R & Tabular Data Analysis

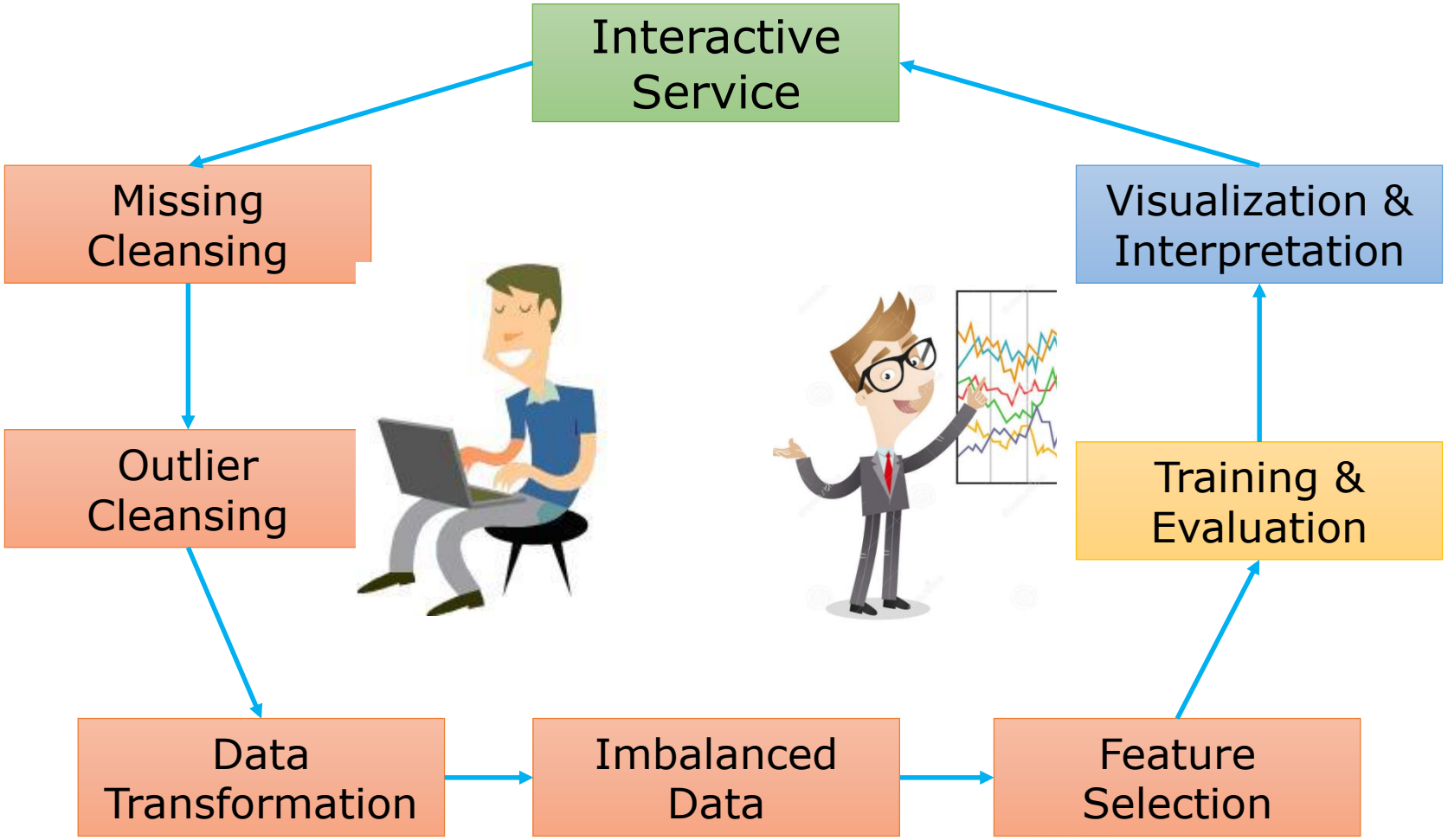
史春奇

2016/12

Agenda

- 交互式分析流程
  - Missing Cleansing
  - Outlier Cleansing
  - Data Transformation
  - Imbalanced Data
  - Feature Selection
  - Training & Evaluation
  - Visualization & Interpretation
  - Interactive Service

交互式快速迭代


Interactive Service
Missing Cleansing
Outlier Cleansing
Data Transformation
Imbalanced Data
Feature Selection
Training & Evaluation
Visualization & Interpretation

Missing Cleansing 算法

- 忽略Missing所在的Record
- 拟合Missing值
  - Mean / Median
  - Impute / Regression
- Missing算法
  - Surrogate：CART树，随机森林
  - expectation-maximization with bootstrapping (EMB)

Missing Cleansing 实现

- 替换字符含义的Missing

```
> x <- c("a", "B", NA, "NA")
> is.na(x)
[1] FALSE FALSE  TRUE FALSE
> x[x=="NA"] <- NA
> is.na(x)
[1] FALSE FALSE  TRUE  TRUE
>
```

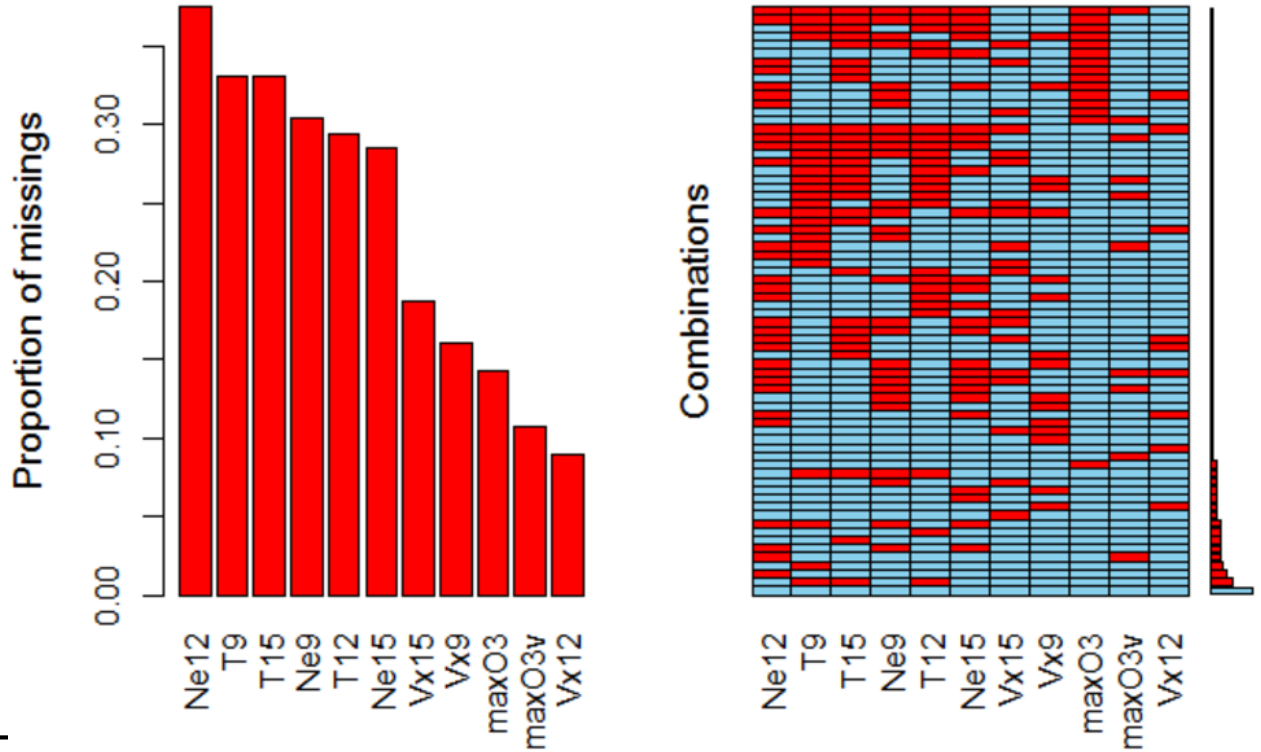
- 用均值中值回归Impute

mice :
Hmisc: Fisher's optimum scoring
imputeTestbench :

VIM: missForest: Amelia: mi:

Missing Cleansing 实现

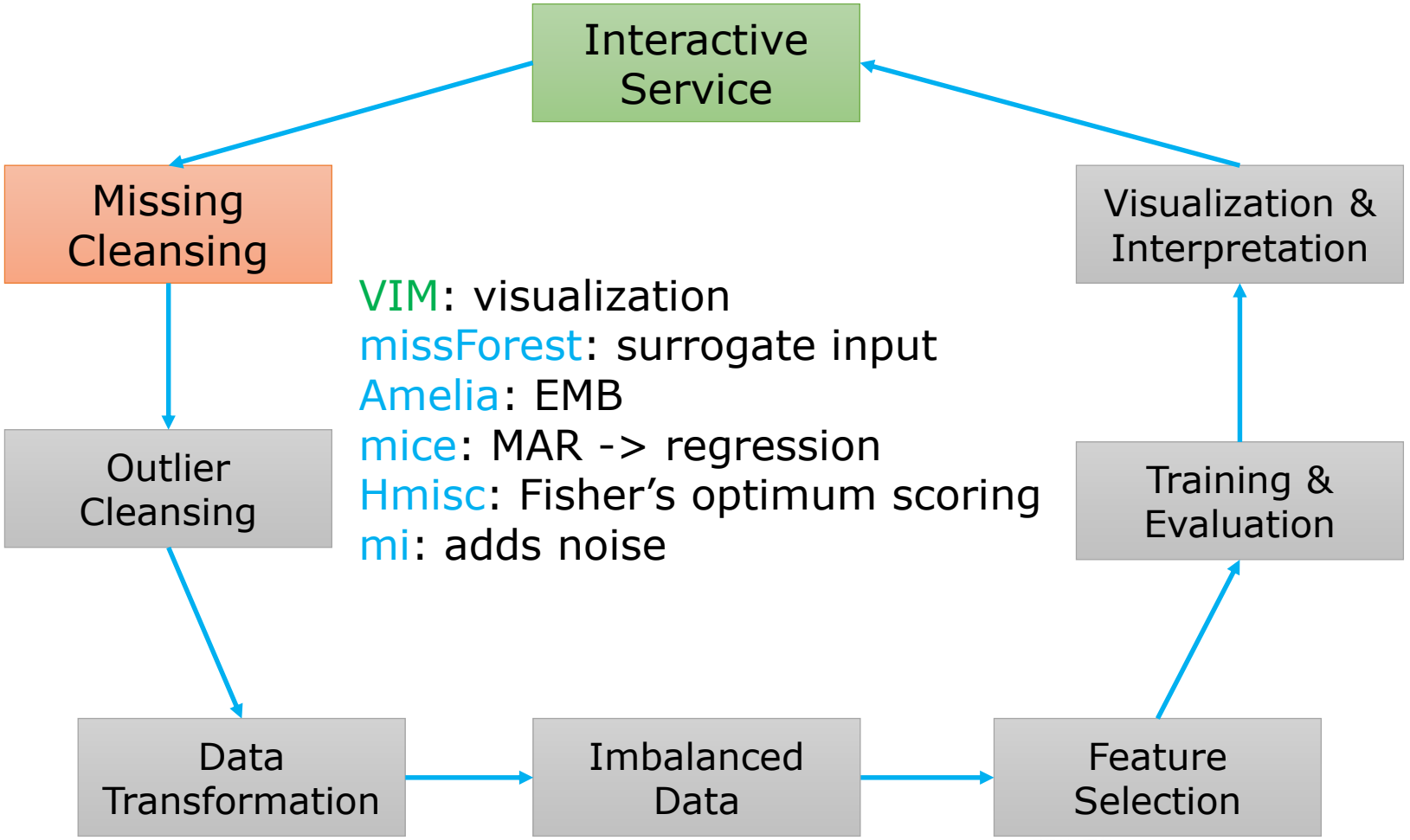
- 可视化
  - VIM
  - aggr(x, delimiter = NULL, plot = TRUE, ...)



Missing Cleansing 实现

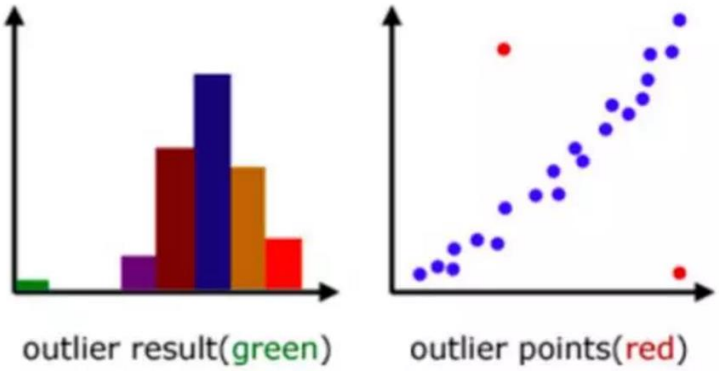
- Surrogate
  - missForest
- EMB: expectation-maximization with bootstrapping
  - Amelia
- Noised Impute
  - mi

