2018.12.1.SAT

GUIで簡単！モダンなデータ解析

株式会社ef-prime

鈴木 了太　@efprime_jp

ef-prime, inc.

自己紹介

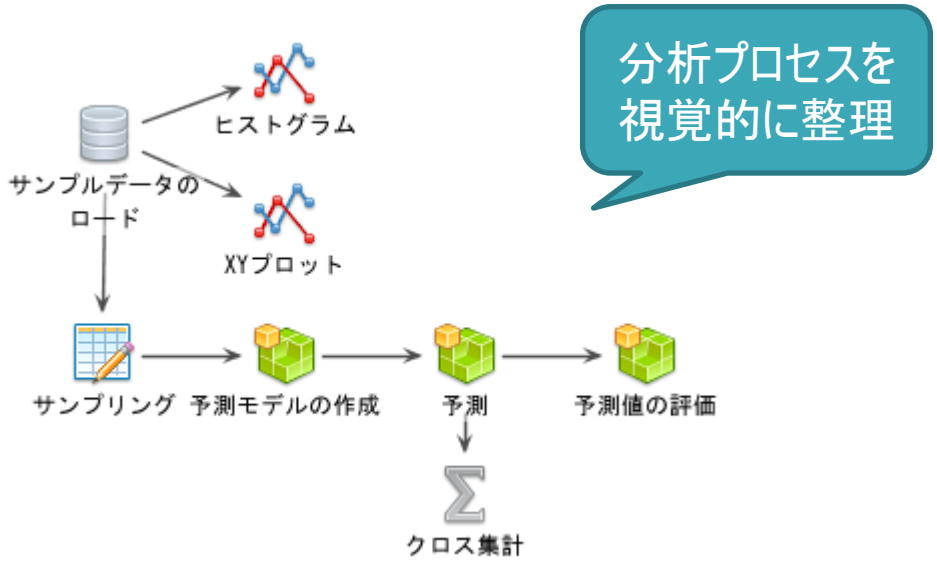
- 鈴木 了太
  - 株式会社ef-prime代表
  - Rユーザー歴17年くらい
    - 開発CRANパッケージ：pvclust
    - 書籍：Useful R 10（共立出版、第IV部のみ）
    - 統数研のRユーザー会やUseR!にも出没しています

- 株式会社ef-prime（エフプライム）
  - データ分析コンサルティング（主に企業向け）
  - 2006年設立、多くのプロジェクトでRを活用
  - データ分析ツールを開発・公開

ef-prime, inc.

R AnalyticFlow

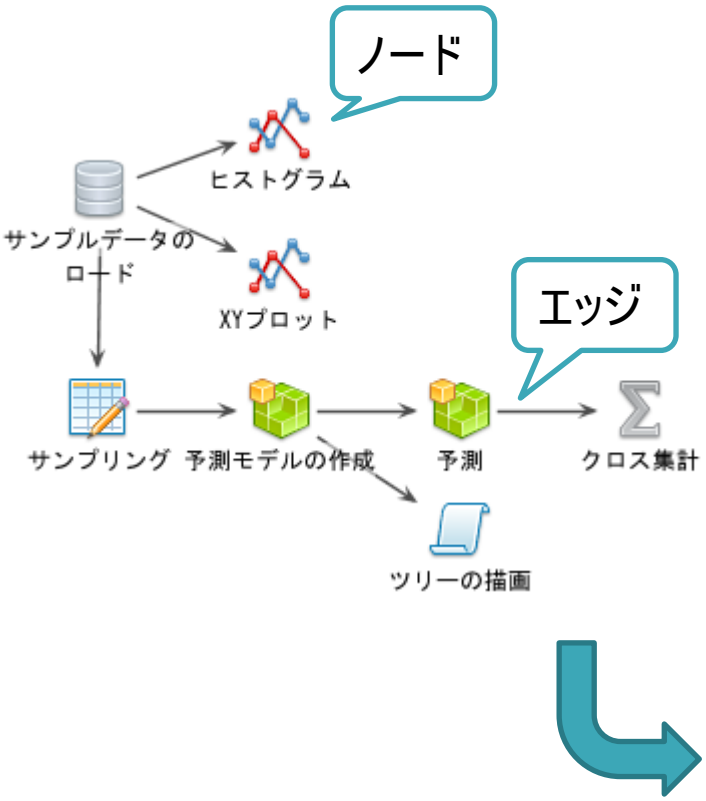
- データ解析のためのR GUI
  - 分析プロセスをワークフローで表現
  - オープンソース、無償公開
  - Javaで開発、マルチOS対応
    - Windows / Mac / Linux



