Faculty Research Day 2019

Nov 15th, 1:20 PM

How Unstructured Data from the Data Warehouse Can be Used with Machine Learning and Visualization to Develop Novel Medical Technologies

Alfred A. Cecchetti
Marshall University, cecchetti@marshall.edu

How Unstructured Data From The Data Warehouse Can Be Used With Machine Learning And Visualization To Develop Novel Medical Technologies

Alfred Cecchetti, PhD, MSc, MSc IS
Director, Division of Clinical Informatics (DCI)
Research Assistant Professor
Department of Clinical and Translational Sciences (DCTS)
Joan C. Edwards School of Medicine
1600 Medical Center Drive, Room 276
Huntington, WV 25701
Office Phone 304-691-1585

Marshall Informatics Platform

Multi-Institutional Data Storage

- *Structured Data*
- *Unstructured Data*
- *Validation*

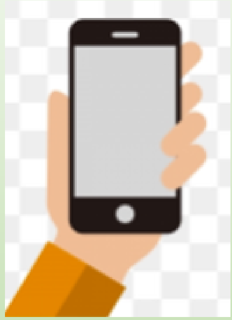
Machine Learning

- *Classification*
- *Prediction*
- *What if Scenarios*

Visualization

- *Data Microscope*
- *Drill Down/Drill Up*
- *Interactive Displays*

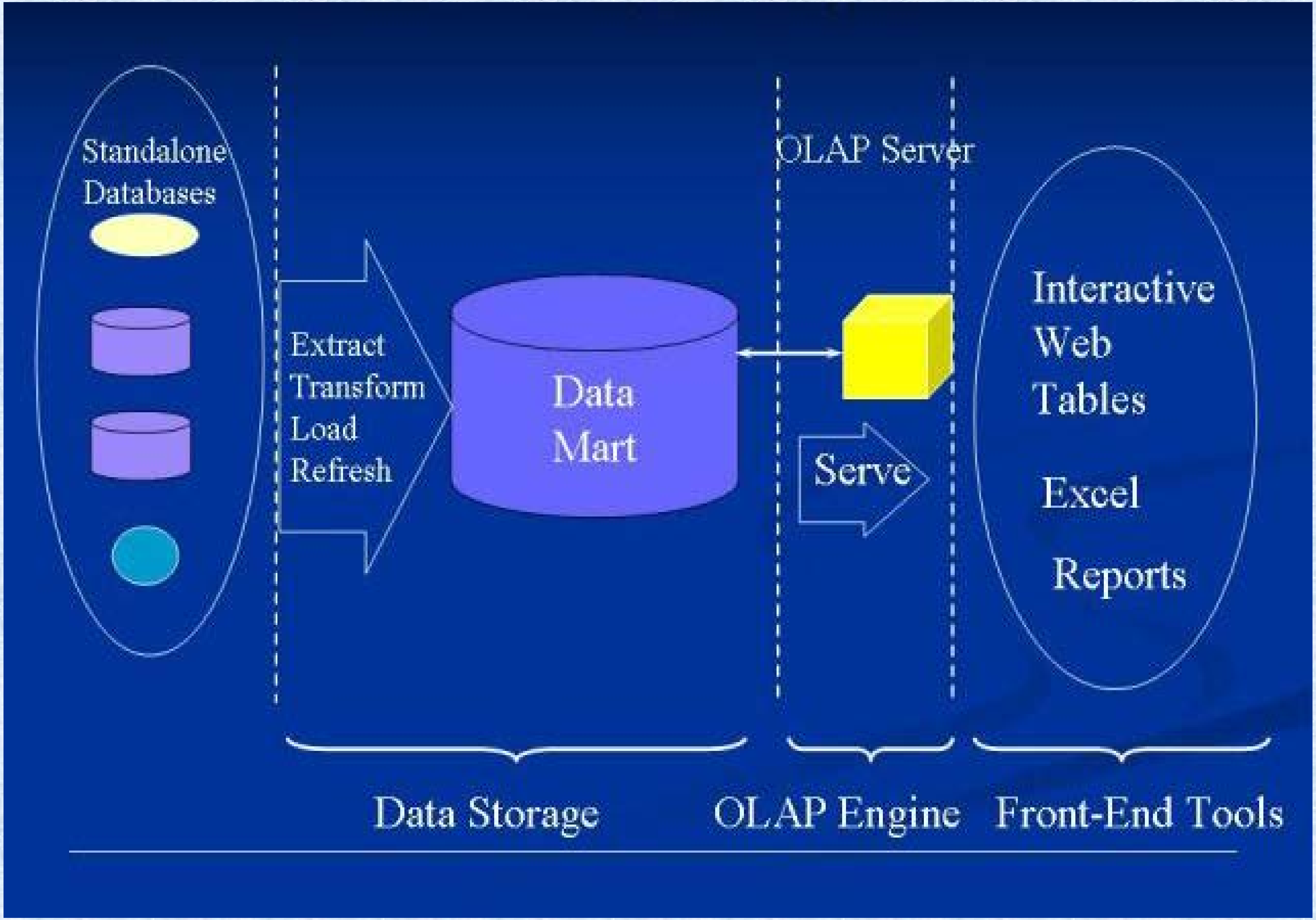
Programming



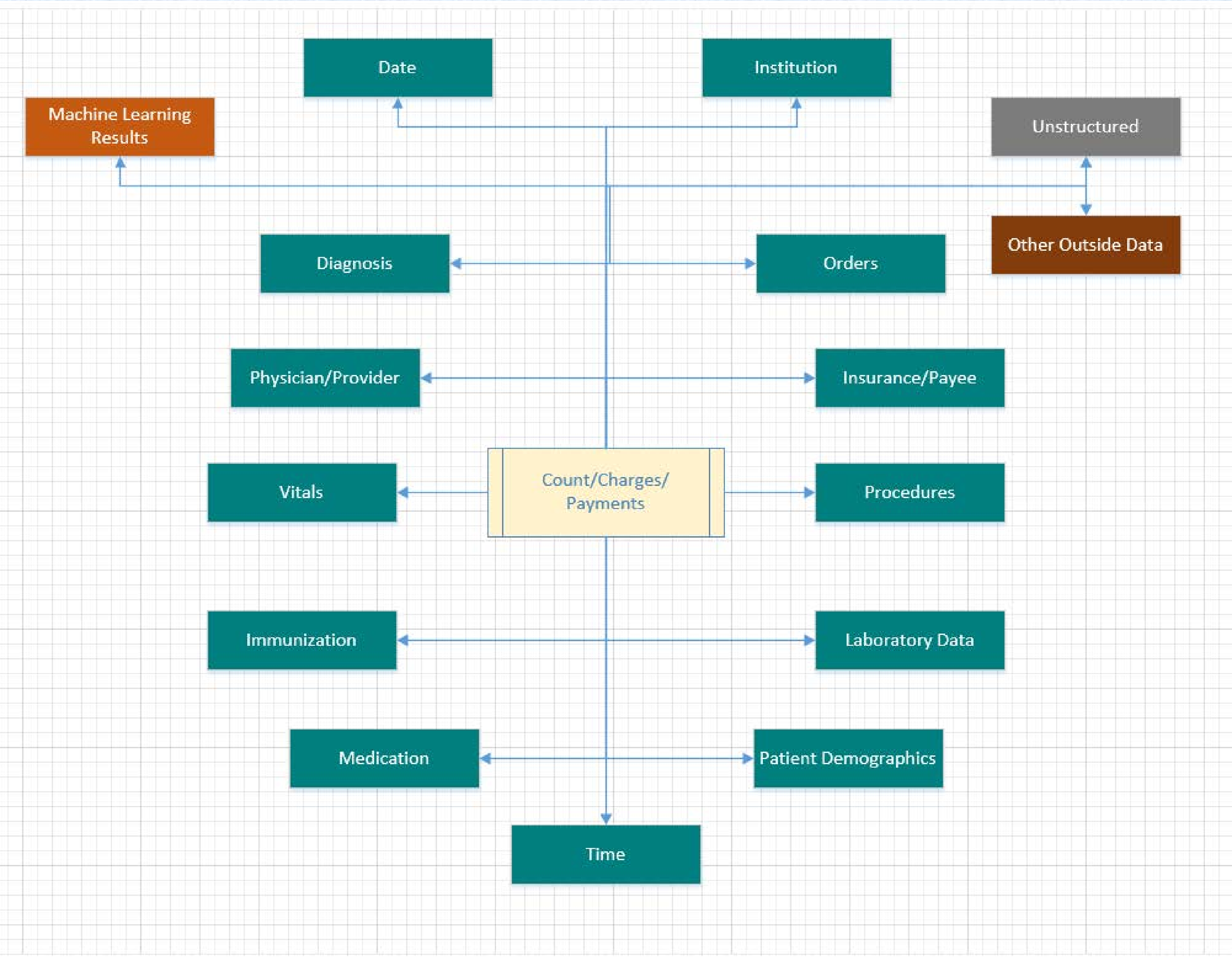
- ***Manage healthcare***
- ***Integrate systems***
- ***Healthcare Tools***

Data Warehouse

Combine Different Sources of Data



OLAP CUBE: Hub and Spoke Design



Access to Unstructured Data

The patient has lost appetite for a month anorexic gradually and she cannot tolerate meat she ***lost weight*** about 6-7 pounds for the past 2 months nonintentional

Sudden weight loss