Missing Cleansing

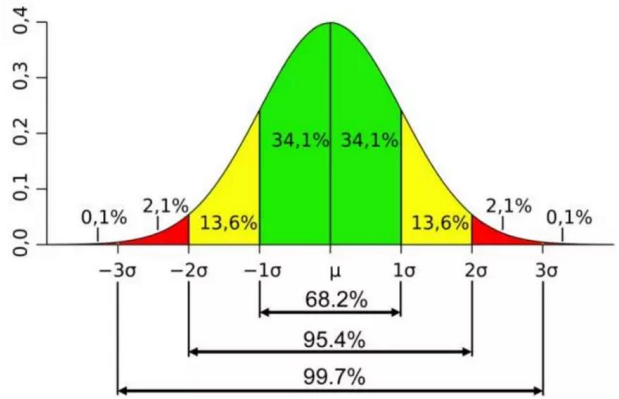


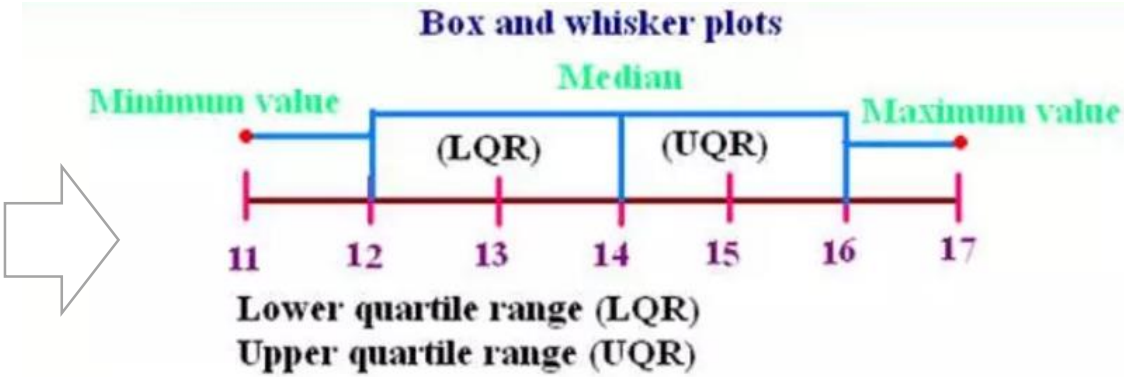
Outlier Cleansing

- outlier是合理的(explainable)小概率事件(rare)
- anomaly是不合理的小概率事件。



- Box-n-Whisker Plot






```
install.packages("extremevalues")
library(extremevalues)
```

Outlier Cleansing 实现

- Grubbs' test or Extreme Studentized Deviate (ESD) test

$$G = \frac{\max_{i=1,\ldots,N} |Y_i - \bar{Y}|}{s}$$

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

  - AnomalyDetection : Seasonal Hybrid ESD (S-H-ESD)

```
install.packages("devtools")
devtools::install_github("twitter/AnomalyDetection")
library(AnomalyDetection)

install.packages("outliers")
library(outliers)
```


Outlier Cleansing 实现

- Chi-squared test
  - Categorical data

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

```
install.packages("outliers")
library(outliers)

install.packages("mvoutlier")
library(mvoutlier)
```


Outlier Cleansing 实现

- Dixon's Q test
  - Non-parameter

$$Q = \frac{|x_{suspect} - x_{nearest}|}{|x_{max} - x_{min}|}$$

Number of values:	3	4	5	6	7	8	9	10
$Q_{90\%}$:	0.941	0.765	0.642	0.560	0.507	0.468	0.437	0.412
$Q_{95\%}$:	0.970	0.829	0.710	0.625	0.568	0.526	0.493	0.466
$Q_{99\%}$:	0.994	0.926	0.821	0.740	0.680	0.634	0.598	0.568

```
install.packages("outliers")
library(outliers)
```


Outlier Cleansing 实现

- Cochran's Q test
  - McNema's test generalization

$$T = k\,(k-1)\,\frac{\sum_{j=1}^{k}\left(X_{\bullet j}-\frac{N}{k}\right)^2}{\sum_{i=1}^{b}X_{i\bullet}\,(k-X_{i\bullet})} \qquad T > \chi^2_{1-\alpha,k-1}$$

```
install.packages("outliers")
library(outliers)
```


Outlier Cleansing 实现

- Local Outlier Factor
  - Concept from DBSCAN

$$\text{reachability-distance}_k(A, B) = \max\{\text{k-distance}(B), d(A, B)\}$$

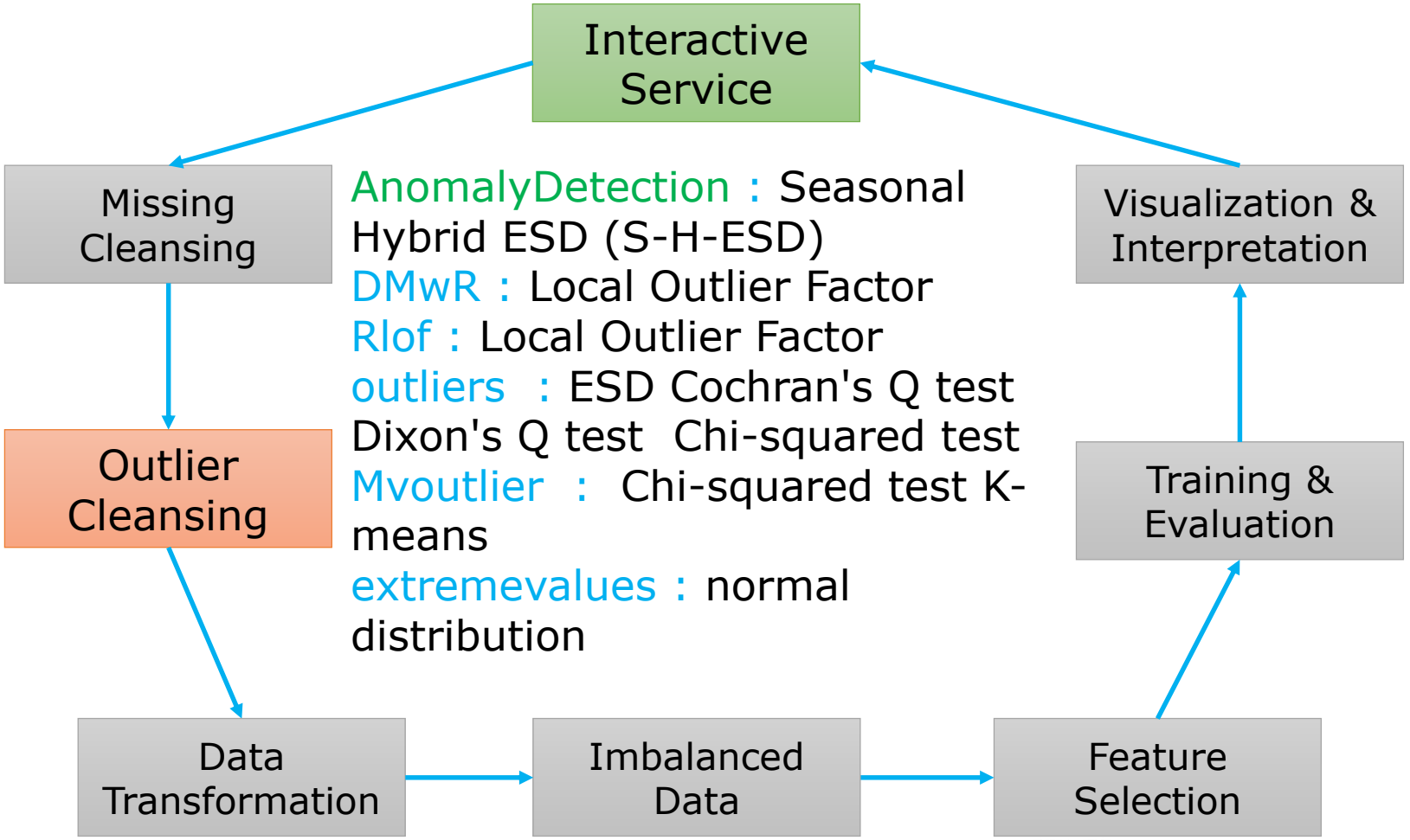
$$\text{lrd}(A) := 1/\left(\frac{\sum_{B \in N_k(A)} \text{reachability-distance}_k(A, B)}{|N_k(A)|}\right)$$

$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}(B)}{|N_k(A)|} / \text{lrd}(A)$$

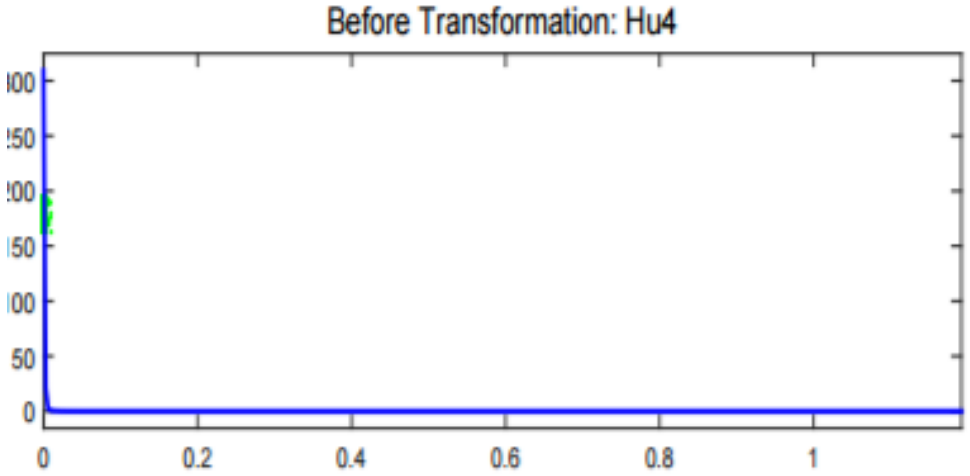
install.packages("DMwR") Data Mining With R
library(DMwR)

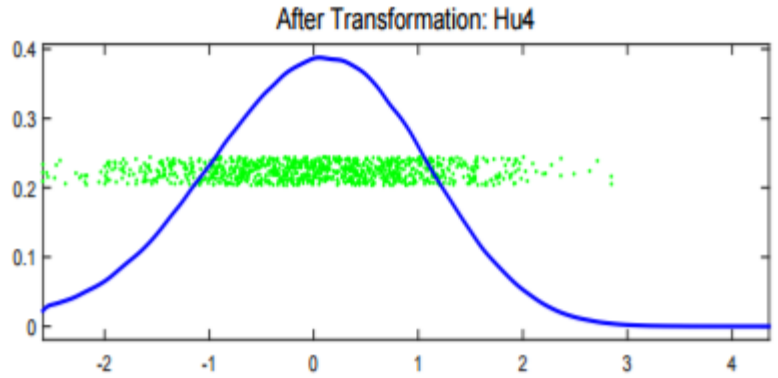
outlier.scores <- lofactor(dataframe, k=5)

