration with other institutions to amass more data.

Lessons Learned

- Real world application of the 9 step KDD process for data mining to the medical physics domain
- Relatively low performance of RF implies more data can improve prediction.
- Larynx may be highly sensitive to dose placement.

Credits

- Fred Brooks presents a "Surgical Team" approach developed by Harlan Mills and F. Baker at IBM.
  - Fumbeya Marungo – Project Lead
  - Hilary Paisley – Project Manager
  - John Rhee – Software Engineer

Publications

- Breiman, Leo. (2001), Random Forests..
- Fayyad, et. al. (1996), From Data Mining to Knowledge Discovery in Databases.
- Kutcher, G. J., & Burman, C. (1989). Calculation of complication probability factors for non-uniform normal tissue irradiation: The effective volume method.
- Lyman, J. T. (1985). Complication probability as assessed from dose-volume histograms.
- Tolosi et al. (2011), Classification with correlated features: unreliability of feature ranking and solutions.

Support by and Acknowledgements

- Dr. Todd McNutt, Dr. Scott Robertson
- The Johns Hopkins Oncology Radiation Department