Family history: ***Grandmother with colon cancer and daughter with lung cancer*** and metastasis Hypertension and diabetes from other side they called regarding the patient she is

Family history Cancer

OB/GYN: Hysterectomy at age of XX ***she is a mother of X kids X boys and X girls*** denies any bleeding and discharge

Mother with large family

Social history: She lives with her daughter and granddaughter, … ***4 dogs and 1 cat***, the patient reported HIV testing and she was negative

Pet Owner

Data Extraction

How do we extract Data SQL

```sql
SELECT DISTINCT * FROM
(SELECT A.[DIM_EMPI_VALUE],
        DATEDIFF(MONTH, C.[DOB], C.ADMITDATE) AS AGEATADMIT,
        [SEX],
        LEFT([ZIP],3) AS ZIP3,
        c.[ADMITDATE]
FROM    DevelopmentSource.dbo.AffinityDiagnosis A
JOIN    DevelopmentSource.dbo.DIM_NAMES B
ON      A.DIM_EMPI_VALUE = B.DIM_EMPI_VALUE
JOIN    DevelopmentSource.dbo.AffinityPROCEDURE C
ON      B.DIM_EMPI_VALUE = C.DIM_EMPI_VALUE
AND     CAST(A.ADMITDATE AS DATE) = CAST(C.ADMITDATE AS DATE)
WHERE   (ICDCode LIKE '466.1%' OR ICDCode LIKE 'J21%')
AND     (DeptName IN ('5 SOUTH - PEDIATRICS','CHH/MH PEDS','CFMC PED AFTER HOURS','PICU')))D
WHERE AGEATADMIT <= 24
AND CAST(ADMITDATE AS DATE) BETWEEN CAST('2015-07-01' AS DATE) AND  CAST('2017-06-30' AS DATE)
GO
```

Description: “Children 24 months or younger who had a diagnosis of Acute bronchiolitis and were admitted between the 1 July 2015 and 30 June 2017 in the specified departments.”

