An intelligent system for churn prediction and customer retention: a case of telecommunications company

Parth Sarangi

Computer Science and Information Management
Asian Institute of Technology Thailand

Thesis Proposal, April 2017

Table of Contents

Overview

- Telecom industry is highly competitive
- Government deregulation policies
- Affordable handsets and Technological advancements
- Disruptive plans and services by rival companies
- Database reveals trends of service usage
- Data mining to identify churn customers

Overview(contd.)

Operators	Customer Count in Aug 2016	Increase or Decrease in Period				Customer Count in Dec 2016
		Aug - Sep	Sep – Oct	Oct – Nov	Nov – Dec	
Airtel	257	2	2	1	2	265
Vodafone	200	0.5	1	0.8	1.8	204
Tata Indicom	58	- 1	- 1	- 1	- 1.6	52
Reliance Jio	0	15	19	16	20	72

- TRAI - Telecom regulatory authority of India
- Reports subscribers at end of every month
- Table shows Relance Jio acquiring 72 million at end of 4 months
- Launch of 4G services by Jio have jolted the revenues of Airtel, Vodafone, Tata Indicom.
- High customer churn noticed

Problem Statement

- Constant product marketing by competitors
- Various cost effective data schemes. Night time free or high speeds, or unlimited usage plans
- Proactive mindset of incumbent services provider to identify unfaithful customers
- ARPU of retaining customer is higher
- High investment cost of acquiring new customer
- Not a fully integrated system developed for churn prediction

Objectives

Overall objective - develop an intelligent system for churn prediction and customer retention ICPCR.

Specific objectives :

- Design models and evaluate prediction performance for churn.
- Build the system of intelligent churn prediction and customer retention system.
- Evaluate the system for reliable performance.

Limitations and Scope

- Many models for churn prediction.
- Scope of this thesis is tentatively limited to build ICPCR with 3 models - Decision tree , Support Vector Machine , Artificial Neural Network