https://r.analyticflow.com

R AnalyticFlow

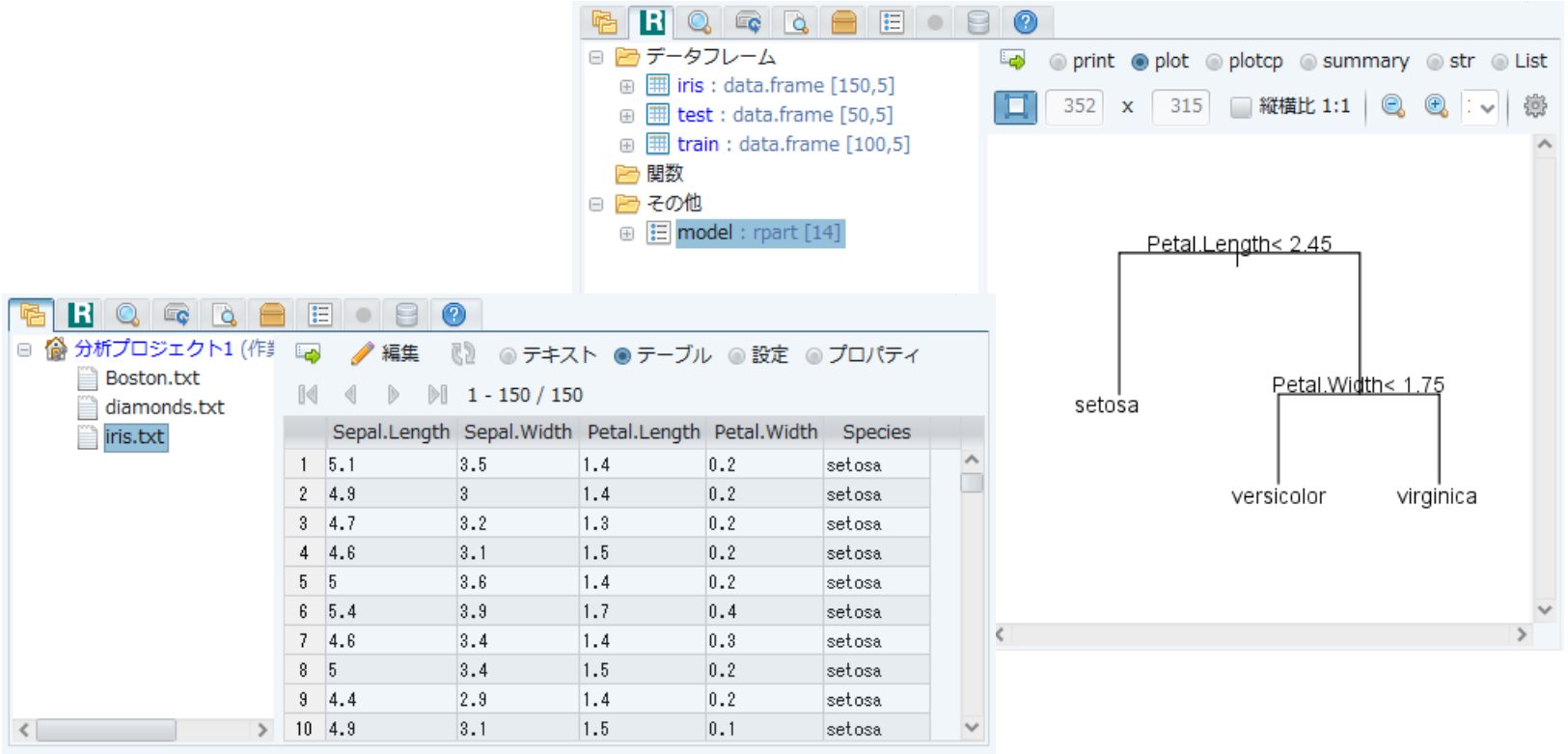
ノードをエッジで繋いだ分析フローを作成、Rスクリプトを生成して実行




```r
# 1. データの読み込み
data(iris)

# 2. 探索的分析
plot(iris[, 1:4], col =
as.integer(iris$Species) + 1)
boxplot(Petal.Length ~ Species, data =
iris, col = 3, main = "Petal.Length")

# 3. モデリング
library(rpart)
rp <- rpart(Species ~ ., iris)

# 4. モデルの確認
plot(rp, margin = 0.1, branch = 0.3)
text(rp, fancy = T, all = T, use.n = T)

# 5. 予測および評価
pred <- predict(rp, type = "class")
xtabs(~pred + iris$Species)
```