AGEATADMIT	SEX	ZIP3	ADMITDATE
23	F	255	2015-11-28 00:00:00.000
24	M	411	2015-12-30 00:00:00.000
23	M	412	2016-01-04 00:00:00.000
21	F	257	2015-12-17 00:00:00.000
22	F	255	2016-01-18 00:00:00.000
17	F	411	2015-09-29 00:00:00.000

Common Language Runtime User Defined Functions

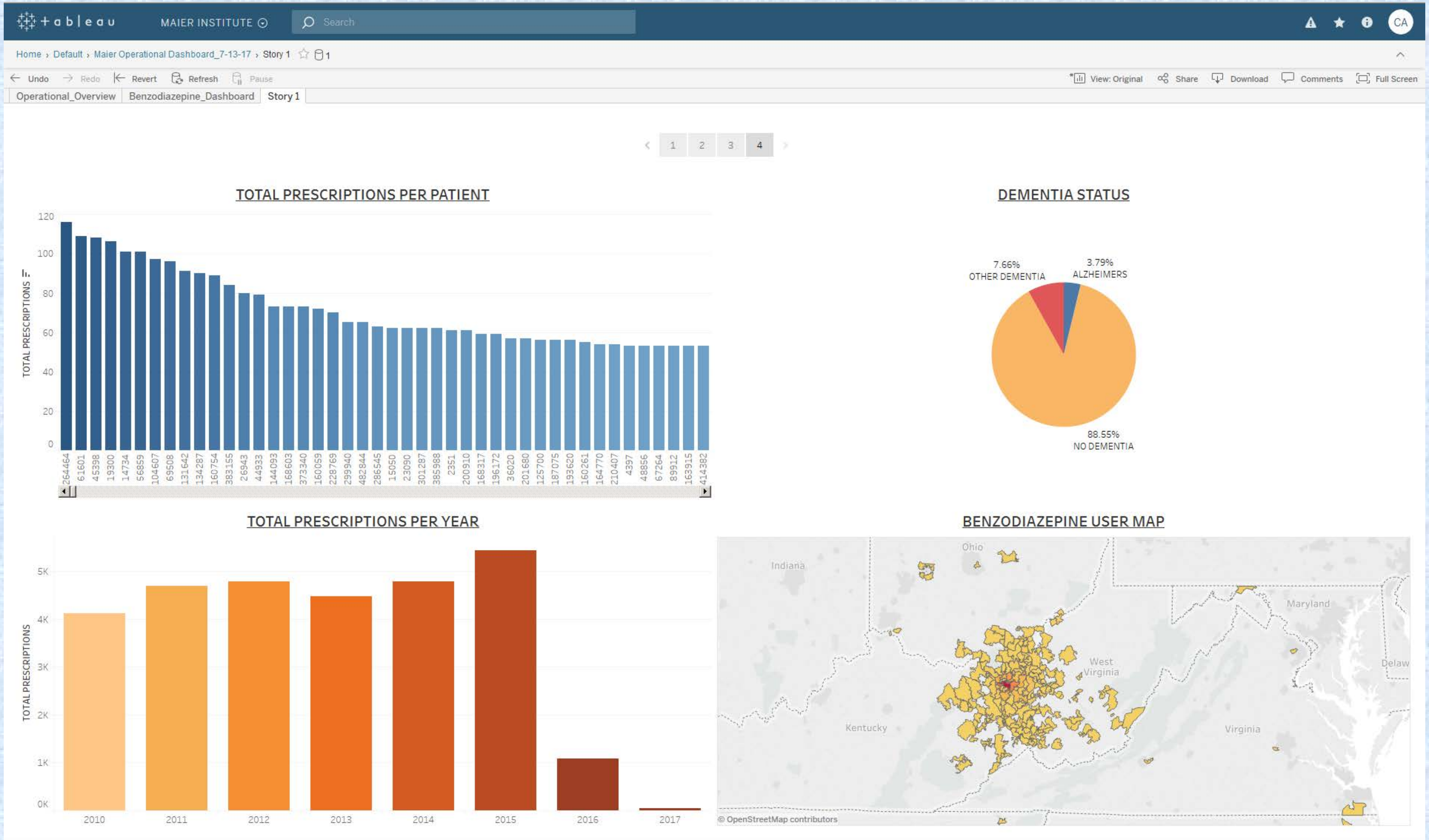
SqlFunction1.cs*

CLR1 | UserDefinedFunctions

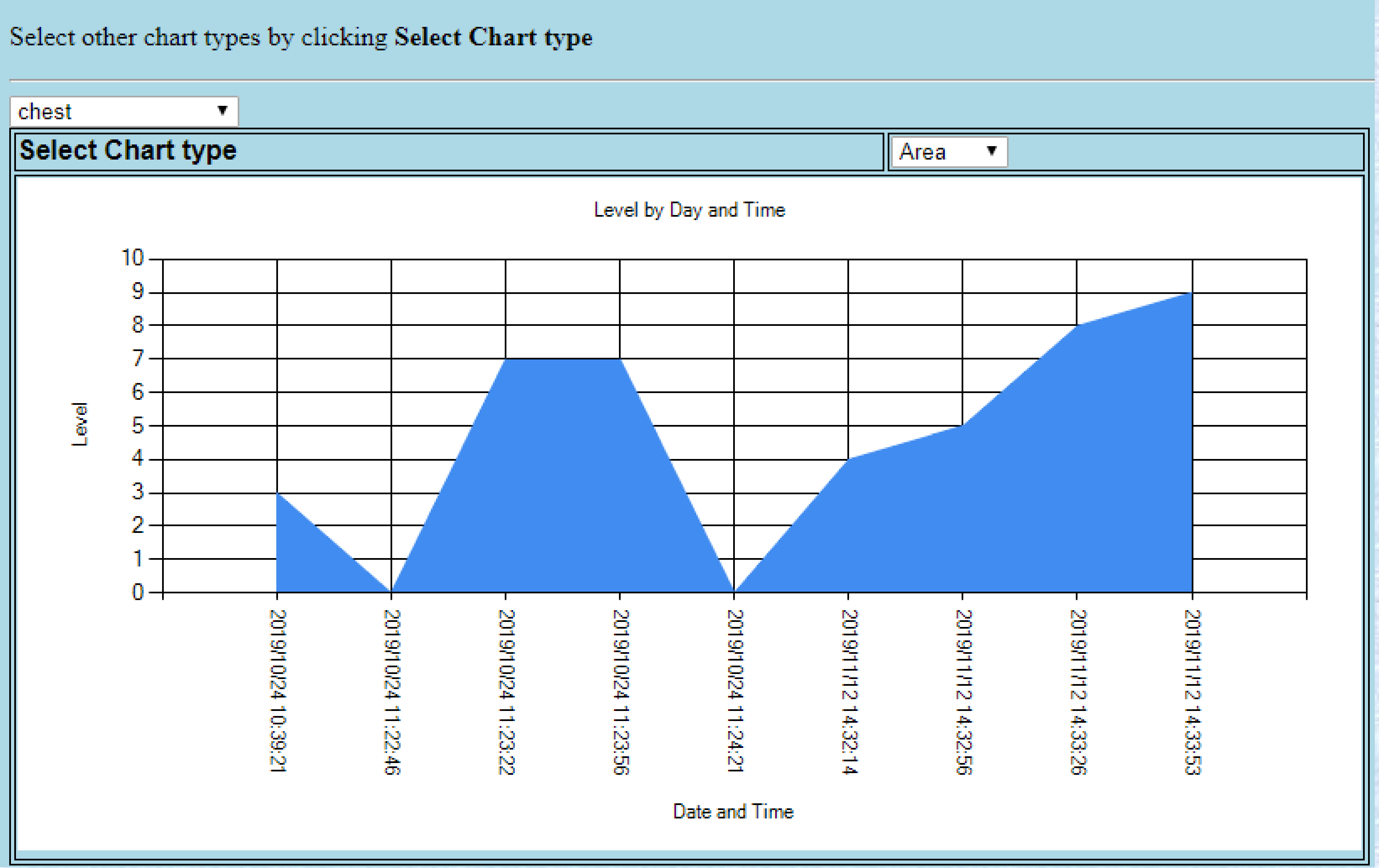
```csharp
using System;
using System.Data;
using System.Data.SqlClient;
using System.Data.SqlTypes;
using System.Text.RegularExpressions;
using Microsoft.SqlServer.Server;

public partial class UserDefinedFunctions
{

    [Microsoft.SqlServer.Server.SqlFunction]
    public static SqlString OutsideWords(string theposition, string mystring, string theword)
    {
        string outputstring = "";
        mystring = "aaa " + mystring.Trim() + " aaa";

        string pattern = @"(?<before>\w+) " + theword + @" (?<after>\w+)";
        MatchCollection matches = Regex.Matches(mystring, pattern);

        for (int i = 0; i < matches.Count; i++)
        {
            if (theposition == "before")
            {
                outputstring = outputstring + matches[i].Groups["before"].ToString();
                if (outputstring.Trim() == "aaa")
                {
                    outputstring = "";
                }
            } else
            {
                outputstring = outputstring + matches[i].Groups["after"].ToString();
                if (outputstring.Trim() == "aaa")
                {
                    outputstring = "";
                }
            }


        }

        return new SqlString(outputstring);
    }

}
```

Visualization

Historical Graphics



Real Time Graphics



Machine Learning

Why Machine Learning

Machine Learning (ML) can accurately classify and accurately predict disease as well as other medical events.

- Classifier models: Used for differential diagnosis, outcome prediction, etc.

- Regression models: patient survival, length of stay, laboratory values, etc.