Table of Contents

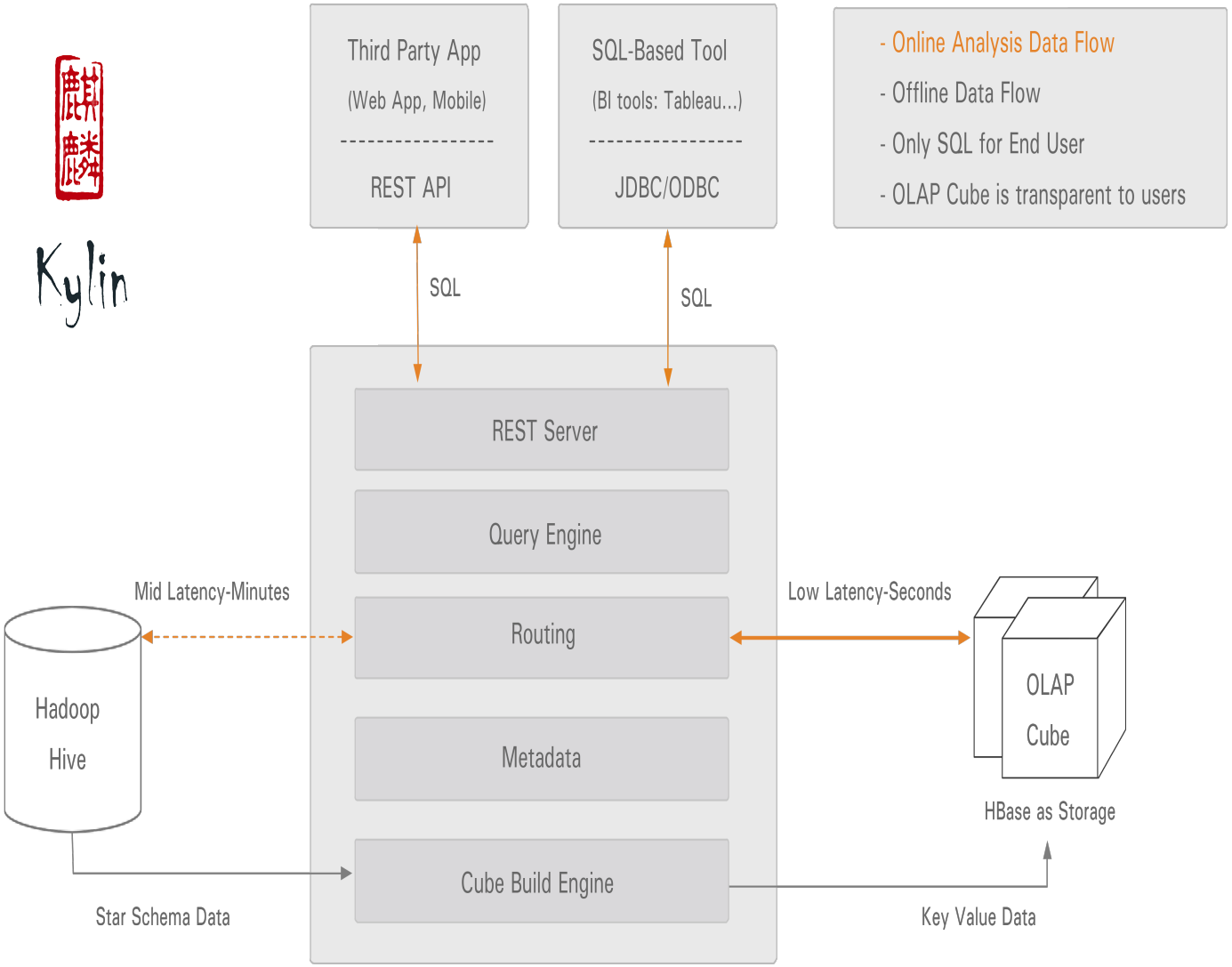
Customer Churn & Retention

- In the research paper it was found that Companies profit if they can retain customer
- Customers are most valuable asset.
- Long serving customers influence new customers to buy contracts
- Average Revenue Per User ARPU for telecom company is high if a customer stays
- If customers churn there is a loss of revenue
- Also acquiring new customer is expensive
- Forbes predicted a 10% swing in revenue if customers are retained

OLAP & Datawarehouse

- Data warehouse is a collection of Data marts
- Data marts are generally summarized tables of important data from Business units
- OLAP - Online analytical processing
- OLAP cube is the heart of an OLAP system
- There are two types - MOLAP & ROLAP. MOLAP is most common
- Apache Kylin is an open source OLAP solution

OLAP & Datawarehouse(contd.)



Data Mining

- John Naisbett (author of famous 'Megatrends') said "We are drowning in information but starved for knowledge"
- Data mining techniques can broadly be classified into two categories
  - Supervised learning
  - Un-Supervised learning

Data Mining(contd.)

Supervised Learning : The dependent and control variables are known. Classification and regression algorithms

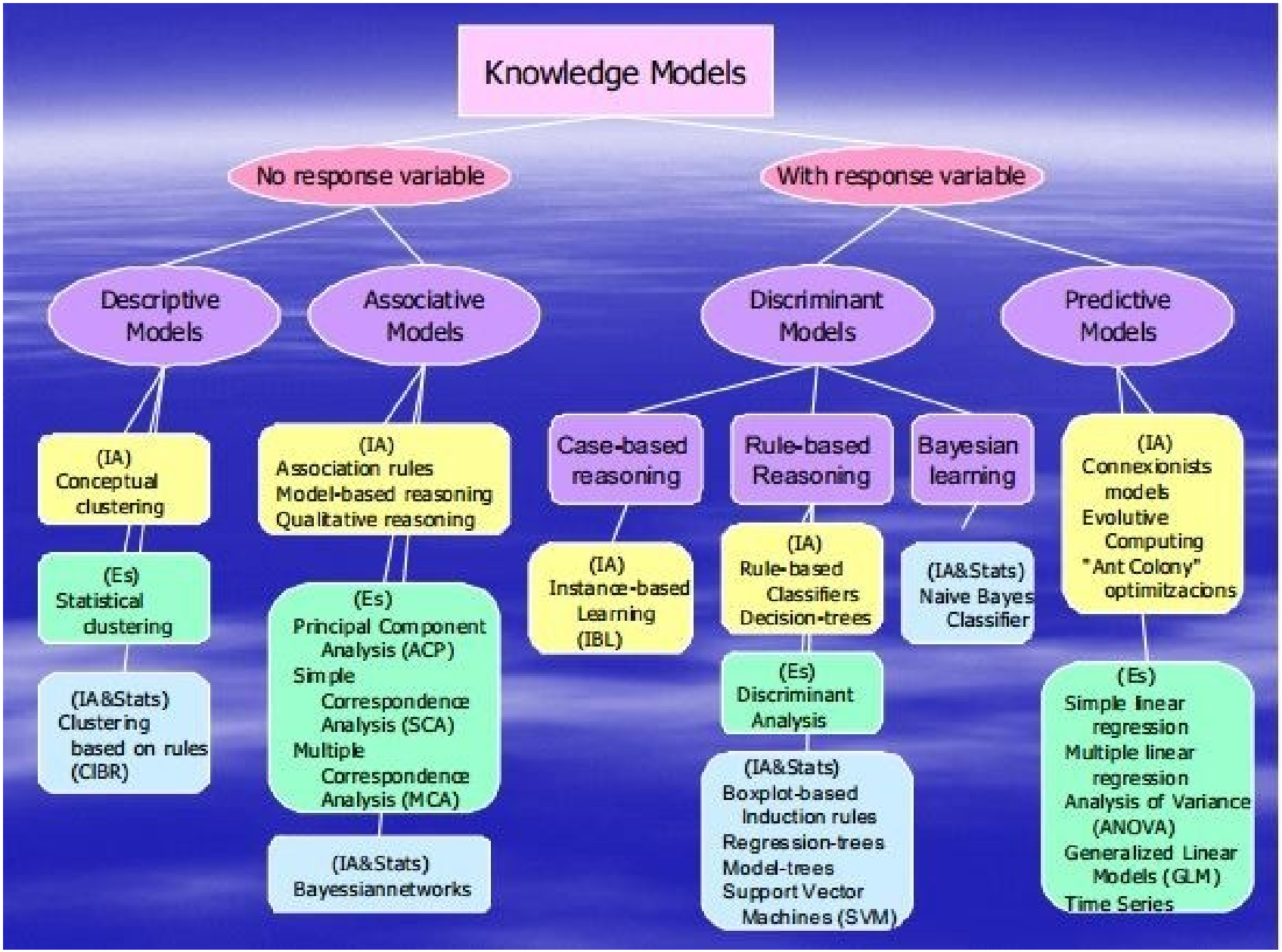
- Linear
- Multiple
- Nonlinear
- Logistic
- Decision tree
- Random forest