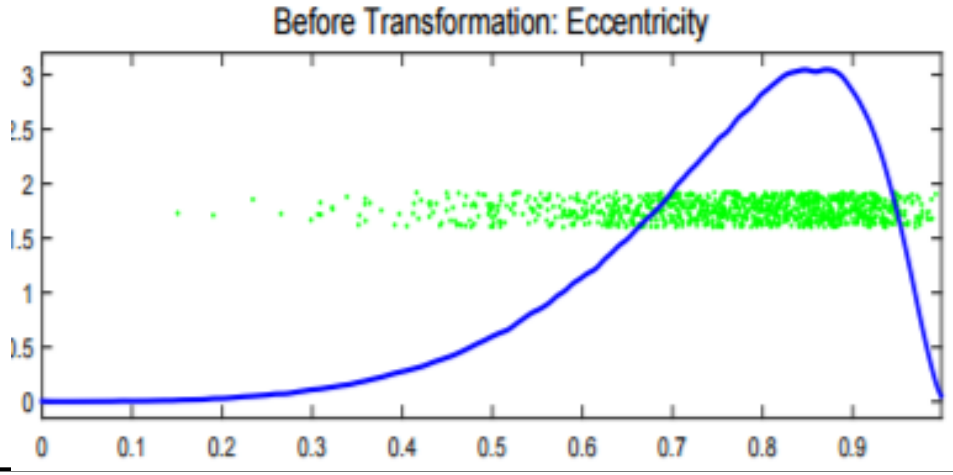
Outlier Cleansing

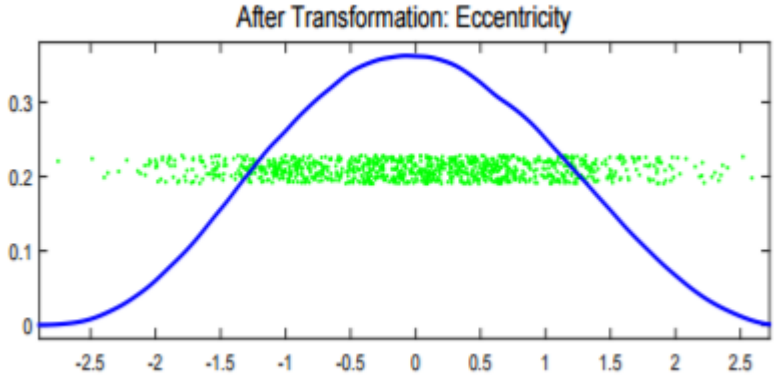


Data Transformation





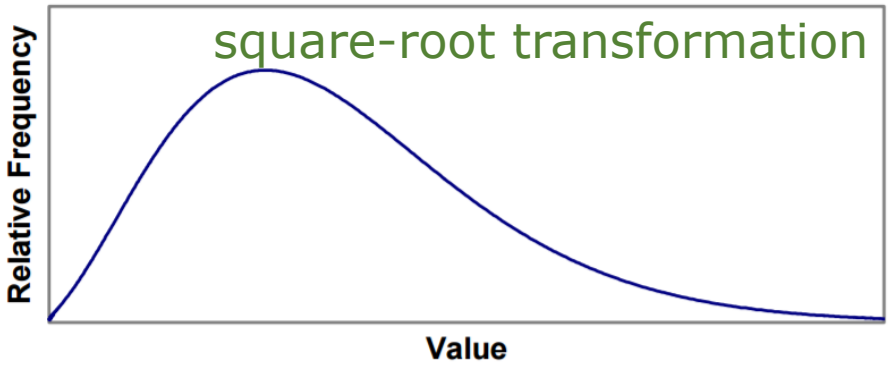


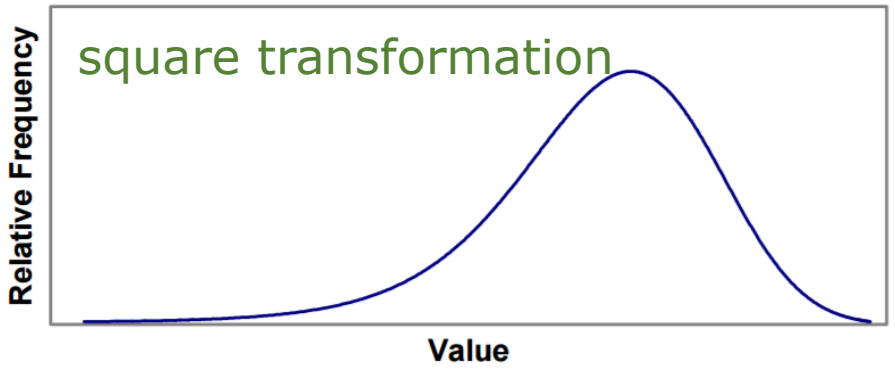


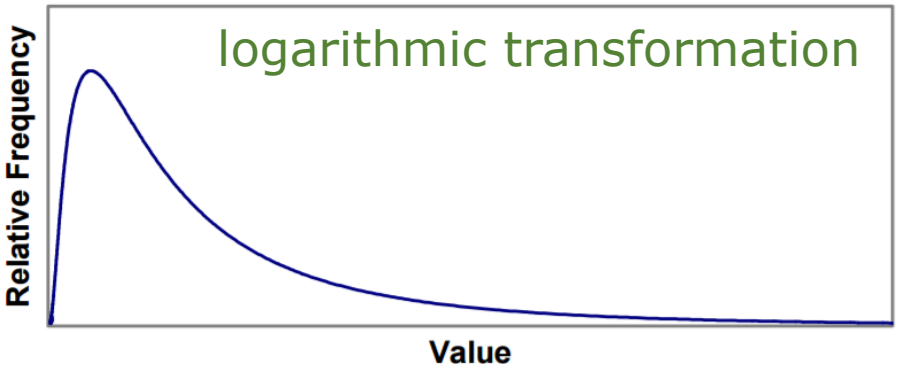
Data Transformation

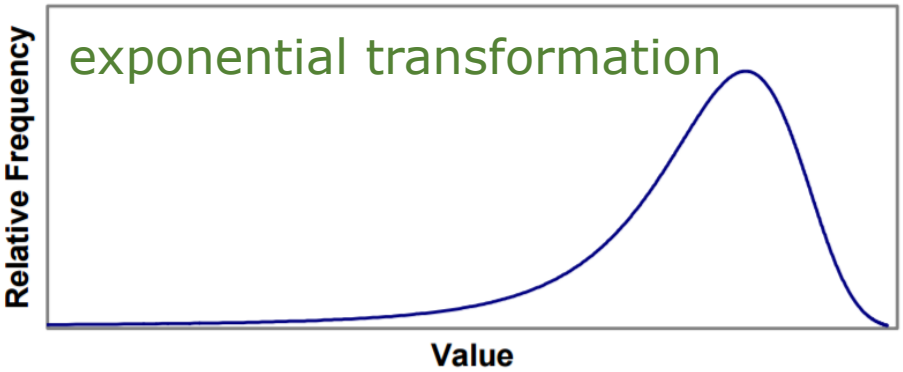
Method	Transformation(s)	Regression equation	Predicted value ($\hat{y}$)
Standard linear regression	None	$y = b_0 + b_1x$	$\hat{y} = b_0 + b_1x$
Exponential model	Dependent variable = log(y)	$\log(y) = b_0 + b_1x$	$\hat{y} = 10^{b_0 + b_1x}$
Quadratic model	Dependent variable = sqrt(y)	$\text{sqrt}(y) = b_0 + b_1x$	$\hat{y} = (b_0 + b_1x)^2$
Reciprocal model	Dependent variable = 1/y	$1/y = b_0 + b_1x$	$\hat{y} = 1 / (b_0 + b_1x)$
Logarithmic model	Independent variable = log(x)	$y= b_0 + b_1\log(x)$	$\hat{y} = b_0 + b_1\log(x)$
Power model	Dependent variable = log(y) Independent variable = log(x)	$\log(y)= b_0 + b_1\log(x)$	$\hat{y} = 10^{b_0 + b_1\log(x)}$

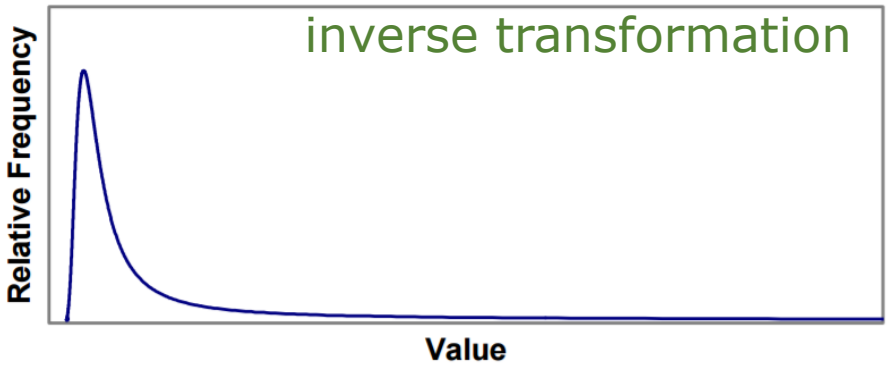
Data Transformation











Data Transformation 实现

```
data.dat$trans_Y <- (data.dat$Y)^3
data.dat$trans_Y <- (data.dat$Y)^(1/9)
data.dat$trans_Y <- log(data.dat$Y)
data.dat$trans_Y <- log10(data.dat$Y)
data.dat$trans_Y <- exp(data.dat$Y)
data.dat$trans_Y <- sqrt(data.dat$Y)
```


Data Transformation

Interactive Service

Missing Cleansing

Outlier Cleansing

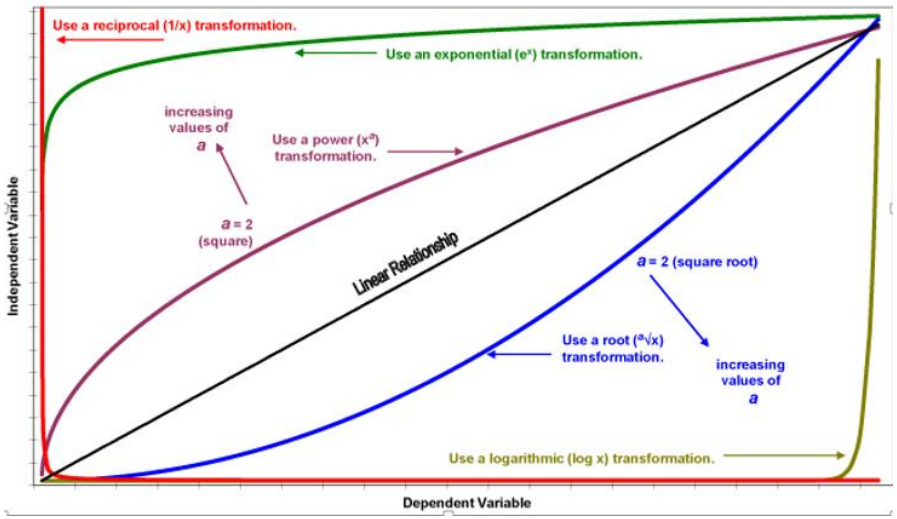
Data Transformation

Imbalanced Data

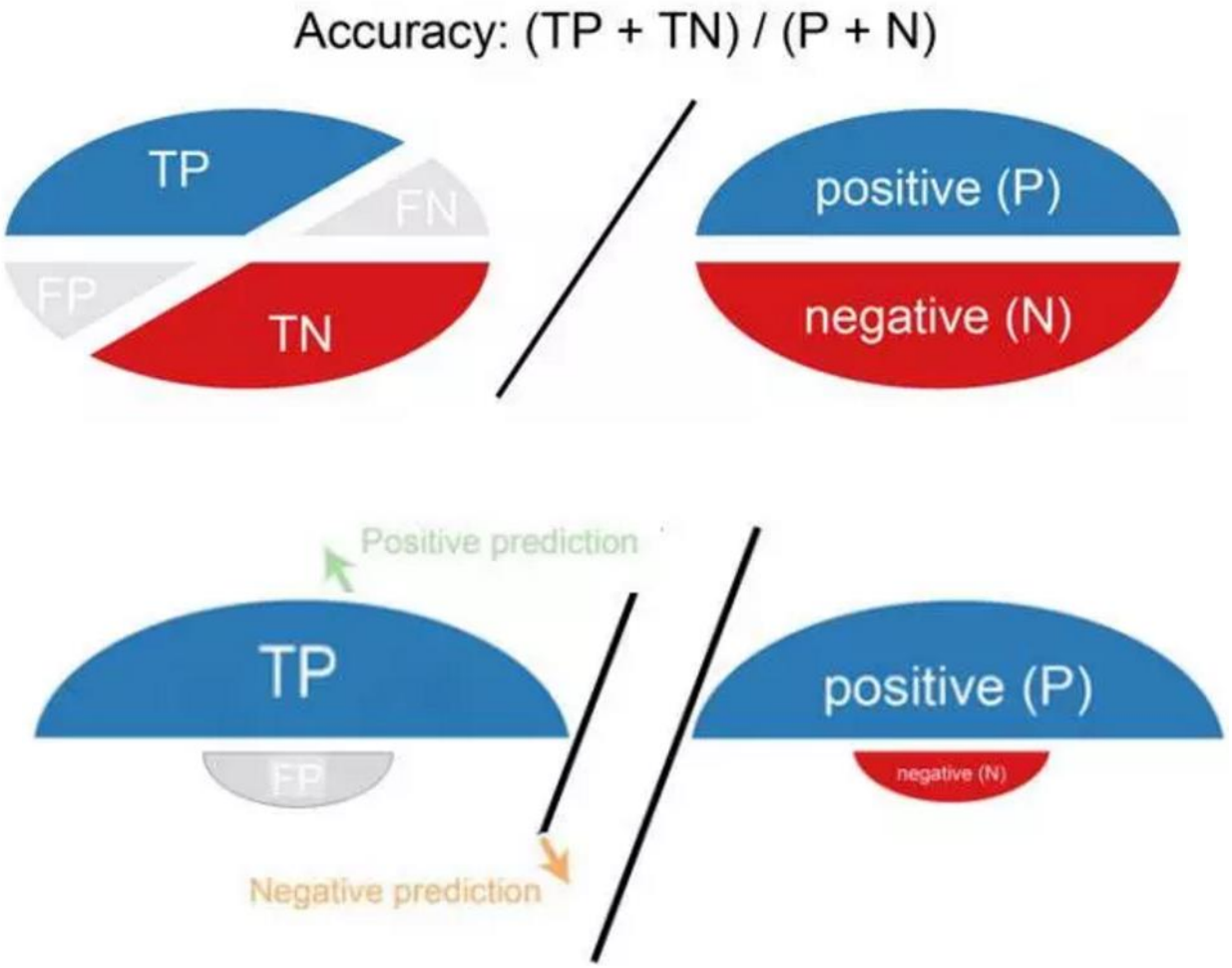
Feature Selection

Training & Evaluation

Visualization & Interpretation

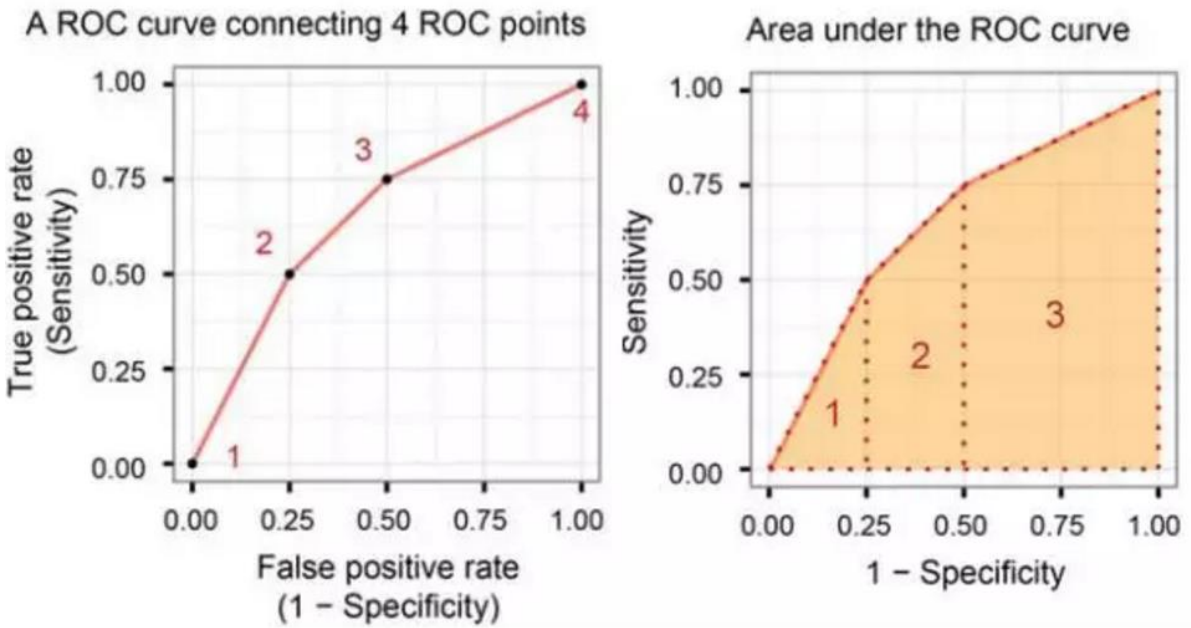


Imbalanced Data



Imbalanced Data

- ROC & AUC

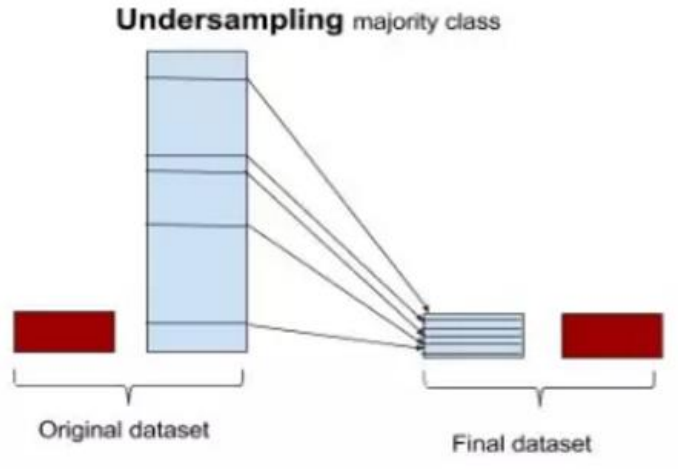


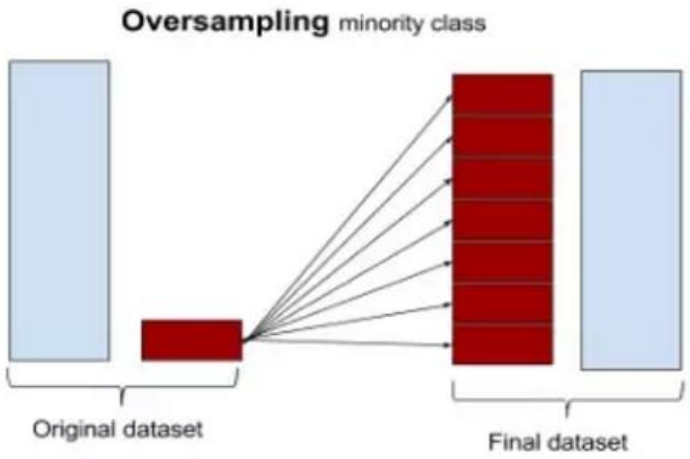

```
install.packages("AUC")
library(AUC)
```

```
install.packages("pROC")
library(pROC)
```