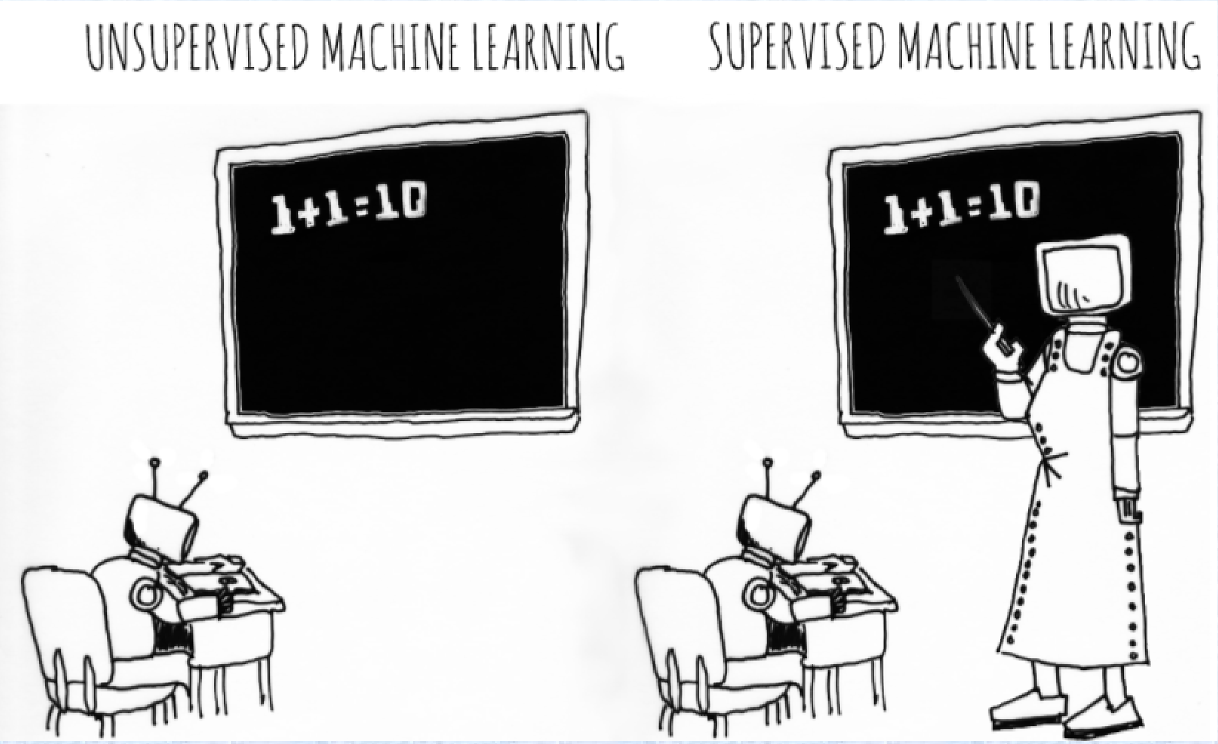
How do Computers Learn

Supervised learning

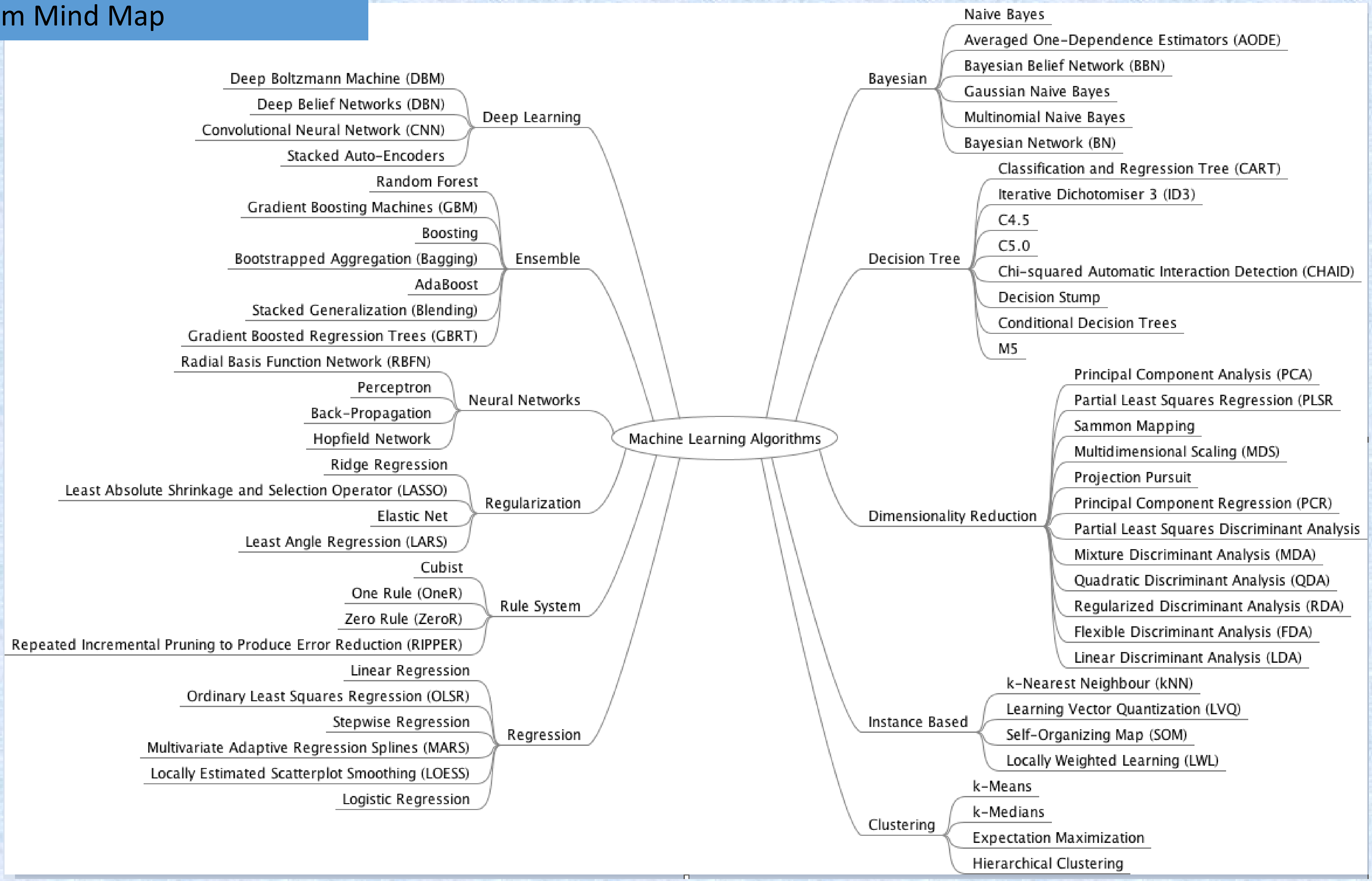
- Prediction
- Classification (discrete labels),
- Regression (real values)

Unsupervised learning

- Clustering
- Probability distribution estimation
- Finding association (in features)
- Dimension reduction



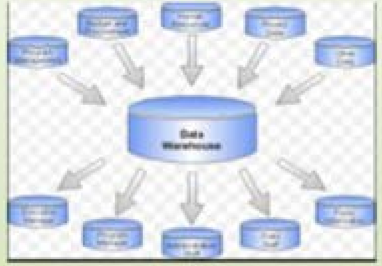
Algorithm Mind Map



Brownlee (2018). Welcome to Machine Learning Mastery: https://machinelearningmastery.com/

Machine Learning Pipeline

Multi-Institutional Data Storage



- *Structured Data*
- *Unstructured Data*
- *Validation*

DATA

MODEL

Machine Learning

- *Classification*
- *Prediction*
- *What if Scenarios*

Embed Machine Learning in SQL

SCRIPT USE THIS ge...\cecchetti-a (84))*

```
GO

--(@xmodel varbinary(max) OUTPUT)

create procedure dbo.generate_lung_cancer_model1

AS
BEGIN
EXECUTE sp_execute_external_script
  @language = N'R'
 ,@script   = N'
library(RevoScaleR)
library(caret) # show me all the packages in caret # names(getModelInfo())
library(RANN)
library(randomForest)
library(RODBC)
library(doSNOW)
library(quanteda)
library(parallel)


gc()

sms_raw$TYPE <- toupper(as.factor(sms_raw$TYPE))
sms_raw$TEXT <- as.character(sms_raw$TEXT)   ## use it all

train.tokens <- tokens(sms_raw$TEXT , what = "word",
                       remove_numbers = TRUE, remove_punct = TRUE,
                       remove_symbols = TRUE, remove_hyphens = TRUE,
                       remove_url = TRUE)

train.tokens <- tokens_tolower(train.tokens)


### get multiword

multiword <- c("you are","yellow","without *","without","wheezing","went away","weight loss","weight","weakness","weak","warm","want to","vomit

## have multiword

train.tokens <- tokens_compound(train.tokens, pattern = phrase(multiword))
train.tokens <- tokens_select(train.tokens, stopwords(),selection = "remove")

train.tokens <- tokens_wordstem(train.tokens, language = "english")



train.tokens.dfm <- dfm(train.tokens) # bag of words model- create a document feature matrix
train.tokens.dfm <- dfm_trim(train.tokens.dfm, min_docfreq = 40)

trained_model <- data.frame(payload = as.raw(serialize(train.tokens.dfm , connection=NULL)))'
```