Data Mining(contd.)

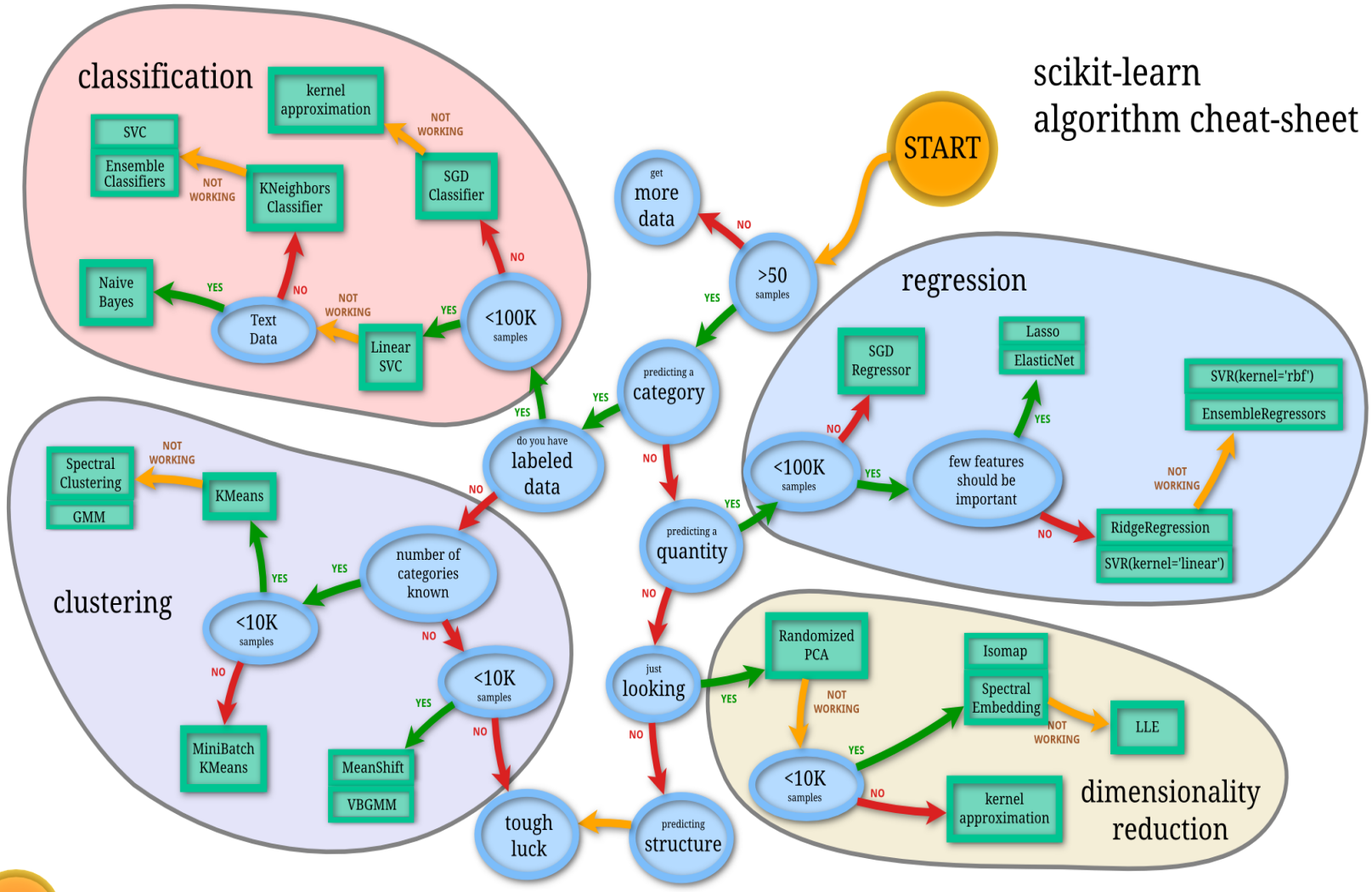
Un-Supervised Learning : The dependent and independent variables are unknown

- k means clustering
- Apriori clustering
- Hierarchical clustering
- Hidden Markov models
- Self Organizing Maps

Choosing the Right Data Mining Technique



Scikit model selection



Back

scikit learn

Data Mining(contd.)