R AnalyticFlow

ファイルやRオブジェクトの一覧をインタラクティブに利用することができる。
データファイルを選択して読み込み処理を作成したり、
実行結果として得られたオブジェクトを確認することができる

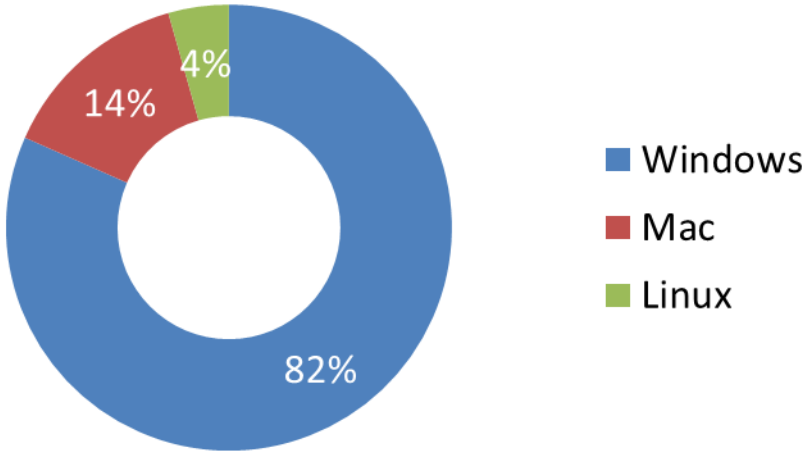


	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

多言語・マルチOS

表示言語として日本語と英語を選択可能
世界146ヶ国でダウンロード（2017年11月時点、直近1年間）
Windows / Mac / Linux をサポート

1. Japan
2. United States
3. India
4. Taiwan
5. China
6. South Korea
7. Peru
8. United Kingdom
9. Germany
10. Italy

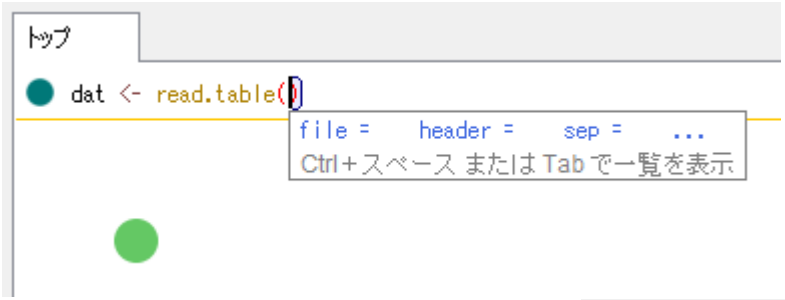


これまでの歩み

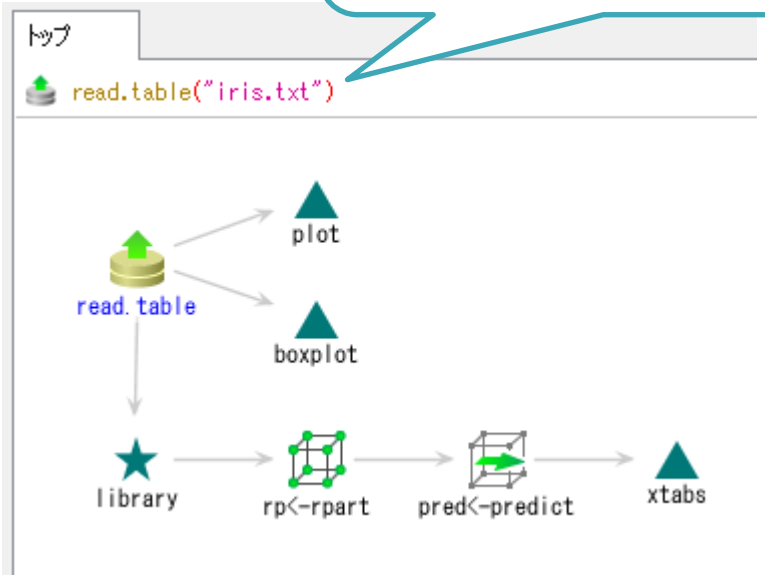
時期	バージョン	詳細
2007年12月	0.1.0	プレビュー版公開（Windowsのみ）
2009年12月	1.0.0	現在に至る基本機能の多くを実装 Windows / Mac / Linuxで動作
2012年12月	2.0.0	タブインターフェースへの統合、デバッグ補助、 自動バックアップ、64bit対応
2015年12月	3.0.0	GUIによる分析機能、新インターフェース、 プロジェクト管理、カスタムUI作成機能
2017年6月	3.1.0	分析機能の強化（予測分析、多変量解析、 仮説検定など）、利便性の向上、 安定性の向上（Rの別プロセス化）
2018年8月	3.1.8	最新版 （問題の修正、RやJavaの更新に対応など）