Imbalanced Data

- Sampling技术

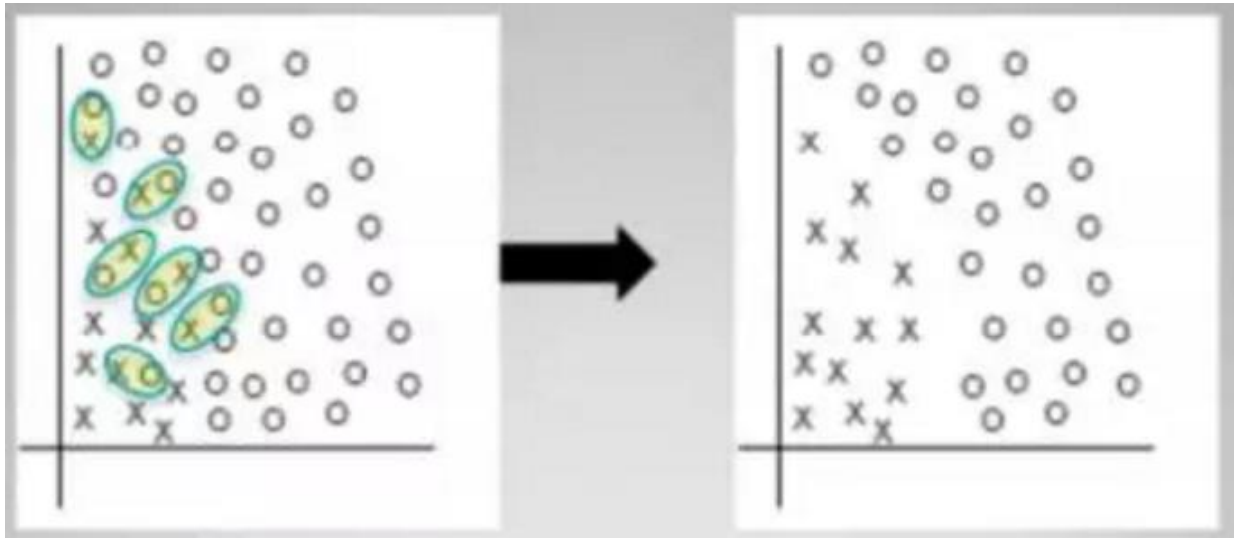





```
install.packages("ROSE")
library(ROSE)
```


Imbalanced Data

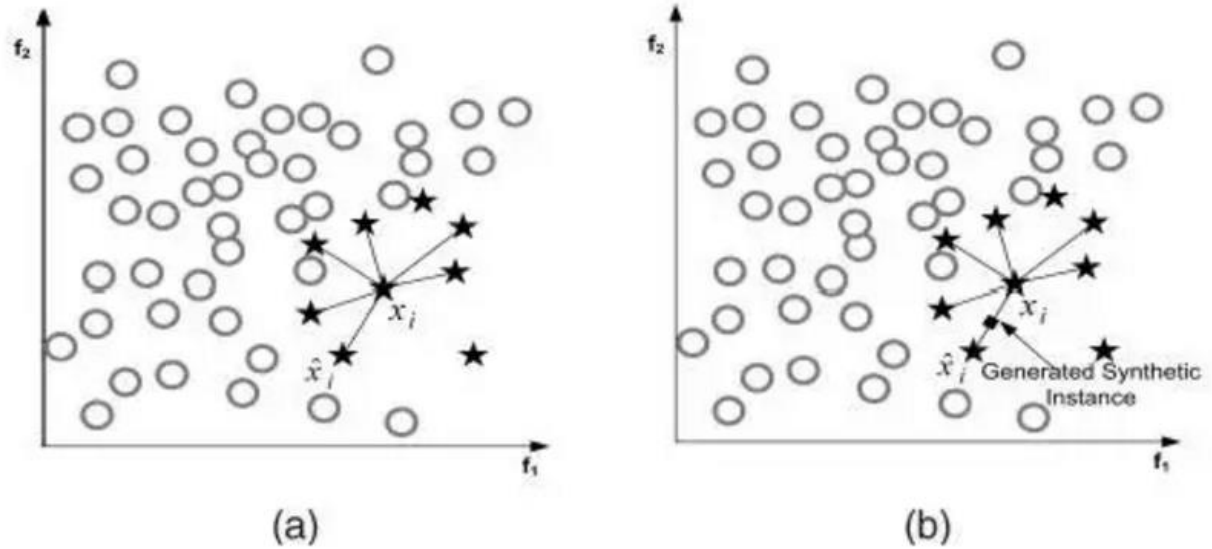
- 带边界清理的 Under-Sampling with Border Cleaning
  - 边界相邻匹配：Tomek Links



```
install.packages("unbalanced")
library(unbalanced)
```


Imbalanced Data

- 带合成的Oversampling技术With synthesize
  - SMOTE - Synthetic Minority Oversampling Technique

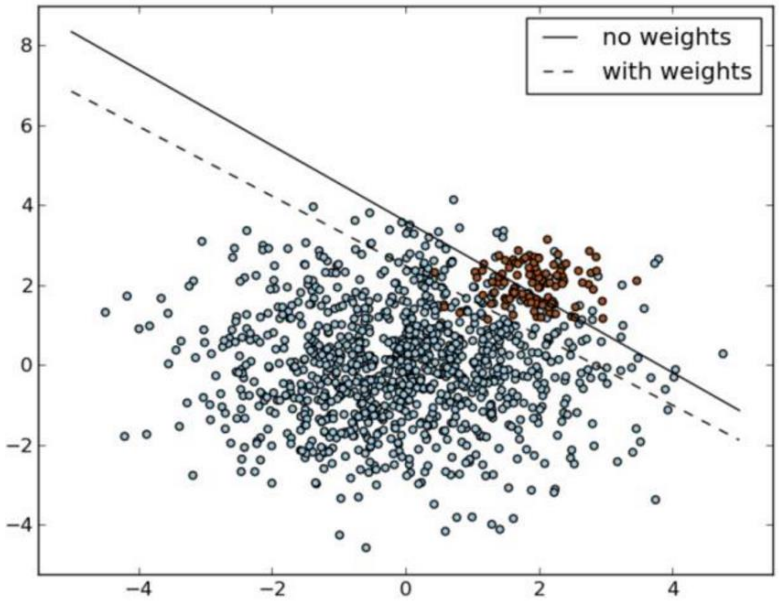



```
install.packages("DMwR")
library(DMwR)
```

```
install.packages("unbalanced")
library(unbalanced)
```

Imbalanced Data

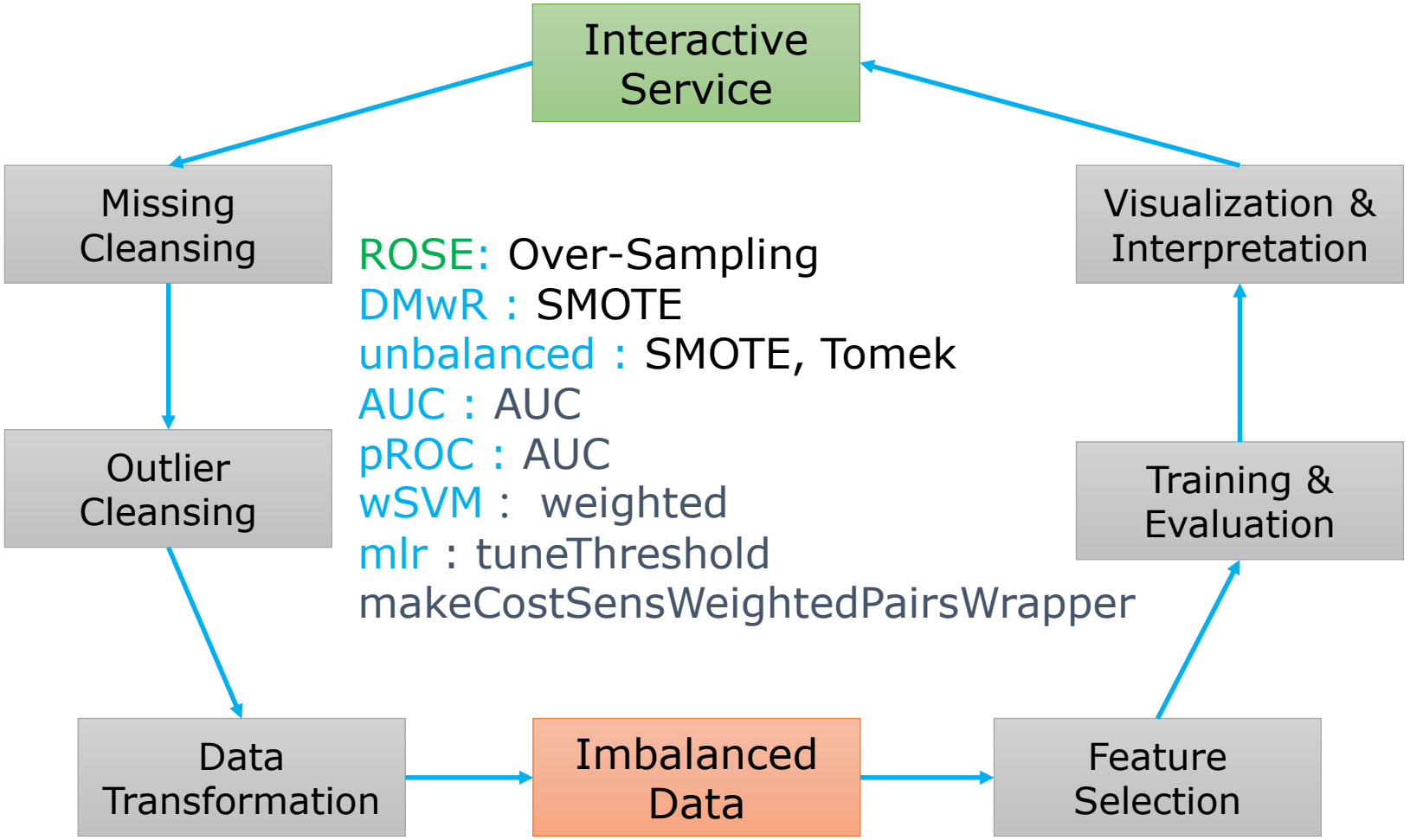
- Cost-Sensitive Learning
  - 权重Weight调整




```
install.packages("wSVM")
library(wSVM)
```

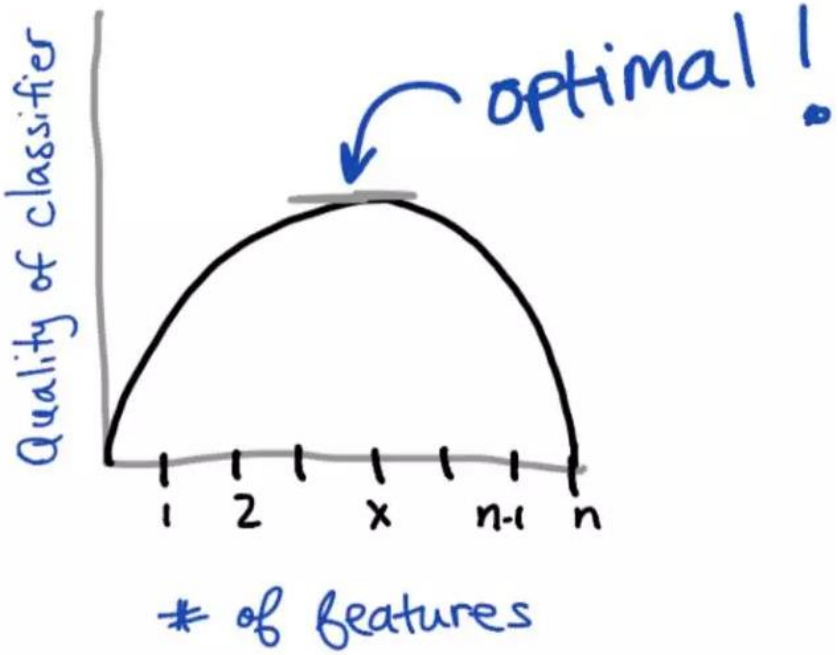
```
install.packages("mlr")
library(mlr)

tune.res = tuneThreshold(pred =
r$pred, measure = wf.costs)

lrn =
makeCostSensWeightedPairsWrap
per(lrn)
```