Softwares

- Weka
- Knime
- Rapidminer

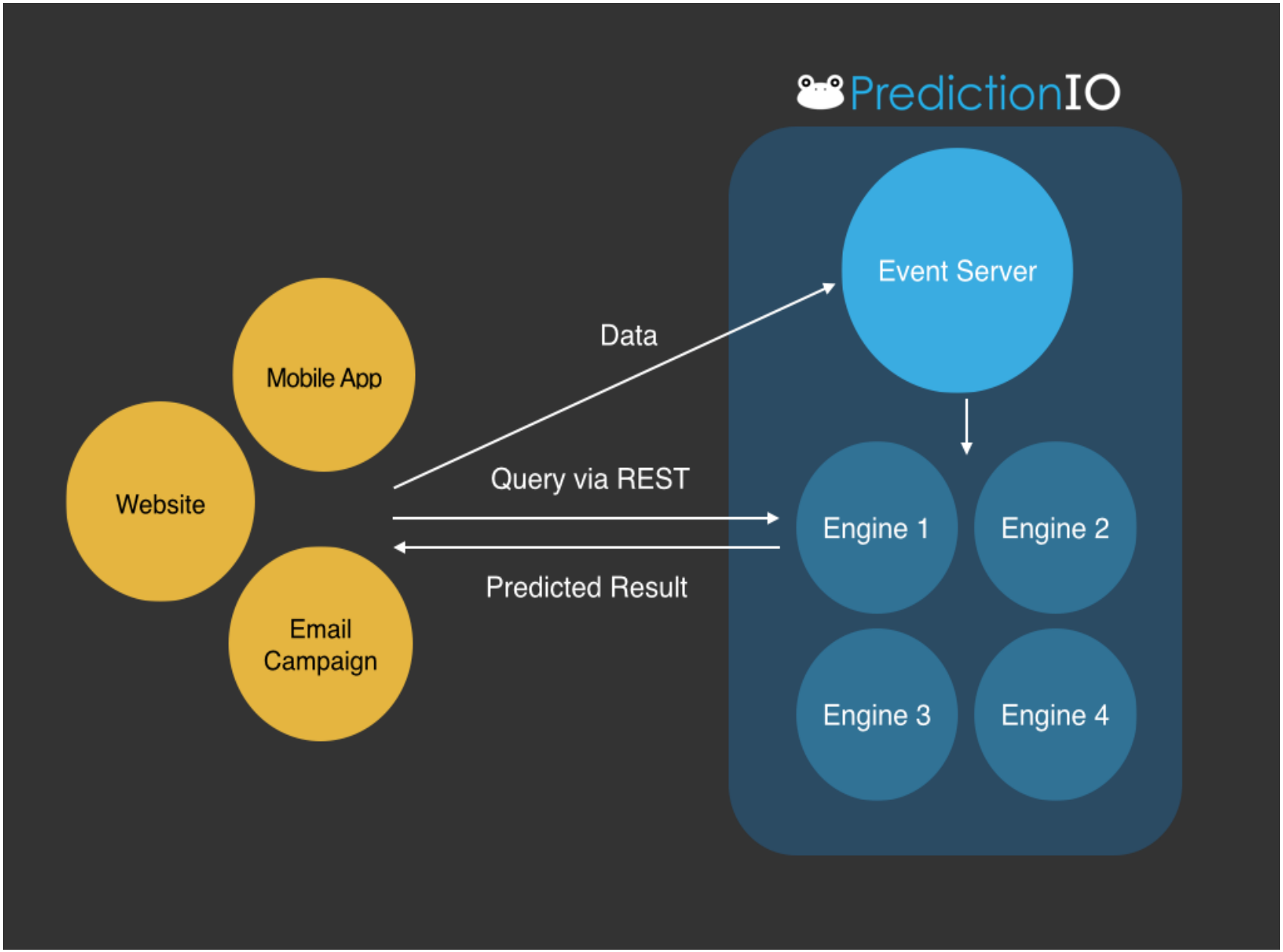
Libraries

- Tensorflow
- mlpack
- H20
- Mlib
- Scikit

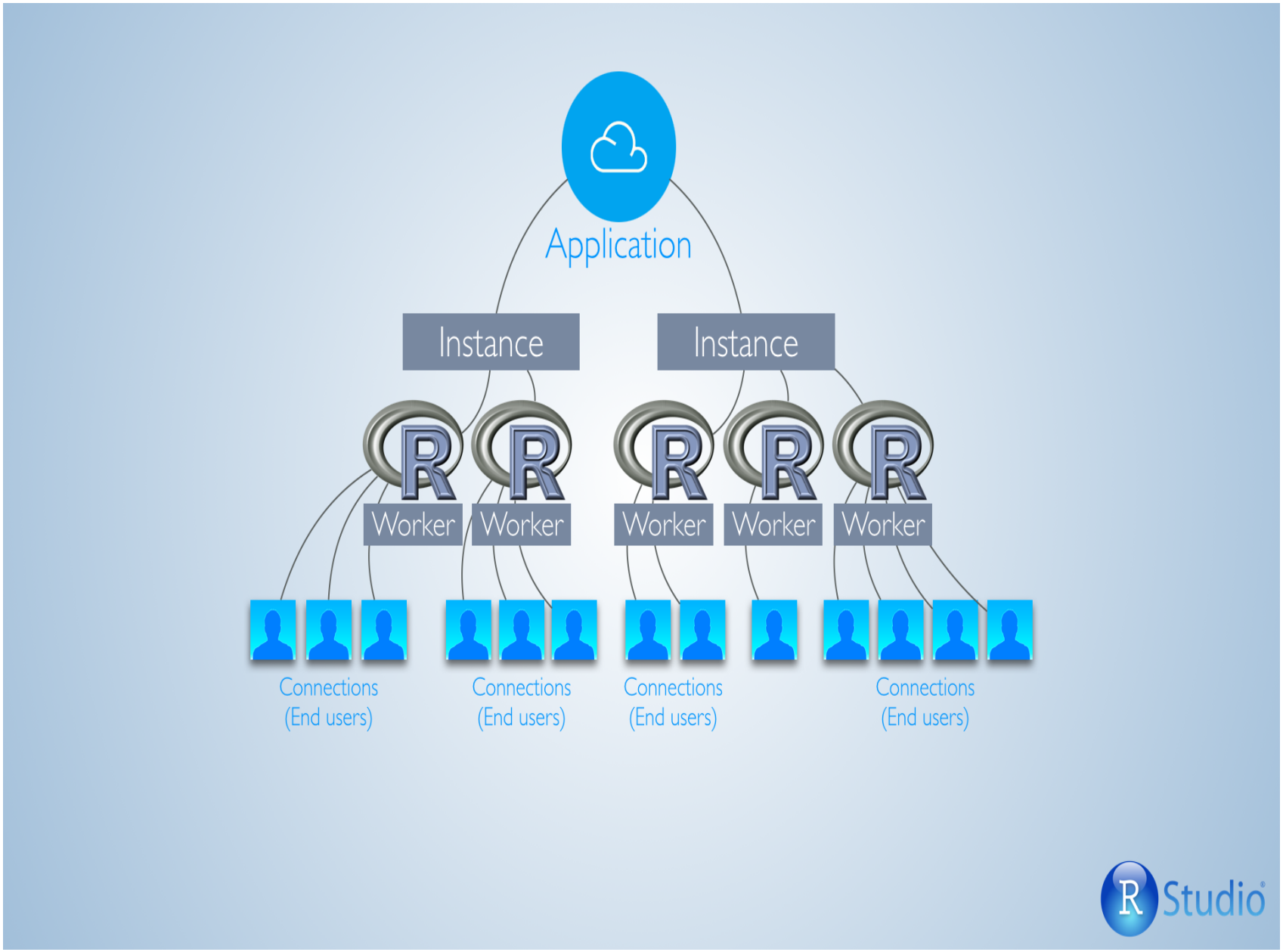
Servers

- DeepDetect
- Apache PredictionIO
- Shiny

Apache PredictionIO



Shiny Architecture



Model Evaluation Metrics

- Holdout technique
- k-fold Cross validation technique
- Sentivity and Specificity

Review of Selected Research Papers

Present 3 papers which are relevant to Churn prediction

- Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry
- A Hybrid Churn Prediction Model in Mobile Telecommunication Industry
- A comparison of machine learning techniques for customer churn prediction

Review of Selected Research Papers(contd.)

Modeling & Simulation of a Predictive Customer Churn Model for Telecommunication Industry

- Technique used by researchers was a Fuzzy inference system
- Combination of Neural network with fuzzy logic
- They modeled Membership functions for the attributes
- Data used was call detail record of 5000 subscribers
- It has 21 attributes and only 9 were selected
- Precision around 80%; Recall of 92.7% and Accuracy 95.8%

Review of Selected Research Papers(contd.)

A Hybrid Churn Prediction Model in Mobile Telecommunication Industry

- Presents a combination of Voted Perceptron and Logistic regression
- Compared performance with Logistic regression and Voted perceptron as individual prediction models
- WEKA software was used to model the predictors
- Call detail records of 2000 customers was sourced from Asian telecom company
- 23 attributes were used for modeling
- Results : Hybrid models performed better than individual models

Review of Selected Research Papers(contd.)

A comparison of machine learning techniques for customer churn prediction

- researchers present a well meted out comparison between the normal model functions and their corresponding boosted models
- performance criteria was based on the F-score
- series of simulations based on the Monte Carlo method
- 5 DM techniques - Back-Propagation algorithm , Support Vector Machines, Decision Trees, Naive Bayes and Logistic Regression.
- churn dataset hosted at UCI Machine learning repository
- 100-fold cross validation technique was used to reduce bias
- Ratio of training to testing set is about 2 : 3

Review of Selected Research Papers(contd.)