Programming

Why Programming

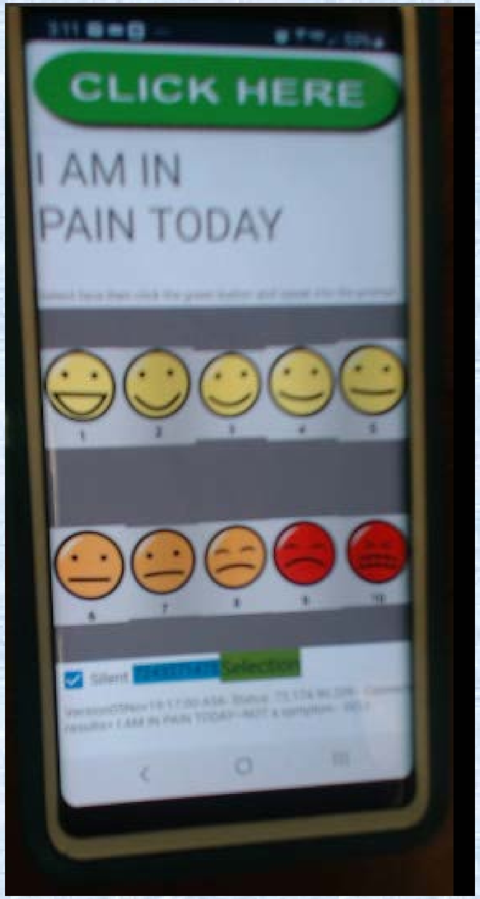
Device Programming, especially smartphone applications, can provide new ways to acquire, transport, store, process, and secure personalized patient data to deliver meaningful results.

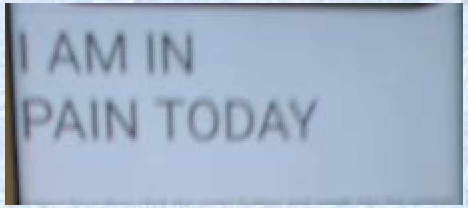
An Example

Extraction of Baseline Data From Hospital Notes

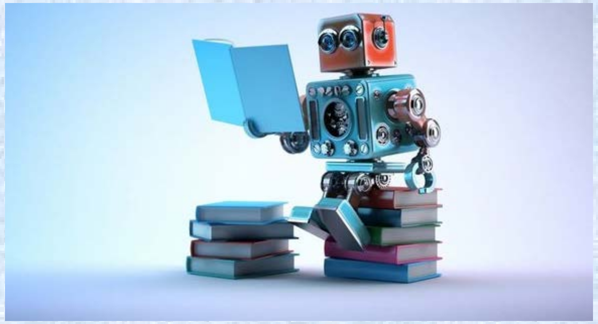
Statement	Symptoms Present	Symptoms Not Present
Patient says he is feeling fatigue for the last 3-4 months	Fatigue	
He has lower abdominal cramping 3 x weekly	Abdominal cramping	
Patient states episodes of nausea	Nausea	
Patient denies heartburn		Heartburn
Patient denies fever		Fever
Patient denies chills		chills

Text mining



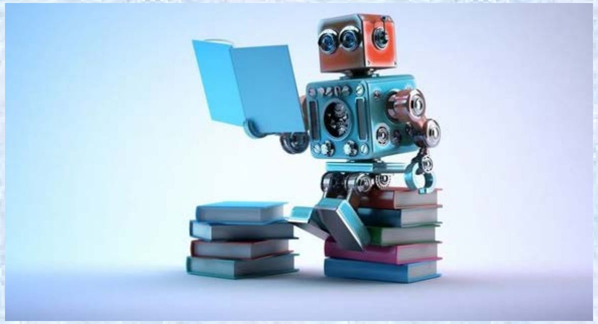


Patient at Baseline



	manageable	little	extreme	pain	committing	suicide	sore	hurts	number	date
Patient 1	manageable			pain				hurts	**5**	21-May-18
Patient 2		little		pain			sore	hurts	4	21-May-18