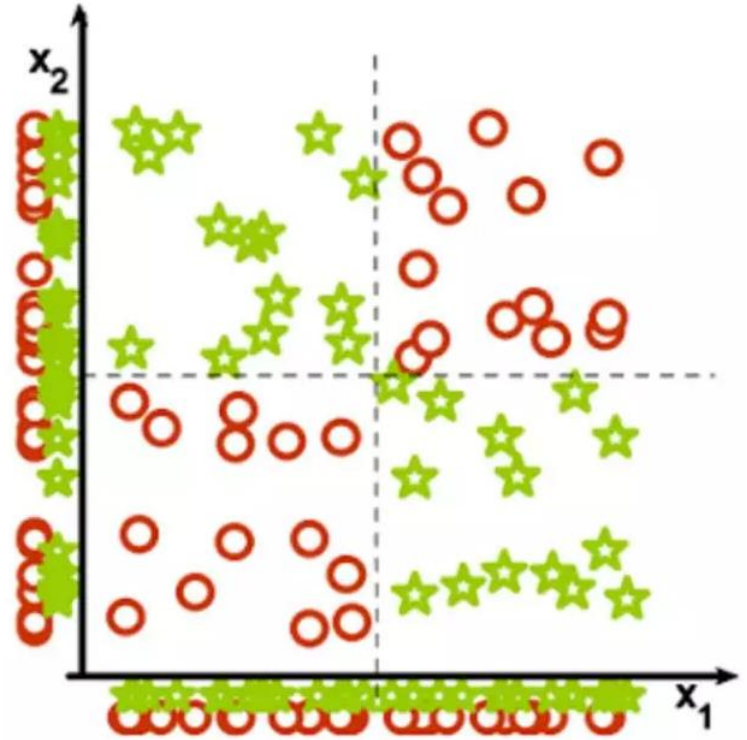

Imbalanced Data

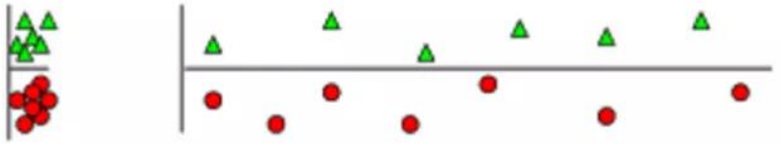


Feature Selection

- 特征选择与分类器性能的关系

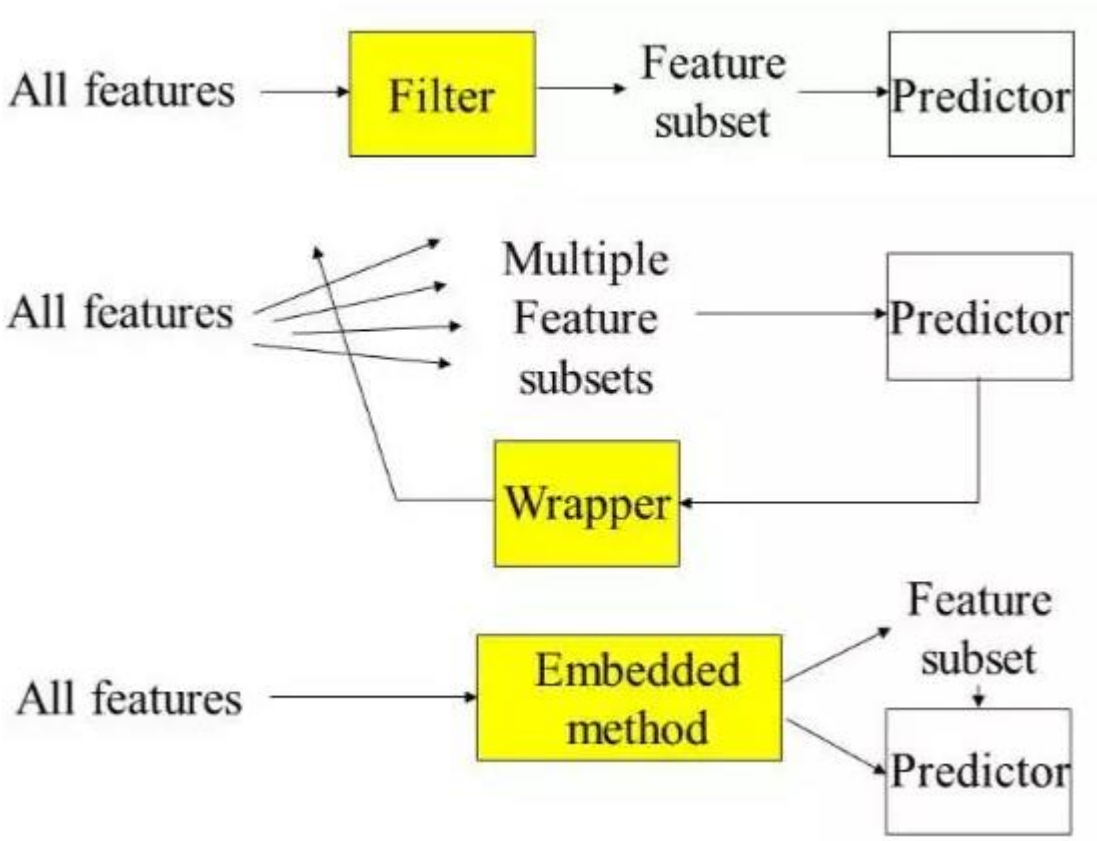






Feature Selection

- 特征选择三大主流



Feature Selection

- Filter 方式
  - 领域知识， 相关性， 距离，缺失， 稳定性
  - 单一特征选择

Welch's t-Test： 来判断两个属性的分布的均值方差距离

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

Chi-Squared test： 计算类别离散值之间的相关性

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

Information Gain：计算两个划分的一致性。

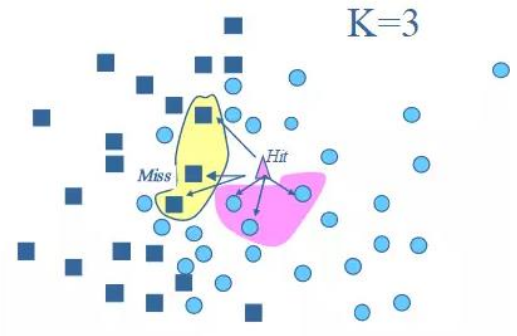
$$IG(f_i, C) = H(f_i) - H(f_i|C),$$

```
install.packages("FSelector")
library(FSelector)
```

```
install.packages("car")
library(car)
```

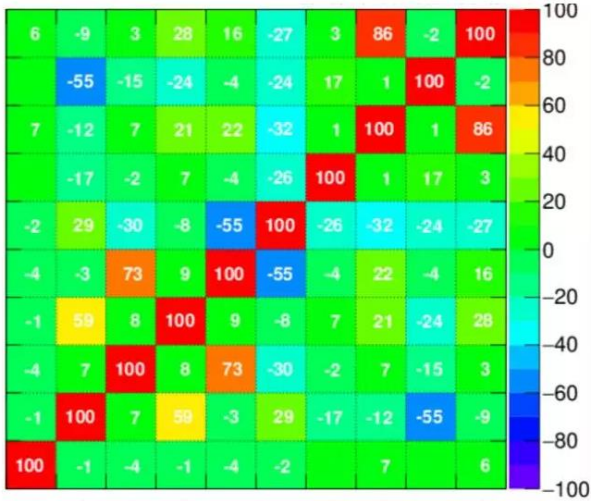

Feature Selection

- Filter 方式
  - 多特征选择



Relief-F： 根据随机选择的样本点，来计算属性之间的相关性

$$S_i = \frac{1}{2}\sum_{k=1}^{l} d(\mathbf{X}_{ik} - \mathbf{X}_{iM_k}) - d(\mathbf{X}_{ik} - \mathbf{X}_{iH_k})$$

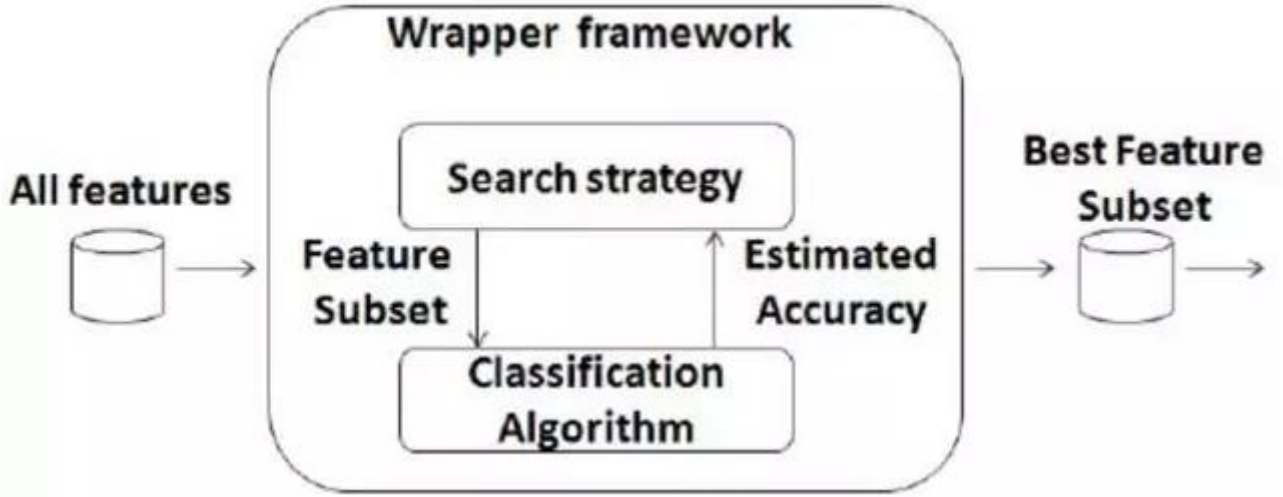


Correlation Feature Selection (CFS)： 利用属性之间的相关性， 进行选择。

```
install.packages("FSelector")
library(FSelector)
```

Feature Selection

- Wrapper 方式




```
install.packages("Boruta")
library(Boruta)
```

```
install.packages("caret")
library(caret)
```

```
install.packages("RRF")
library(RRF)
```

```
install.packages("sisal")
library(sisal)
```

```
install.packages("varSelRF")
library(varSelRF)
```

```
install.packages("rattle")
library(rattle)
```

Feature Selection

- Embedded 方式
  - Lasso

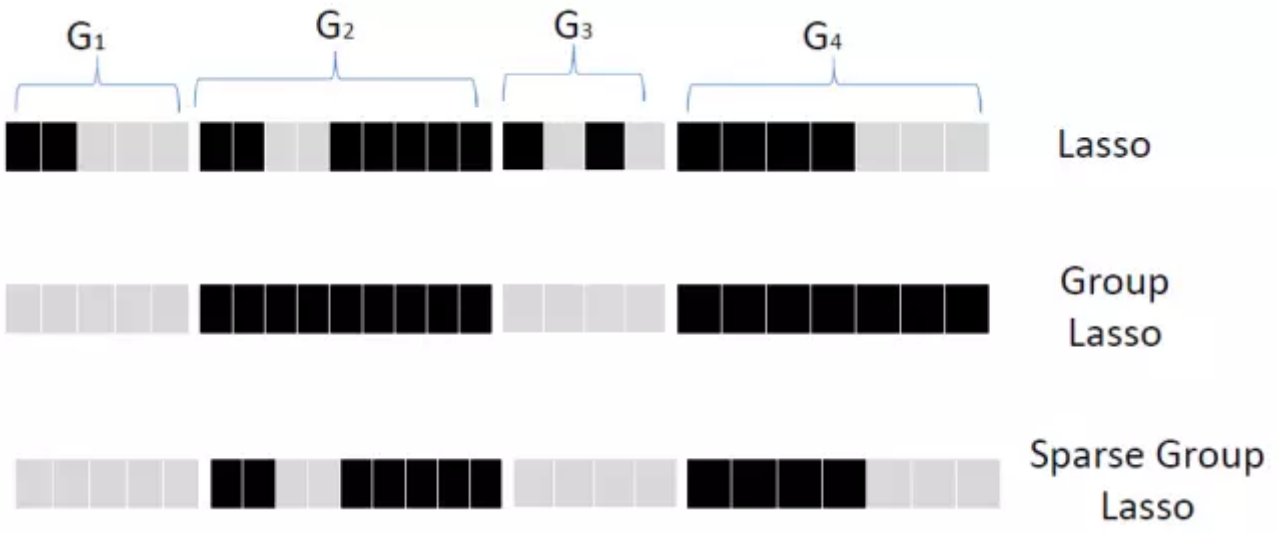
$$\min_{w,b} \mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell\left(y_i, f(\boldsymbol{x}_i)\right)}_{\frac{1}{2}\left(f(\boldsymbol{x}_i)-y_i\right)^2} + \lambda\Omega(\boldsymbol{w})$$

$\|\boldsymbol{w}\|_1$	$\frac{1}{2}\|\boldsymbol{w}\|_2^2$	$\rho\|\boldsymbol{w}\|_1 + (1-\rho)\frac{1}{2}\|\boldsymbol{w}\|_2^2$
LASSO	**Ridge Regression**	**Elastic Net**
Tibshirani, 1996	Hoerl & Kennard, 1970	Zou & Hastie, 2005

```
install.packages("glmnet")
library(glmnet)
```