- boosting algorithm Adaboost Adaboost.M1
- R programming was used for modeling the simulation experiment
  1. tested classifiers run with data and performance of F-score measured
  2. boosting algorithm was applied and performance F-score measured
- Results
  - Best performance : 2 layer BPN with 15 hidden nodes and Decision tree classifier
  - SVM scored lower followed by Naive Bayes and Logit Regression at last.
  - After boosting SVM got best performance accuracy of 97% and F-score 84%

Table of Contents

Research Methodology

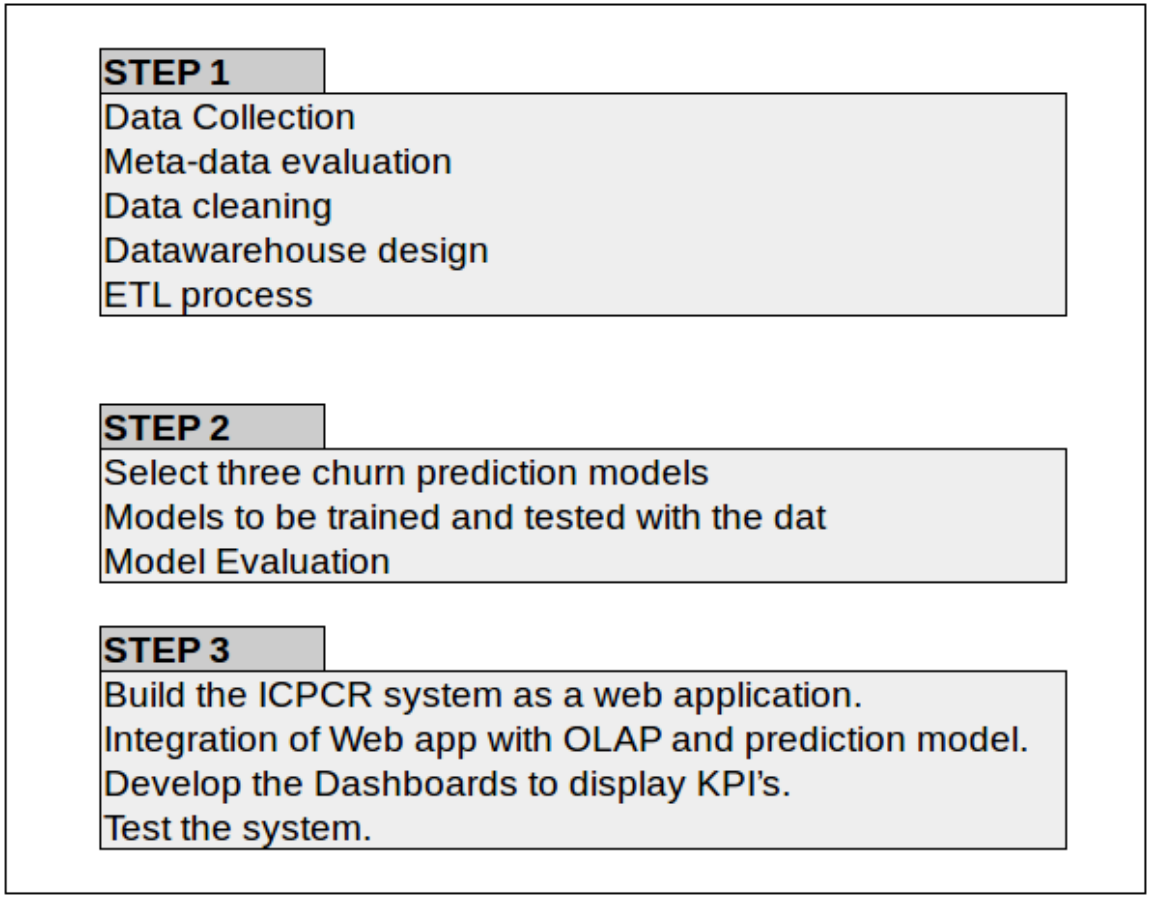


Figure: Research Methodology

Data Preprocessing and Datawarehouse Development

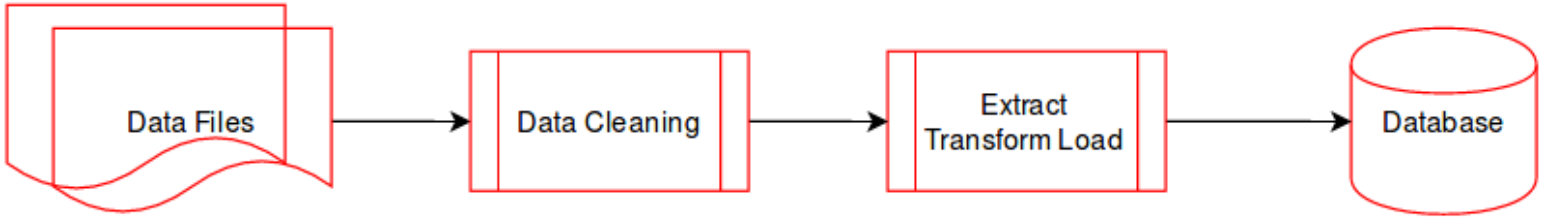


Figure: Data preprocessing

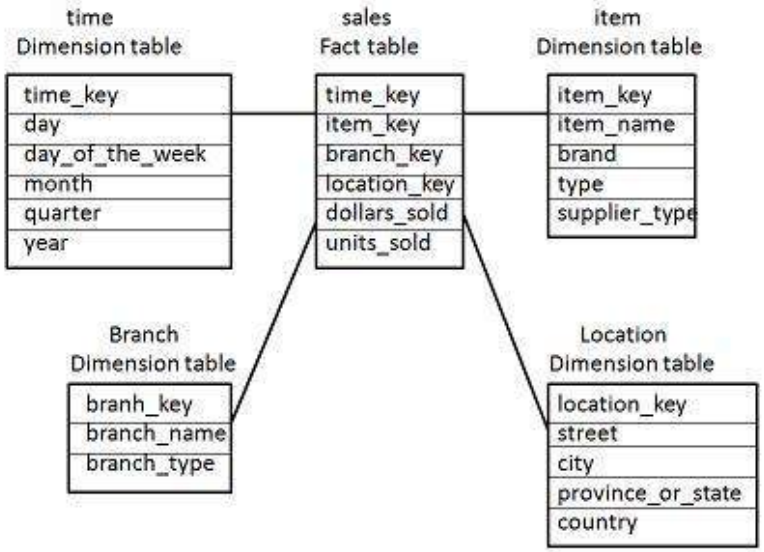


Figure: OLAP Star Schema

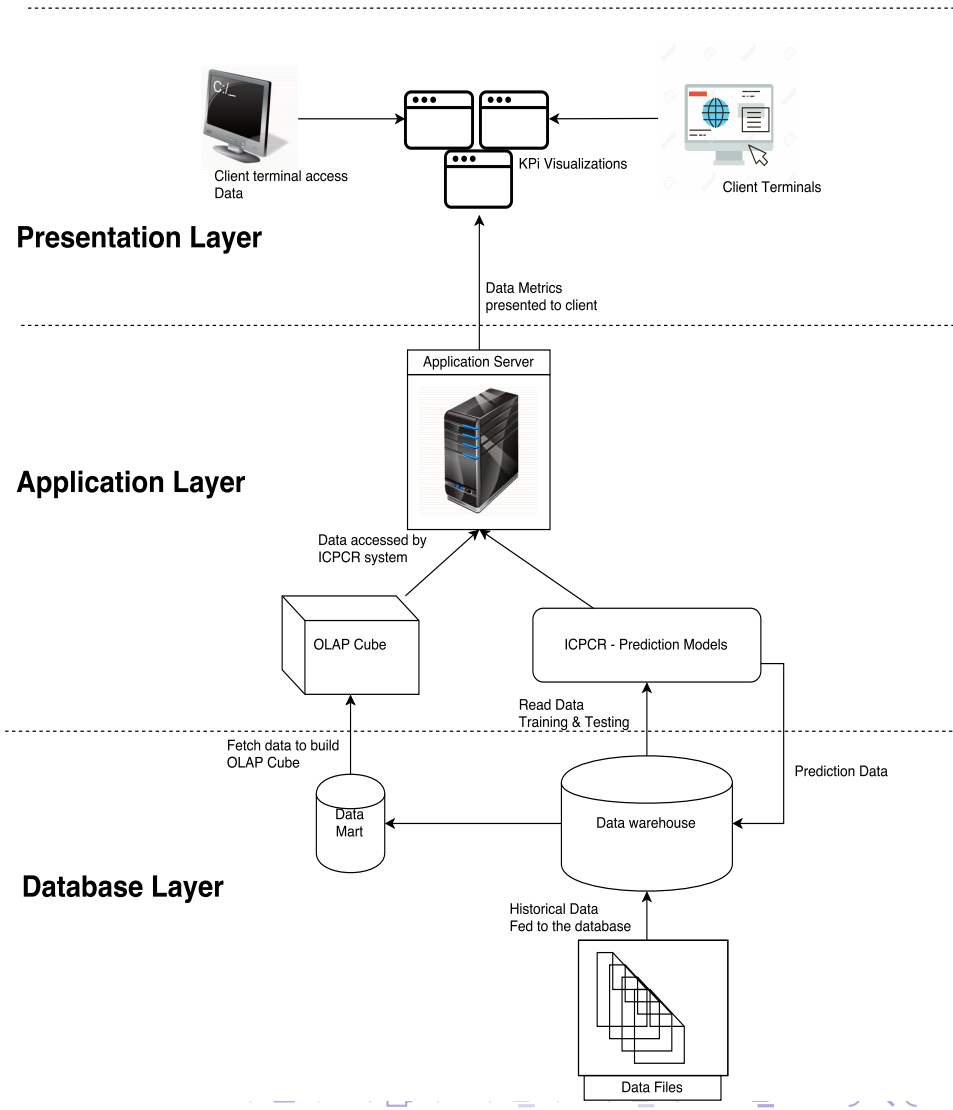
Development and Evaluation of the Prediction Models

