A Machine Learning Approach to Medical Physics Dosimetry

Fumbeya, Marungo, **Hilary Paisley**, John Rhee

Mentors: Dr. Todd McNutt, Dr. Scott Robertson

JOHNS HOPKINS MEDICINE

ciis

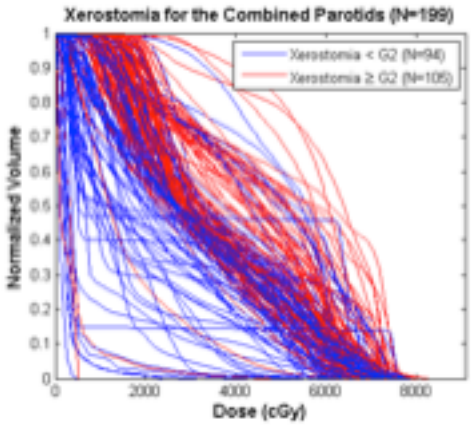
ERC | CISST

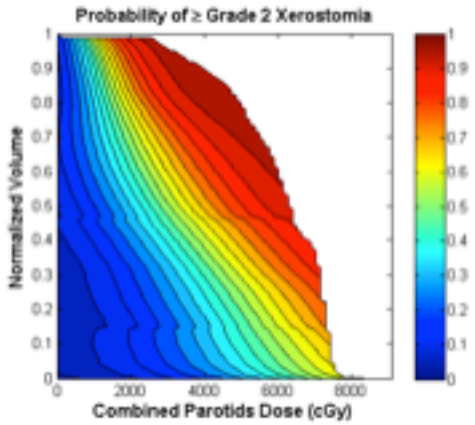
Collaborative Laboratory for BIGSS Biomechanical- and Image-guided Surgical Systems

LABORATORY FOR Computational Sensing + Robotics THE JOHNS HOPKINS UNIVERSITY

Summary of Project IX

- Use “Big Data” techniques to create a toxicity risk model after irradiation of the parotid gland





Images courtesy of Dr. Todd McNutt,
Dr. Scott Robertson

JOHNS HOPKINS MEDICINE

cils

ERC | CISST

Collaborative Laboratory for BIGSS
Biomechanical- and Image-guided Surgical Systems

LABORATORY FOR Computational Sensing + Robotics
THE JOHNS HOPKINS UNIVERSITY

Research Papers

- Fayyad, et. al. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996.
- Breiman, Leo. Random Forests. Machine Learning, 45, 5-32, 2001.

KDD Process

- Knowledge Discovery in Databases (KDD)

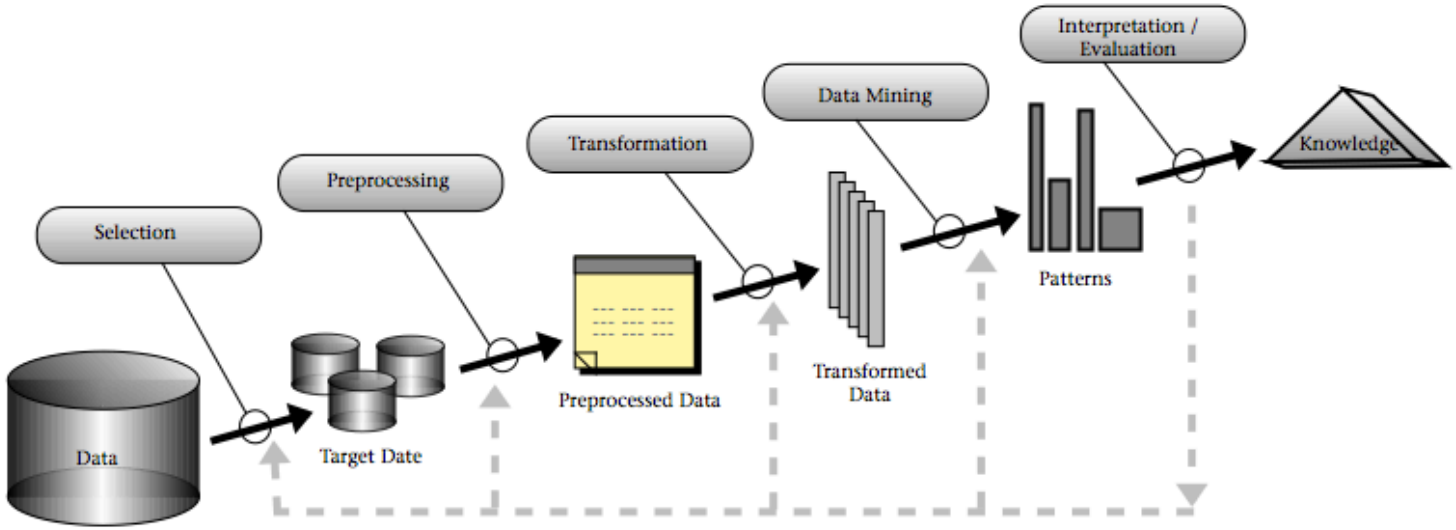


Image courtesy of Fayyad, et. al.