Feature Selection

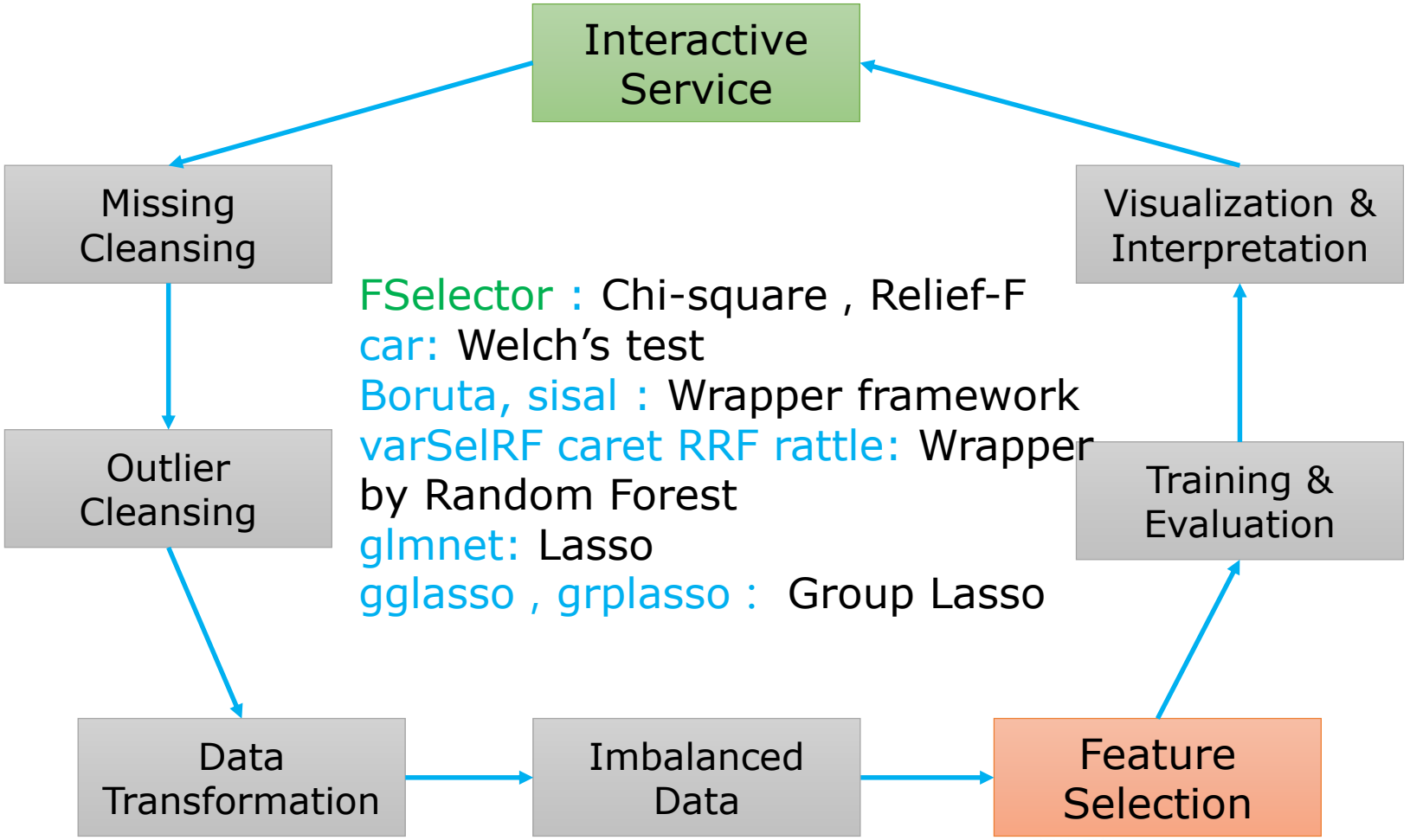
- Embedded 方式
  - Group Lasso




```
install.packages("gglasso")
library(gglasso)
```

```
install.packages("grplasso")
library(grplasso)
```

Imbalanced Data



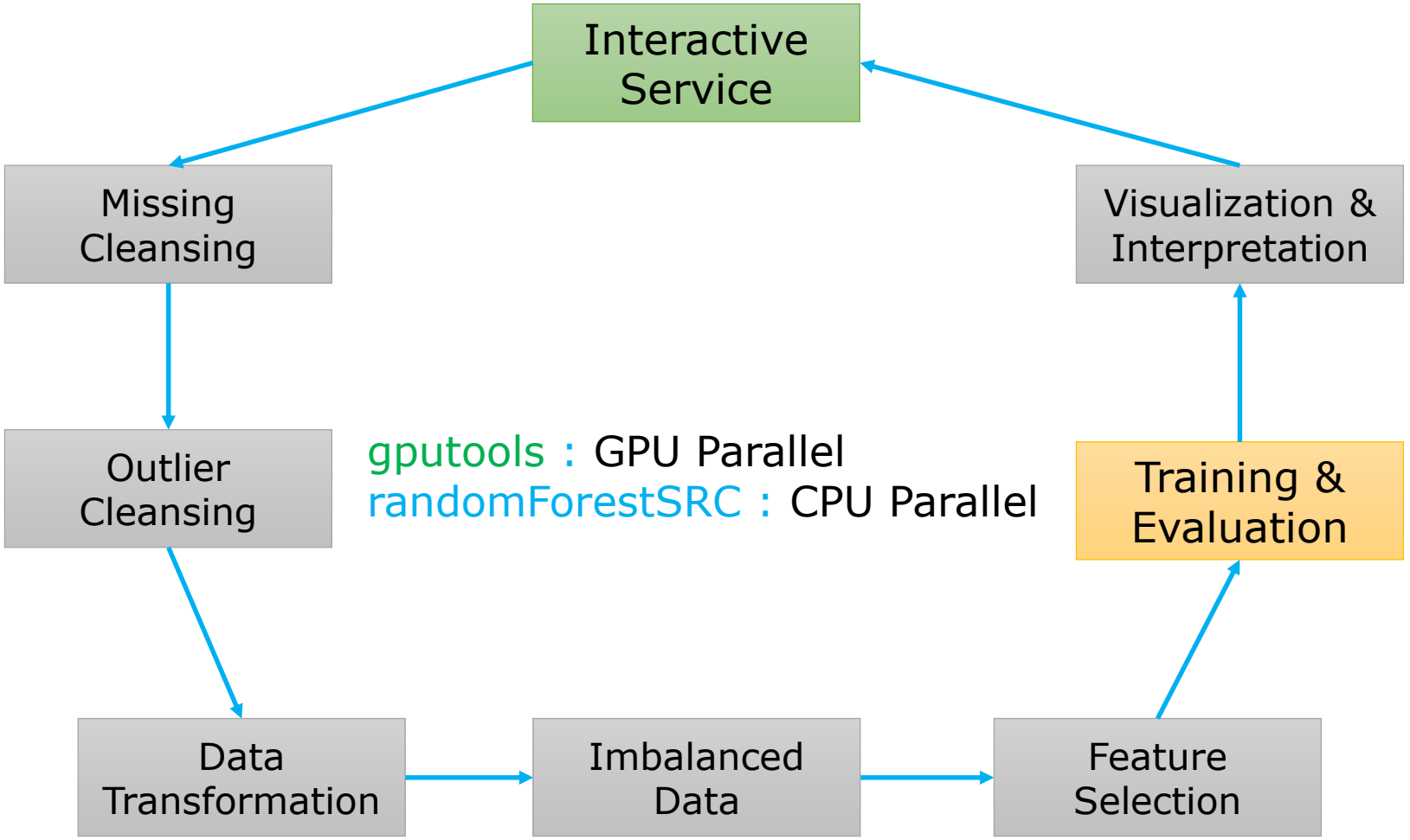
Training & Evaluation

- 利用不可解释模型进行效果上限确认
  - Random Forest, Neural Network
- 利用可解释的模型， 对结果进行分析， 反馈到数据获取阶段
  - Logistic Regression, Decision Tree
- 实时训练
  - 平行算法

```
install.packages("randomForestSRC")
library(randomForestSRC)
```

```
install.packages("gputools")
library(gputools)
```


Training & Evaluation



Visualization & Interpretation

- 可视化相当重要
  - PNG 图

- 交互式可视化更佳
  - JavaScript 图

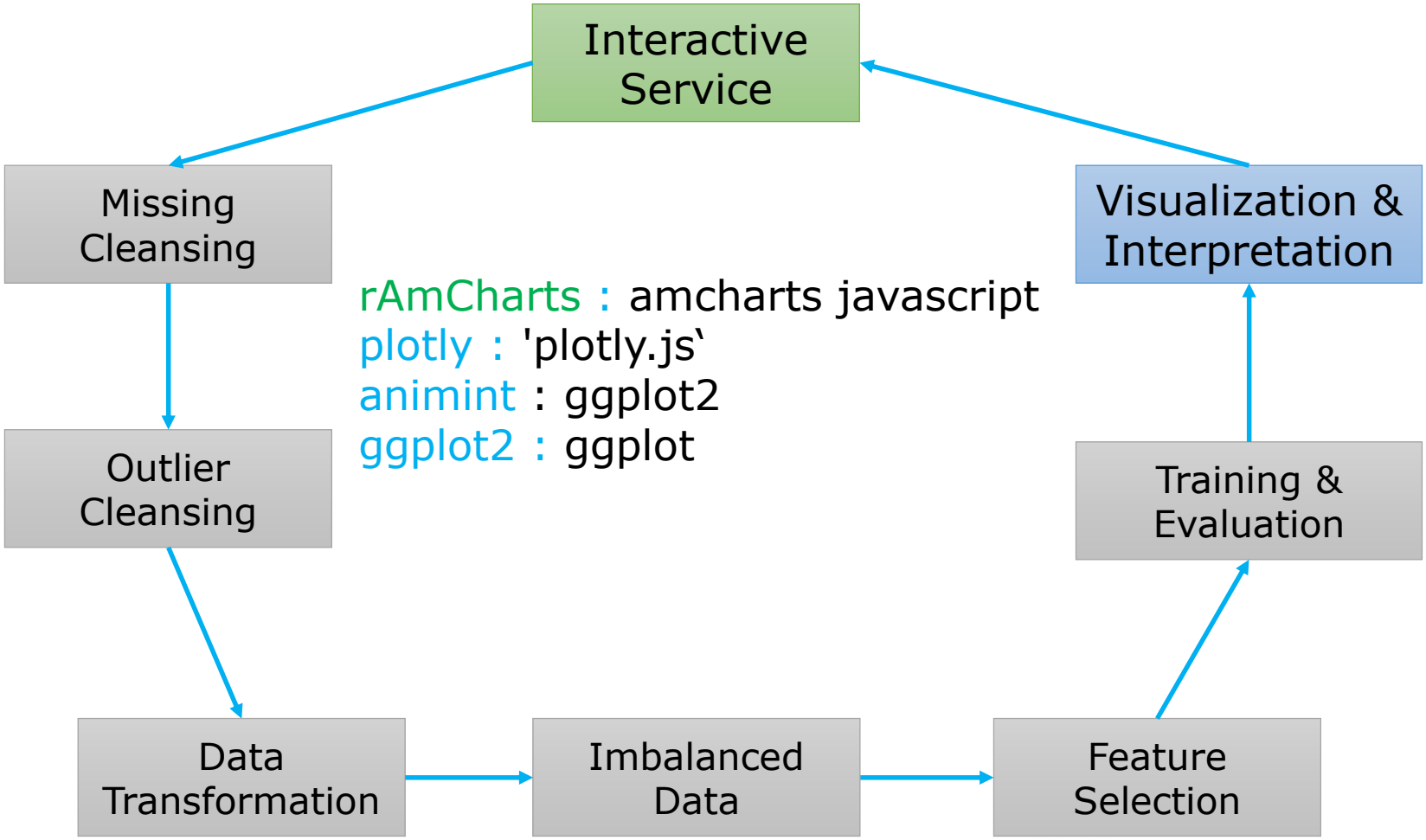
```
install.packages("ggplot2")
library(ggplot2)
```

```
install.packages("plotly")
library(plotly)
```

```
devtools::install_github("tdhock/animint", upgrade_dependencies=FALSE)
library(animint)
```

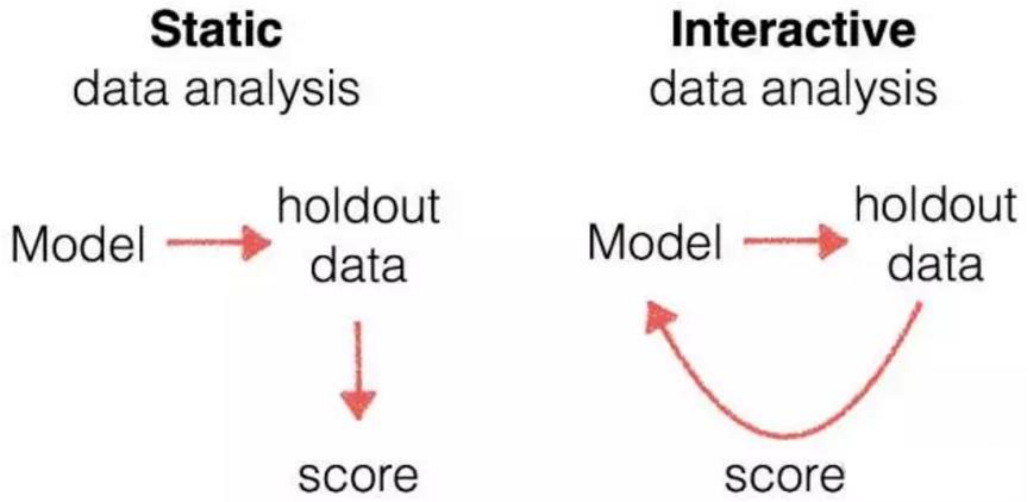
```
install.packages("rAmCharts")
library(rAmCharts)
```