初期のバージョン（0.1～2.x）

ユーザーがRスクリプトを直接記述することで分析フローを作成。
作成したフローは誰でも簡単に実行できる

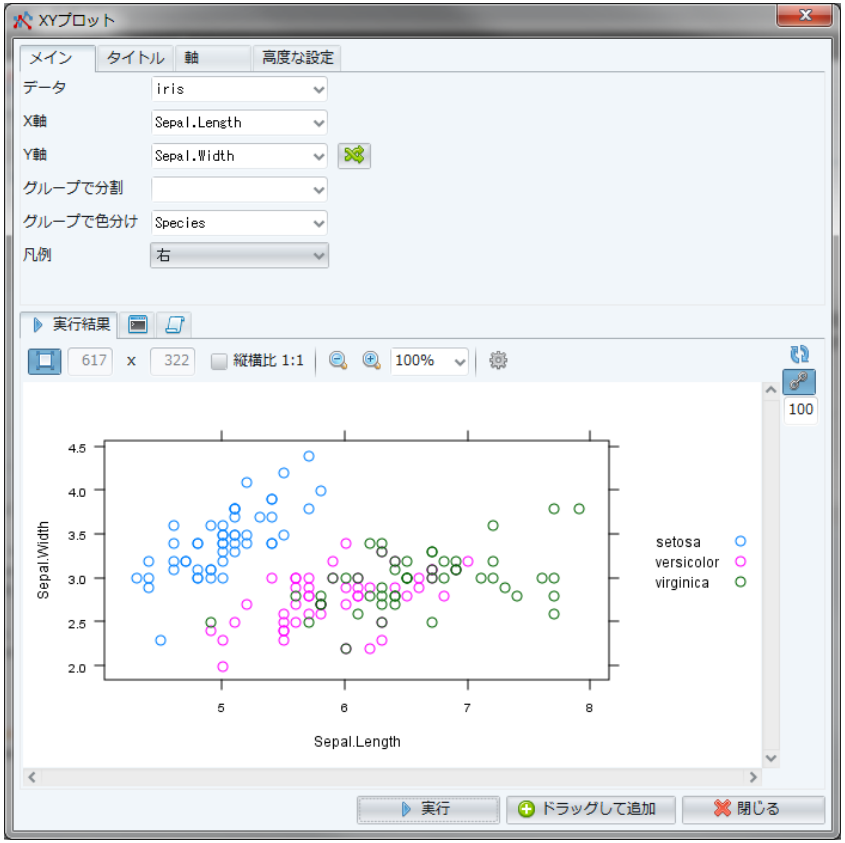


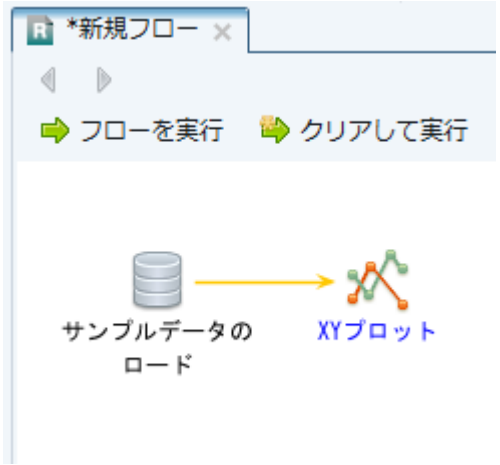
ノードにRスクリプトを直接記述。
関数を検出してタイトルにし、
対応するアイコンを表示する



最近のバージョン（3.x）