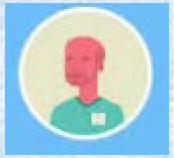
Next Day

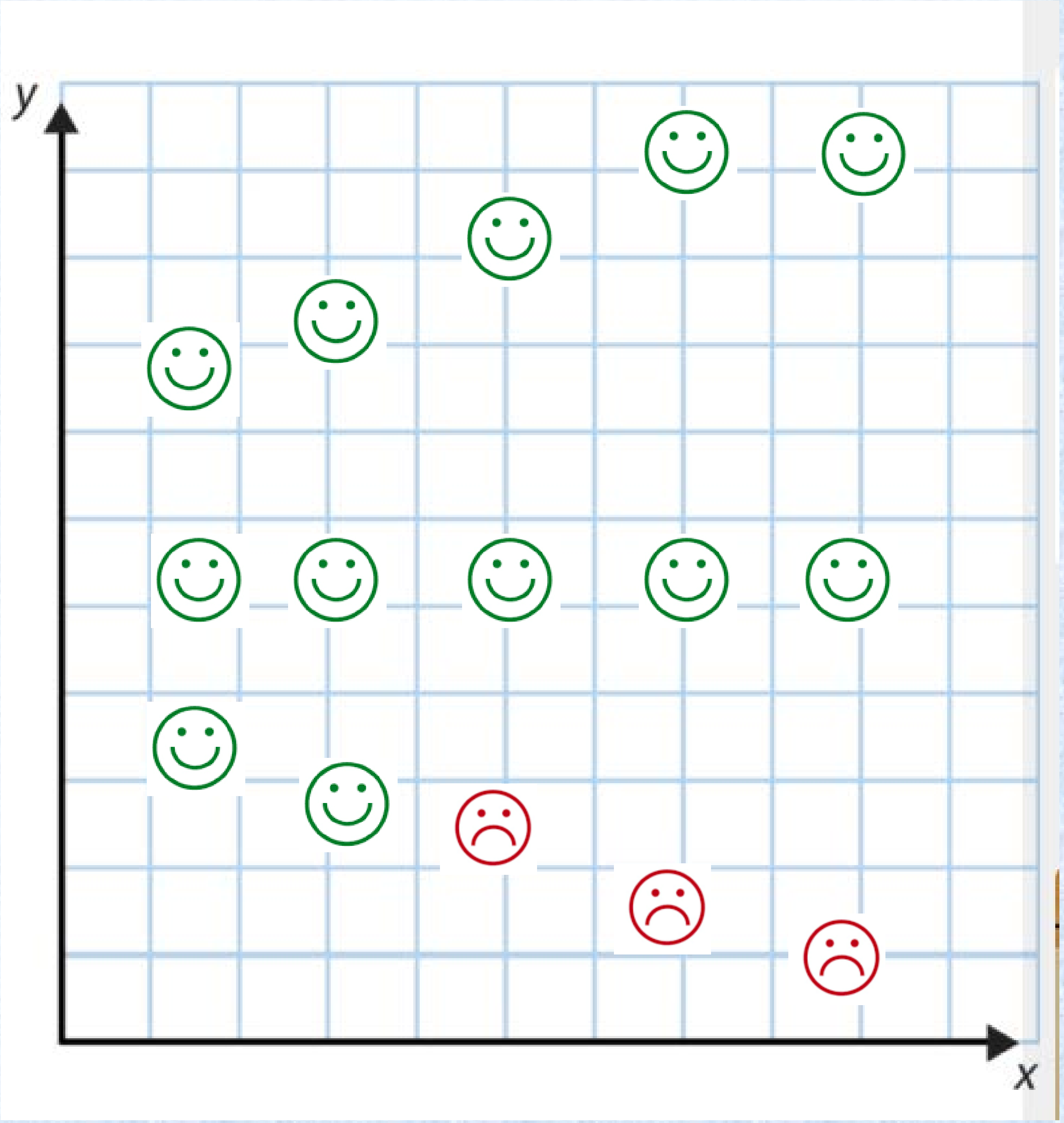


	manageable	little	extreme	pain	committing	suicide	sore	hurts	number	date
Patient 1			extreme	pain		suicide		hurts	**9**	22-May-18
Patient 2		little		pain			sore	hurts	**5**	22-May-18