Visualization & Interpretation

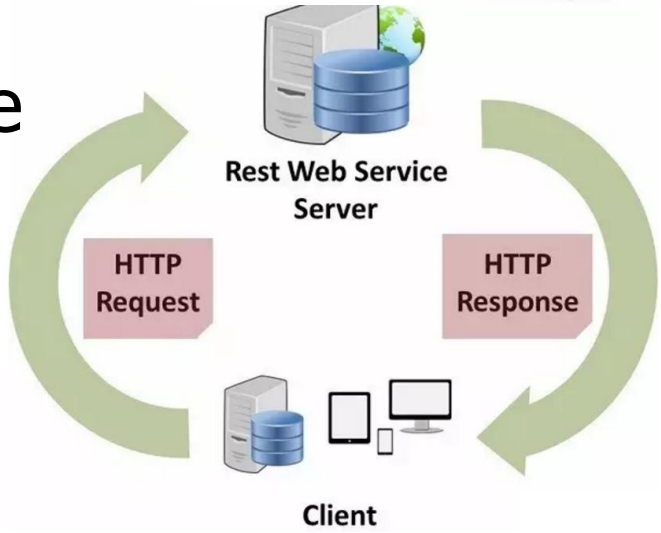


Interactive Service

- 交互式数据分析

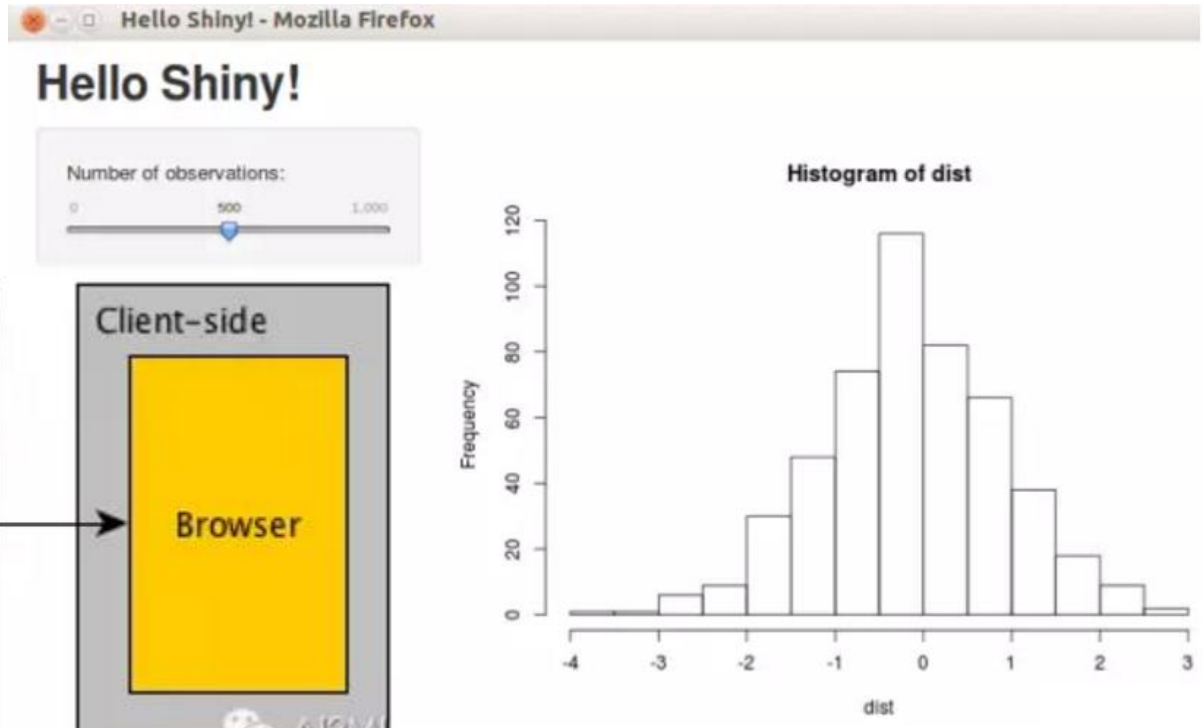


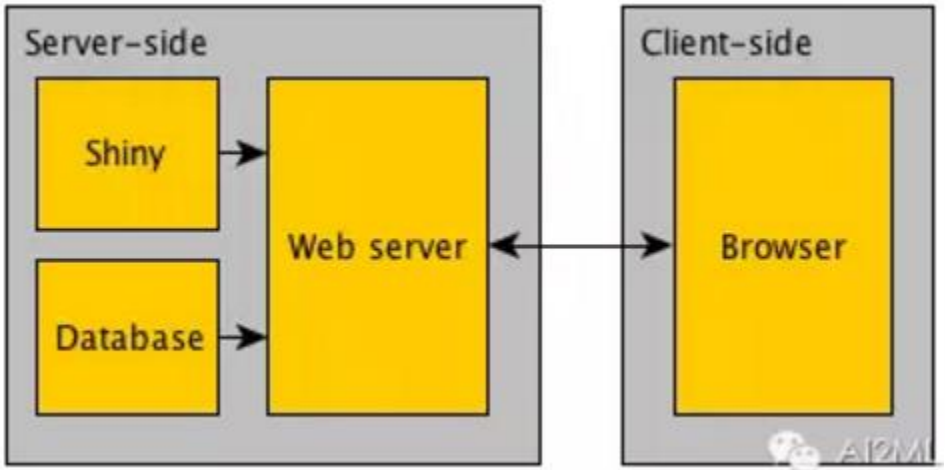
- Web Service



Interactive Service

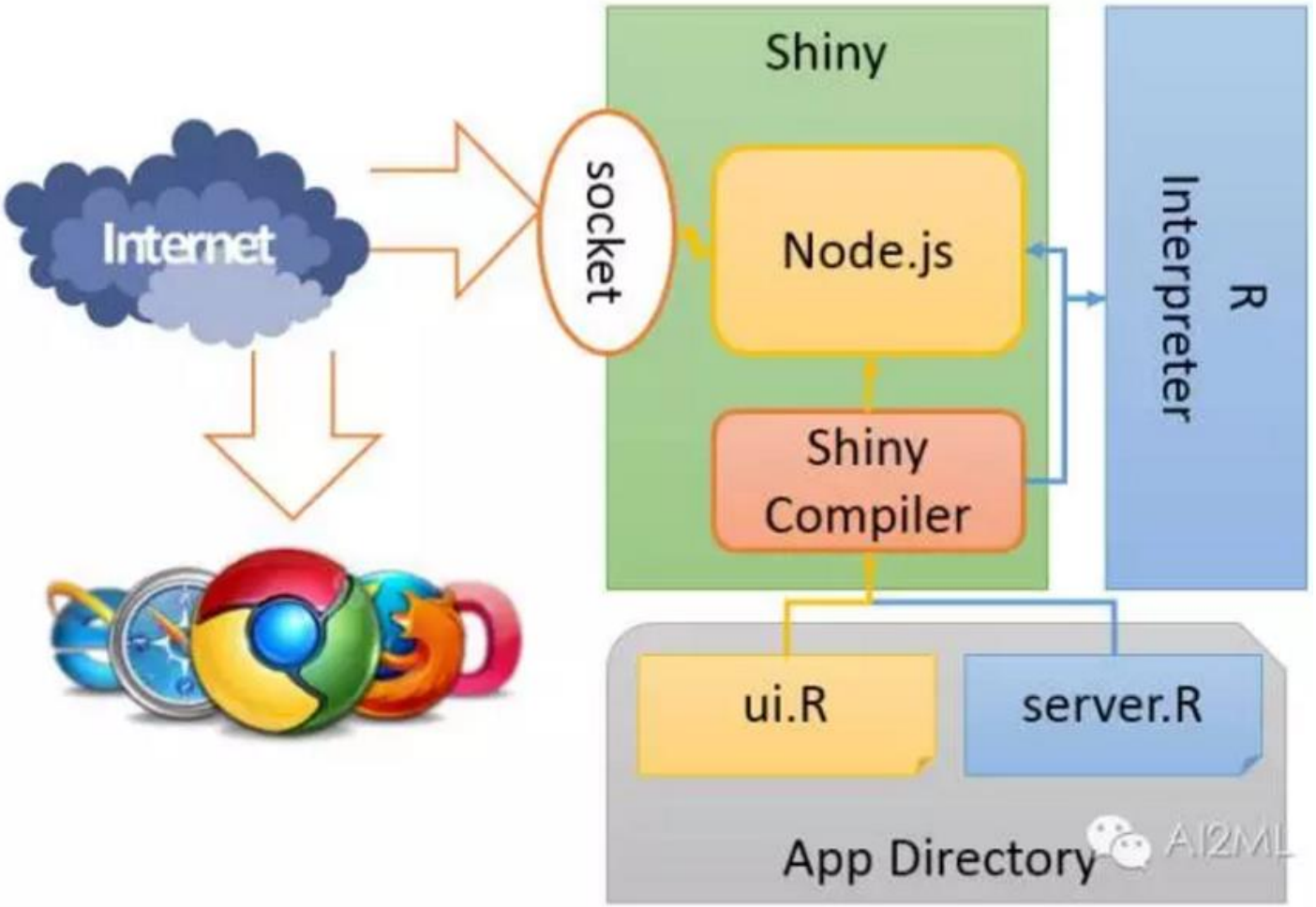
- Shiny: Web application framework






```
install.packages("shiny")
library(shiny)
```

Shiny 工作模式

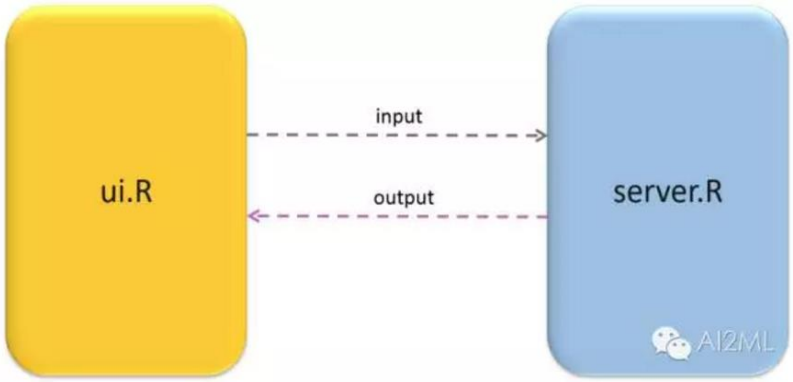

Shiny
Internet
socket
Node.js
R
Interpreter
Shiny
Compiler
ui.R
server.R
App Directory
AI2ML

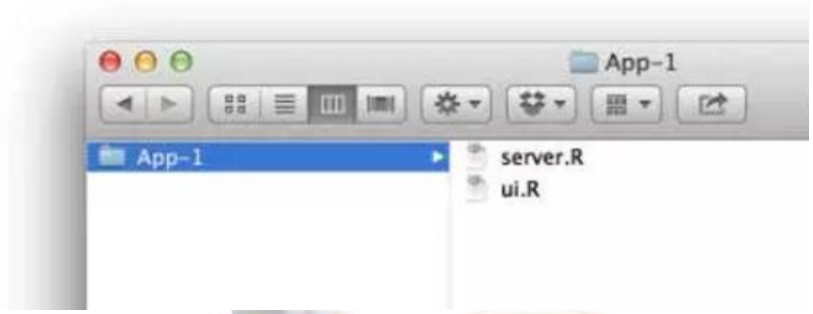
Shiny 代码结构

```r
# ui.R

library(shiny)

shinyUI(fluidPage(

))
```

```r
# server.R

library(shiny)

shinyServer(function(input, output) {

})
```

```
> setwd("Project_HOME")
> library(shiny)
> runApp("App-1")
```





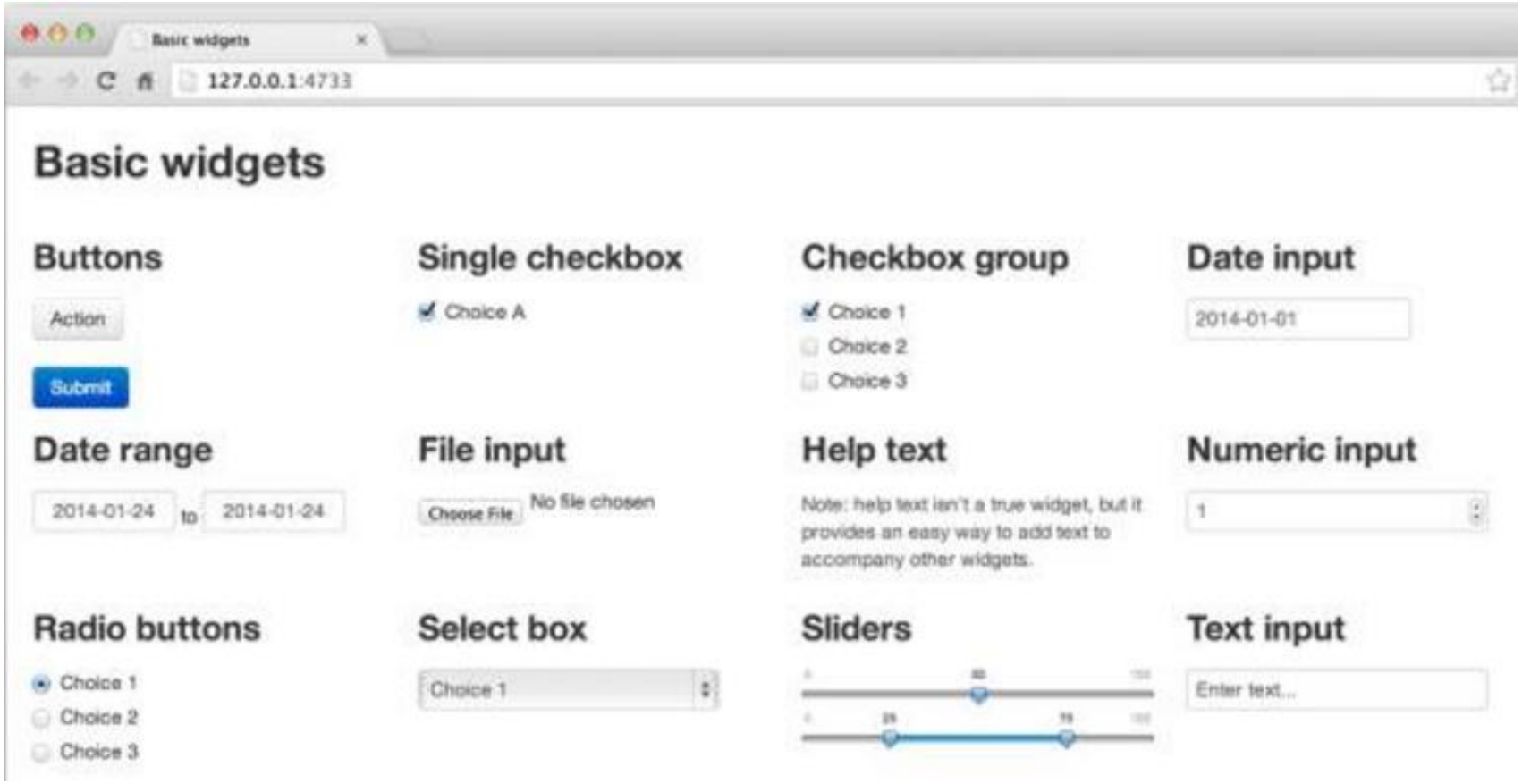
Shiny 输入输出

```
shinyApp(
  ui = fluidPage(sidebarLayout(
    sidebarPanel(
      sliderInput("bins", "Number of bins:", min = 1, max = 75, value = 30)
    ),
    mainPanel(
      plotOutput("distPlot")
    )
  )),
  server = function(input, output, session) {
    output$distPlot <- renderPlot({
      hist(faithful$eruptions, breaks = input$bins)
    })
  })
```


布局(Layout)

```
shinyUI(fluidPage(
    titlePanel("title panel"),

    sidebarLayout(
      sidebarPanel( "sidebar panel",
      selectInput('element_id', label = 'Select one option', choices = LETTERS[1:10]),

      textInput('title_text_box_id', label = 'Enter a title for the plot')),

    mainPanel("main panel",
      h1('The title of some text'),
      p('And here is some content that is put into the first paragraph'),        p(textOut
put('dynamicText')),

      plotOutput('dynamicPlot'))
    )
))
```


构件(Widget)


Basic widgets
127.0.0.1:4733
Basic widgets
Buttons
Action
Submit
Single checkbox
Choice A
Checkbox group
Choice 1
Choice 2
Choice 3
Date input
2014-01-01
Date range
2014-01-24
to
2014-01-24
File input
Choose File
No file chosen
Help text
Note: help text isn't a true widget, but it provides an easy way to add text to accompany other widgets.
Numeric input
1
Radio buttons
Choice 1
Choice 2
Choice 3
Select box
Choice 1
Sliders
Text input
Enter text...