Understanding Application Domain

- First must understand domain and goal of the KDD process
- Our goal is to assess importance of specific voxels after parotid irradiation

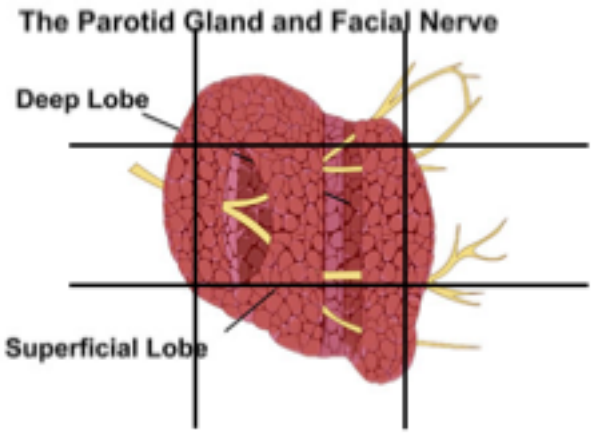


Image courtesy of Dr. Todd McNutt,
Dr. Scott Robertson

Create Target Data Set

- Using the Oncospace data provided, we must find relevant data
  - Dosage
  - Voxel
  - Patient toxicity

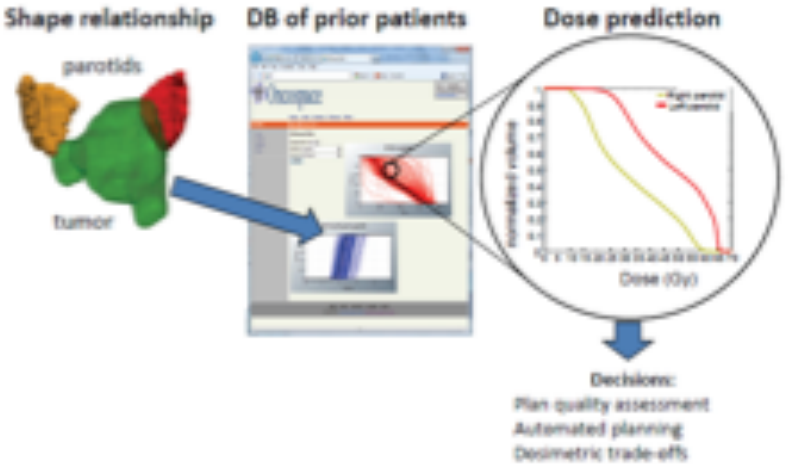


Image courtesy of Dr. Todd McNutt,
Dr. Scott Robertson

Data Cleaning and Preprocessing

- Cleaning involves removing noise, accounting for missing data fields, and being able to handle possible changes in data
- A major problem in medical data is to account for possible movements in patient during irradiation and location error with several treatments

Data Reduction

- Finding useful features to represent the data depending on the goal of the task
- This will require extensive understanding/research of the database provided to us

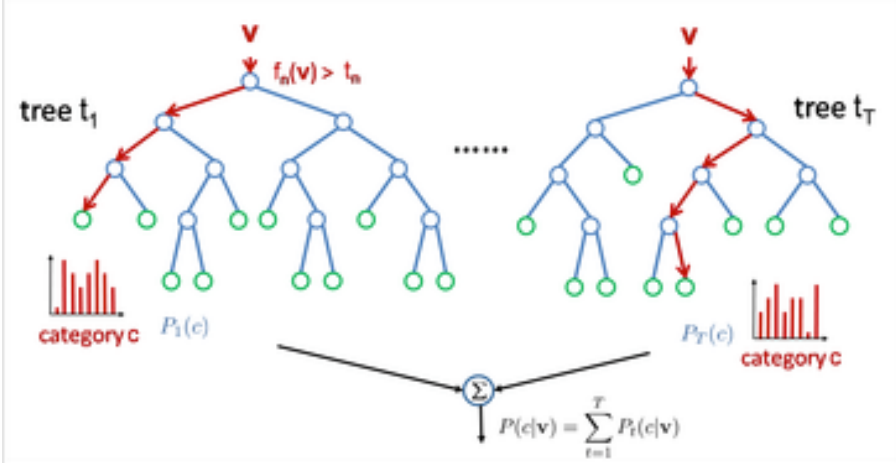
Image courtesy of Dr. Todd McNutt,
Dr. Scott Robertson

Data Mining

- Based on the goal of the KDD process, determine an appropriate data mining technique
- We have decided to try a Random Forest

Random Forest

- A random forest is a classifier consisting of a collection of tree-structures classifiers $\{h(\mathbf{x}, \Theta_k), k=1, \ldots\}$ where the Θ_k are IID random vectors and each tree casts a unit for the most popular class at input **x**



Tree Bagging Application

- For each sample, select random subset of features
- Train a decision tree on the samples
- To predict, use averaging from the individual decision trees

Why Random Forest?

- Computationally efficient
- Convergence/Correlation
- Easy to extract key features

Interpretation

- If the data mining results are poor, return to any of the previous steps for reevaluation of the data
- It is helpful to model the results of the data mining to better interpretation of success

Incorporation

- Using the knowledge gathered for a greater use
- Our goal is to integrate the methods/results to be used for different toxicities or different regions of interest

Potential Pitfalls for the Application

- Overfitting – using a limited set of data
  - Solution: cross-validation, regularization
- Changing data – continue to add/change data
  - Solution: incremental methods for updating the patterns
- Noisy data – missing fields in medical data
  - Solution: identify hidden variables/dependencies and account for them

Relevance of Papers

- Every step of the KDD process is vital to the success of our project (not just the data mining step)
- Random forests may be a good method for the project