Outcome

Good

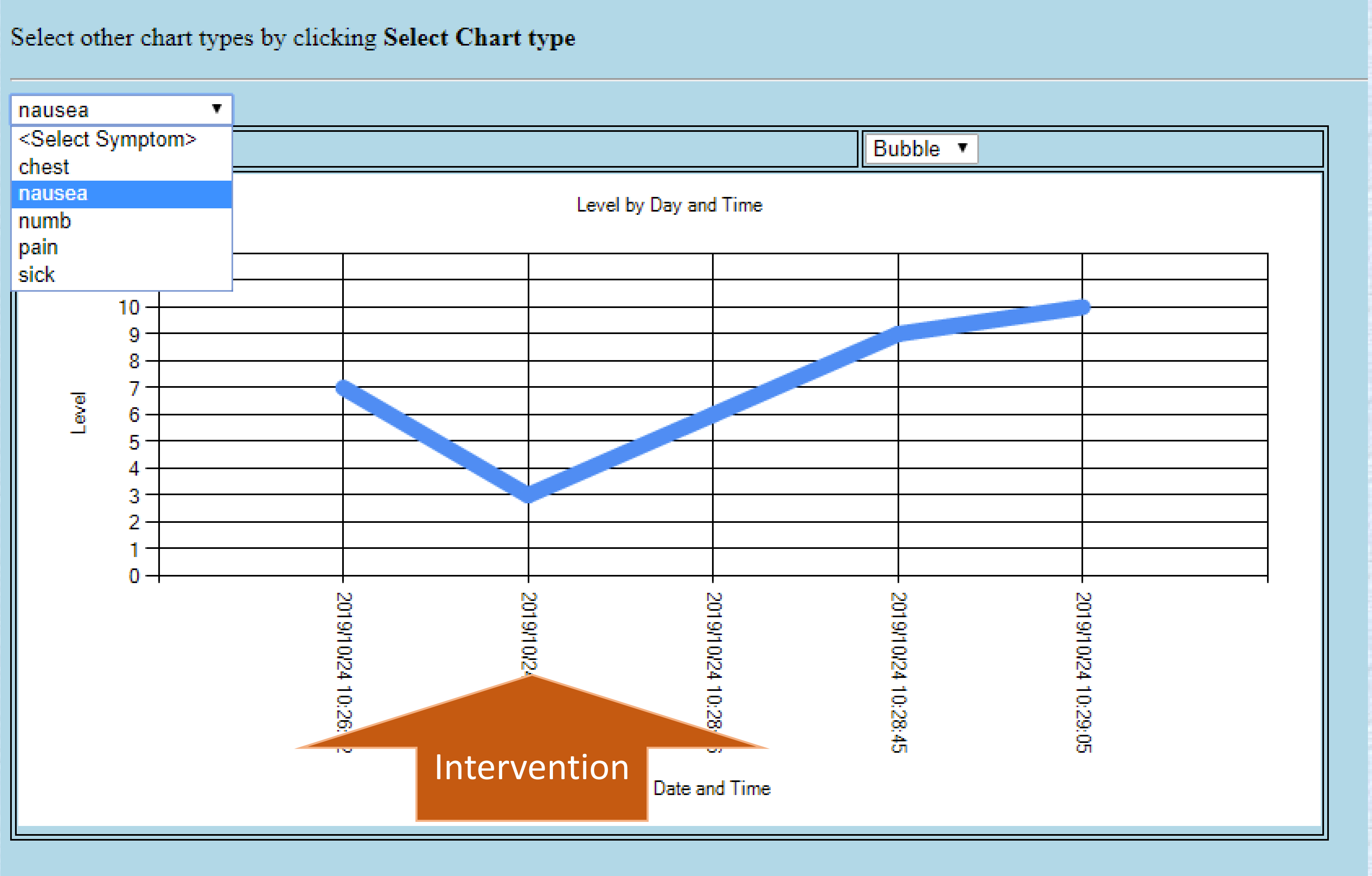




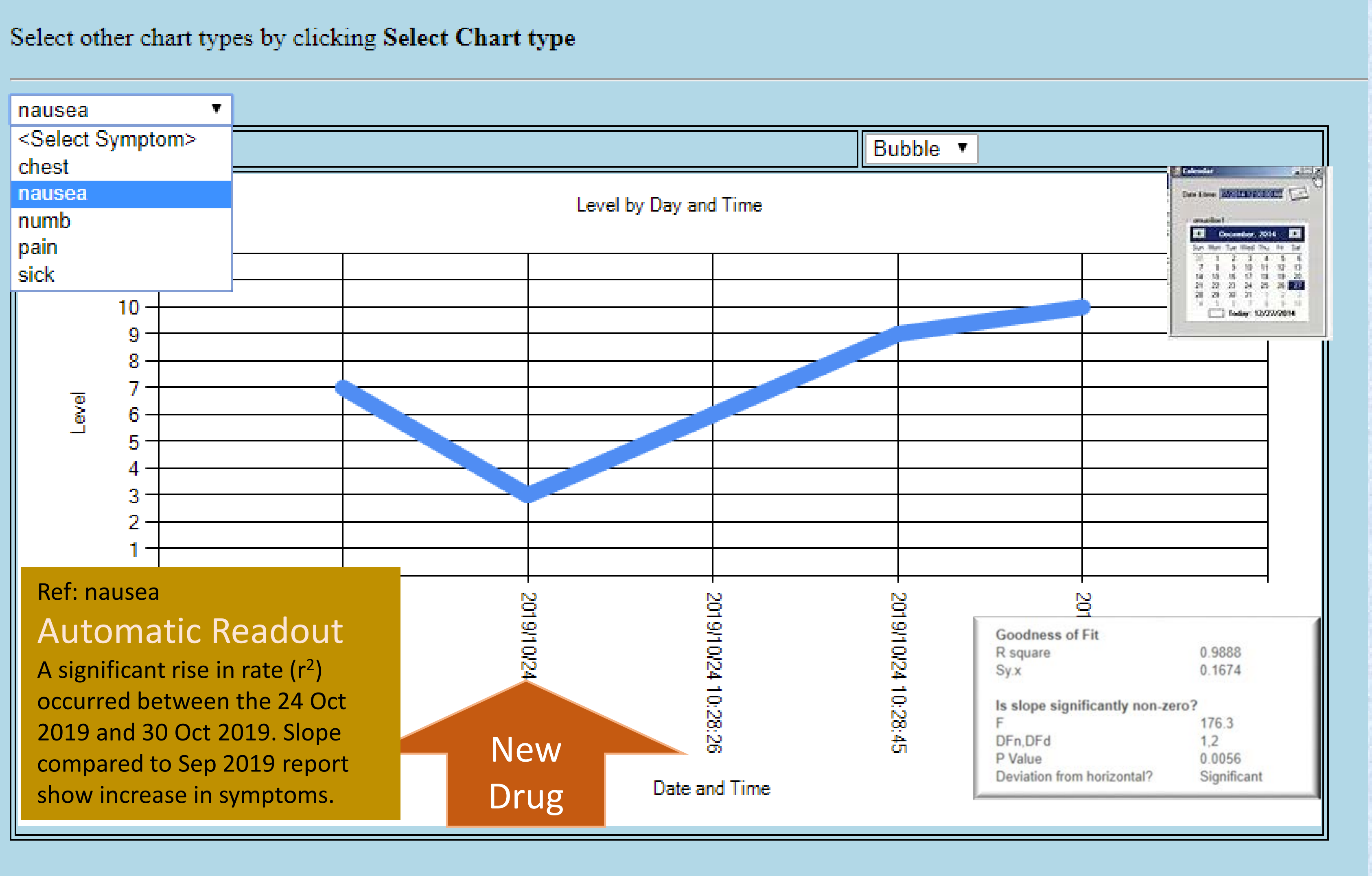
Bad

Time

What Happens During An Intervention



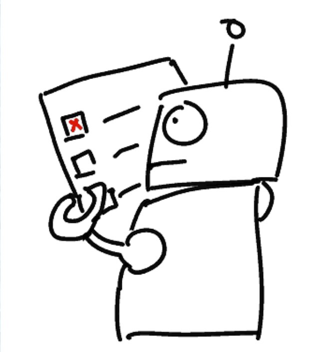
Analyze the Intervention



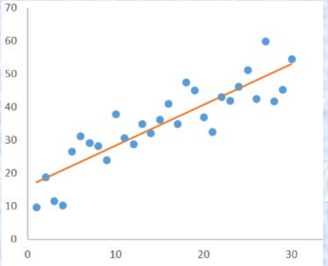
Develop Novel Medical Technologies for Specific Chronic Diseases or Events

Unstructured and Structured data is gathered

Sent to a machine learning algorithm



a data model to predict trends is created



Trends are interpreted as a simplified readout

How Can Novel Medical Technologies Benefit The Appalachian Community

- Remote individuals can now participate in the health care value matrix with minimal costs in ways not possible in the past.

- Algorithms, developed by Appalachian medical experts, can provide standardized guidance for specific chronic conditions at little or no cost.

END