主题（Theme）

```
shinyUI(fluidPage(theme = "bootstrap.css",

  titlePanel("My Application")

  # application UI                    ))
```

```
install.packages("shinythemes")
```

```
## ui.R ##
library(shinythemes)
fluidPage(theme = shinytheme("simplex"),
  ...
)
```


优化(Optimization)

- Reactive Expression

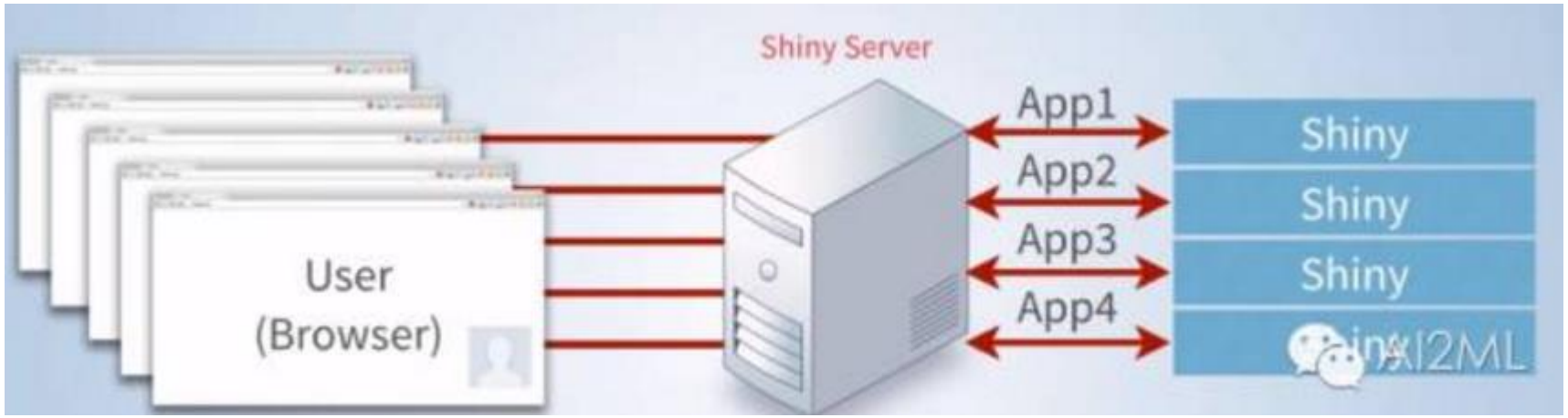
```
shinyServer(function(input, output) {
  output$dynamicPlot <- renderPlot({
    dat = get_data(input$ui_element1, input$ui_element2)
    plot(dat, input$ui_element3)
  })})
```

```
shinyServer(function(input, output) {
  input_data = reactive({
    dat = get_data(input$ui_element1, input$ui_element2)
  })
  output$dynamicPlot <- renderPlot({
    plot(input_data(), input$ui_element3)
  })})
```

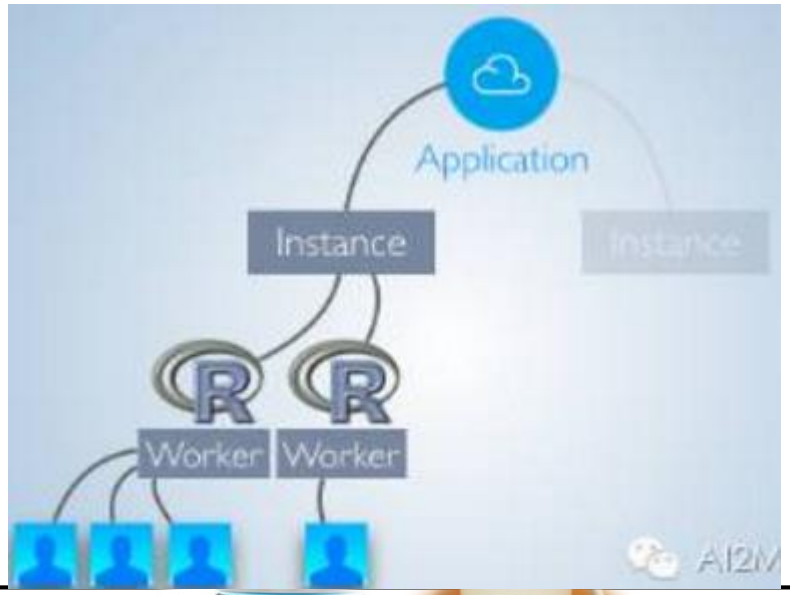

扩展（Extending）

- shinydashboard
- shinyURL
- htmlwidgets：
  - leaflet – Geo-spatial mapping
  - dygraphs – Time series charting
  - MetricsGraphics – Scatterplots and line charts with D3
  - networkD3 – Graph data visualization with D3
  - DataTables – Tabular data display
  - threejs – 3D scatterplots and globes
  - rCharts – Multiple JavaScript charting libraries

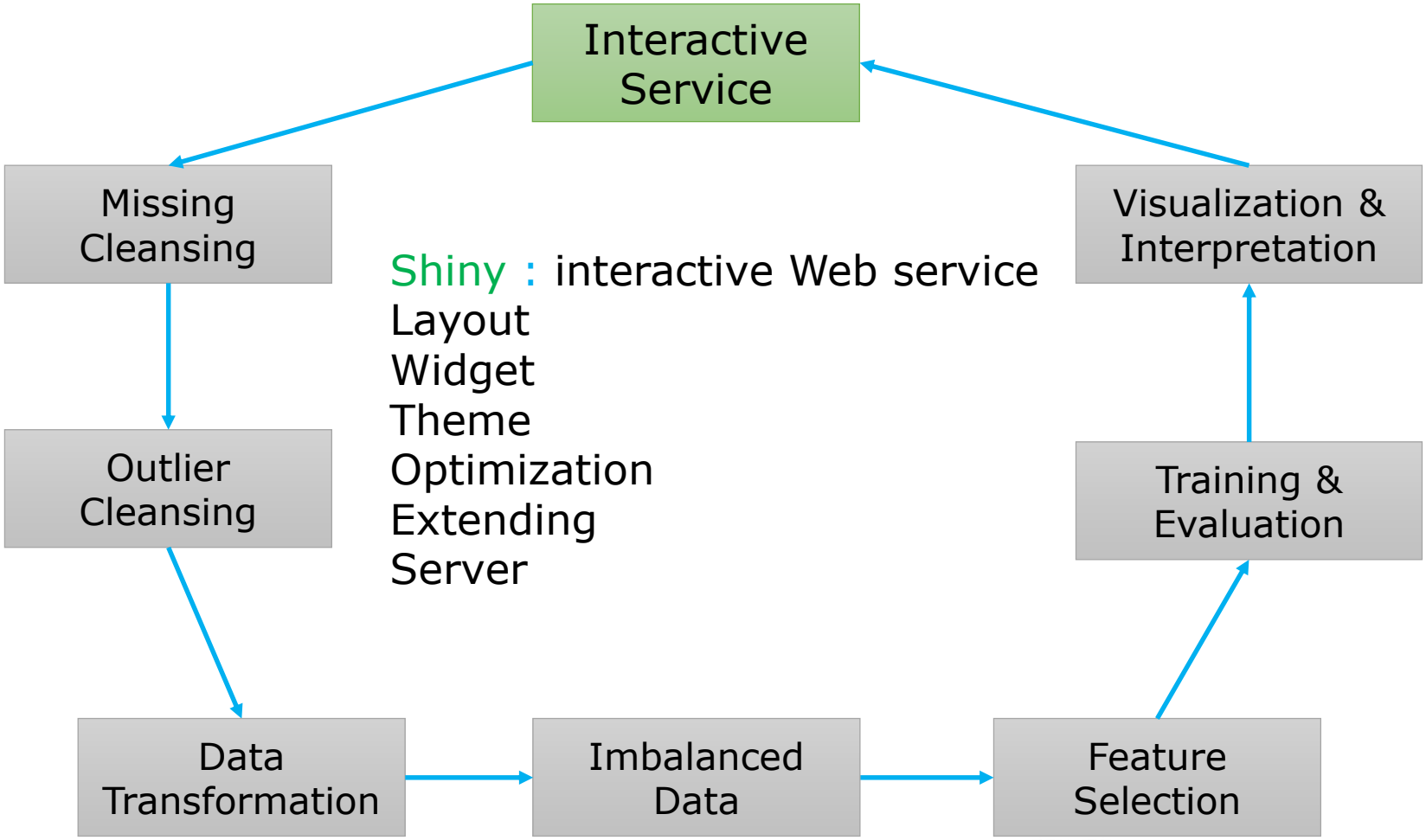
Shiny 服务器



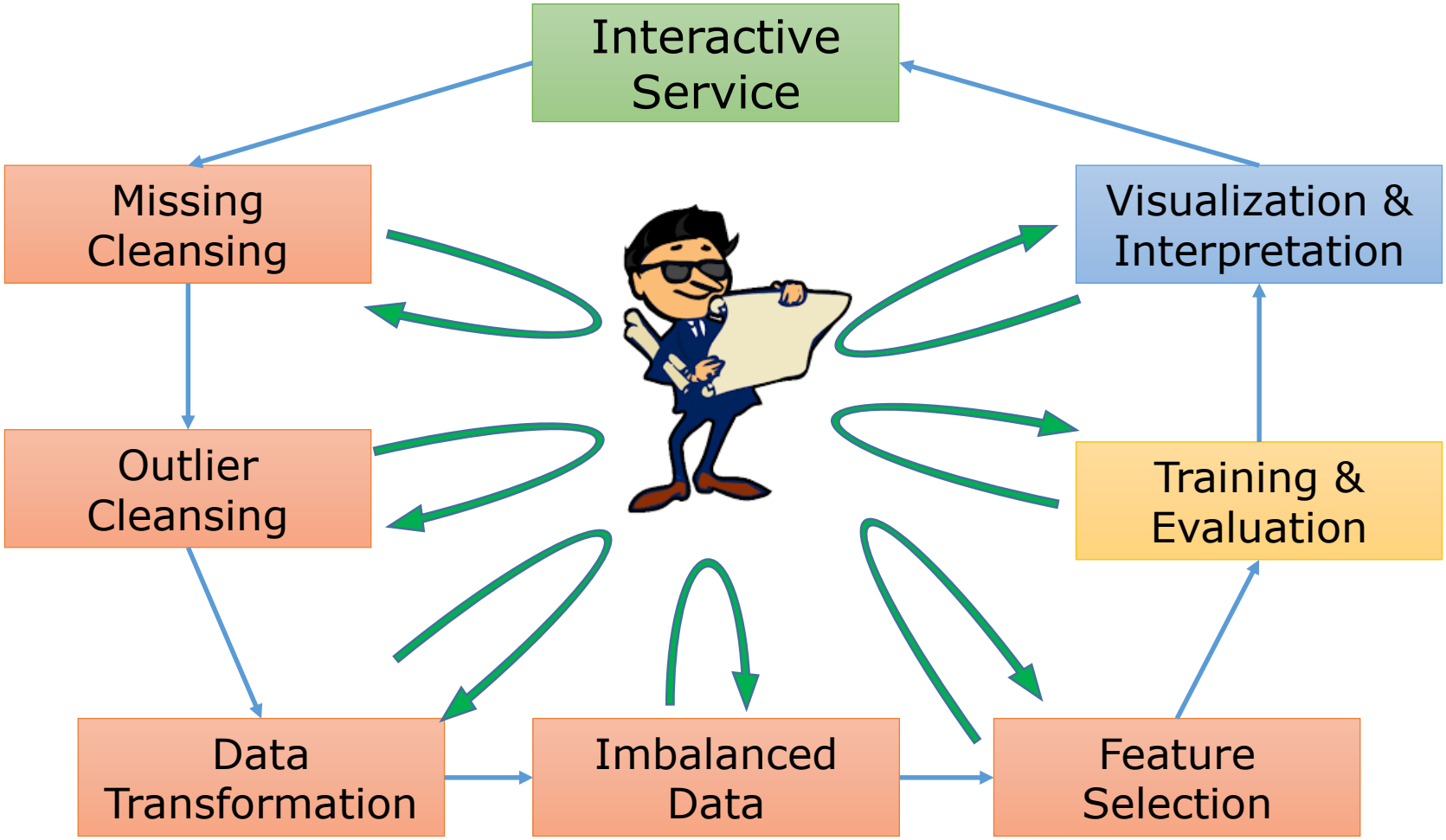
- Shiny服务器
- Shiny云服务
  - “www.shinyapps.io”



Interactive Service



交互式快速迭代


Interactive Service
Missing Cleansing
Visualization & Interpretation
Outlier Cleansing
Training & Evaluation
Data Transformation
Imbalanced Data
Feature Selection

Thank You ~

References

- https://www.analyticsvidhya.com/blog/2016/03/tutorial-powerful-packages-imputing-missing-values/
- http://blog.revolutionanalytics.com/2016/07/user-2016-tutorials-part-2-.html
- https://github.com/twitter/AnomalyDetection
- http://www.rdatamining.com/examples/outlier-detection、
- https://statswithcats.wordpress.com/2010/11/21/fifty-ways-to-fix-your-data/
- https://www.rdocumentation.org/packages/mlr/versions/2.9
- https://cran.r-project.org/web/views/HighPerformanceComputing.html