- Model design
  - Proposal is to model 3 techniques based on Decision Tree, SVM, ANN
  - To use Machine learning libraries of either MLib, Scikit or R
  - Propose to implement a boosting algorithm Adaboost
- Model evaluation
  - K-fold cross validation technique
  - Confusion matrix with scoring of Sensitivity, Specificity, Precision and Recall, F-score

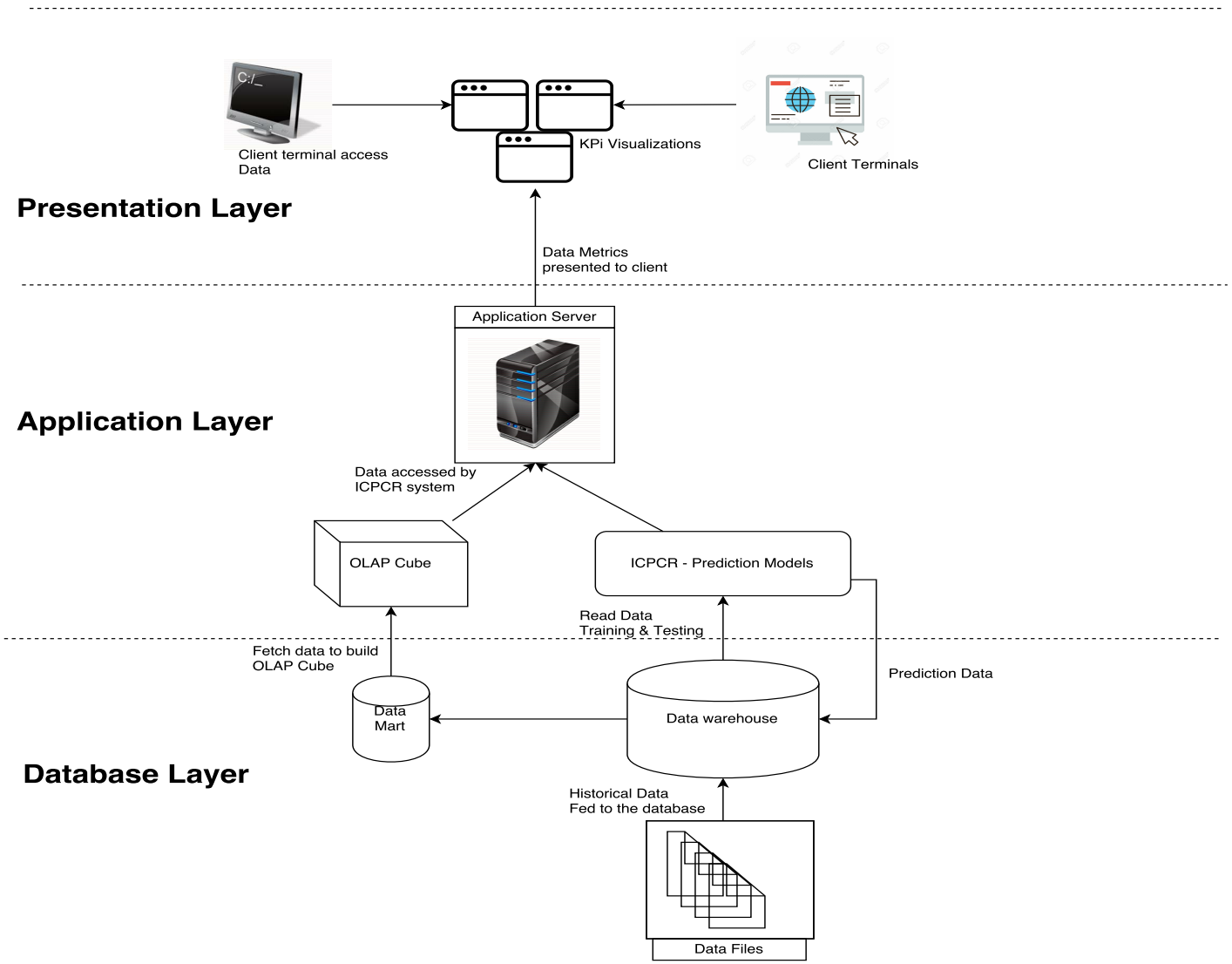
System Development & Evaluation

4 steps to be implemented.

- Presentation Layer
  - GUI for KPI's
  - Plots of predictions
- Application Layer
  - Application processing
  - Predictive model
  - OLAP cube
- Database Layer
  - Data warehouse tables in star schema
  - Data from prediction
- System Testing
  - Unit testing and Latency tests



The Intelligent Churn Prediction Architecture



Timeline

Name	Begin date	End date
Proposal	3/17/17	4/20/17
Data collection and Meta data evaluation	4/10/17	5/5/17
Data cleaning	4/24/17	5/12/17
Data warehouse design	5/15/17	6/23/17
Model selection	5/19/17	6/29/17
Model Training and Evaluation	6/15/17	7/26/17
Model Evaluation	7/3/17	8/11/17
Build ICPCR system	8/1/17	9/11/17
Integration with OLAP and Model	9/1/17	9/28/17
Dashboard development	10/2/17	11/10/17
System testing	10/20/17	11/30/17
Final defence	12/1/17	12/14/17