簡単なマウス操作でRスクリプトを記述せずに分析が可能。
Rの知識が必須ではなくなり、スクリプトの記述量も削減

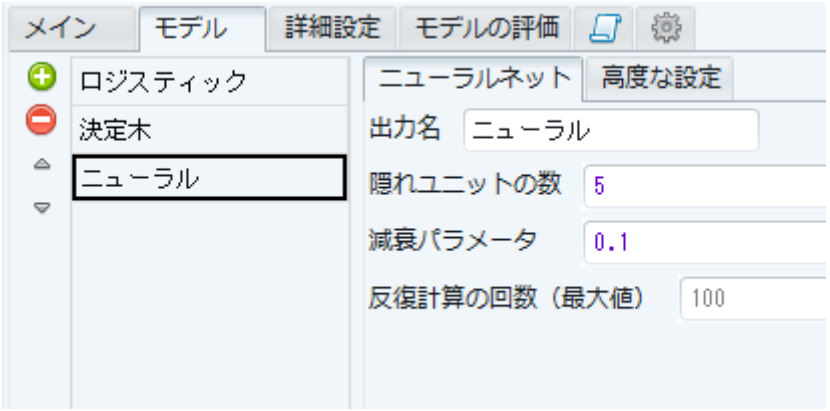


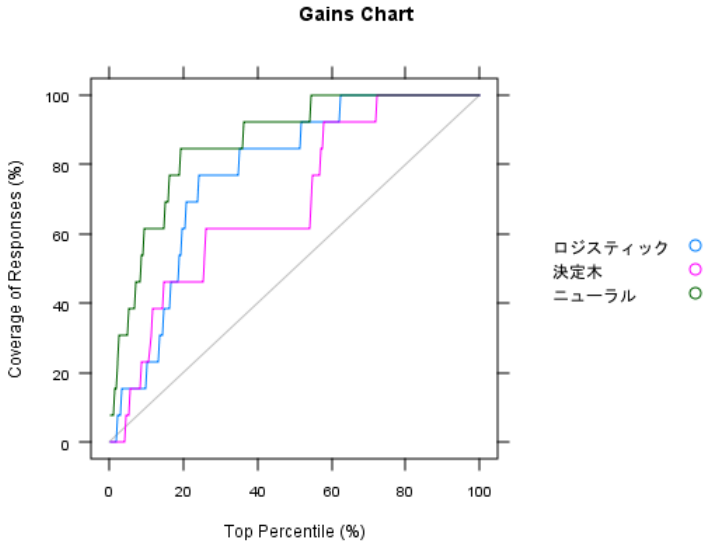


ドラッグ＆ドロップ

機能例：予測分析

複数の予測手法やパラメータ設定のもとで予測モデルを作成し、予測精度を評価して出力





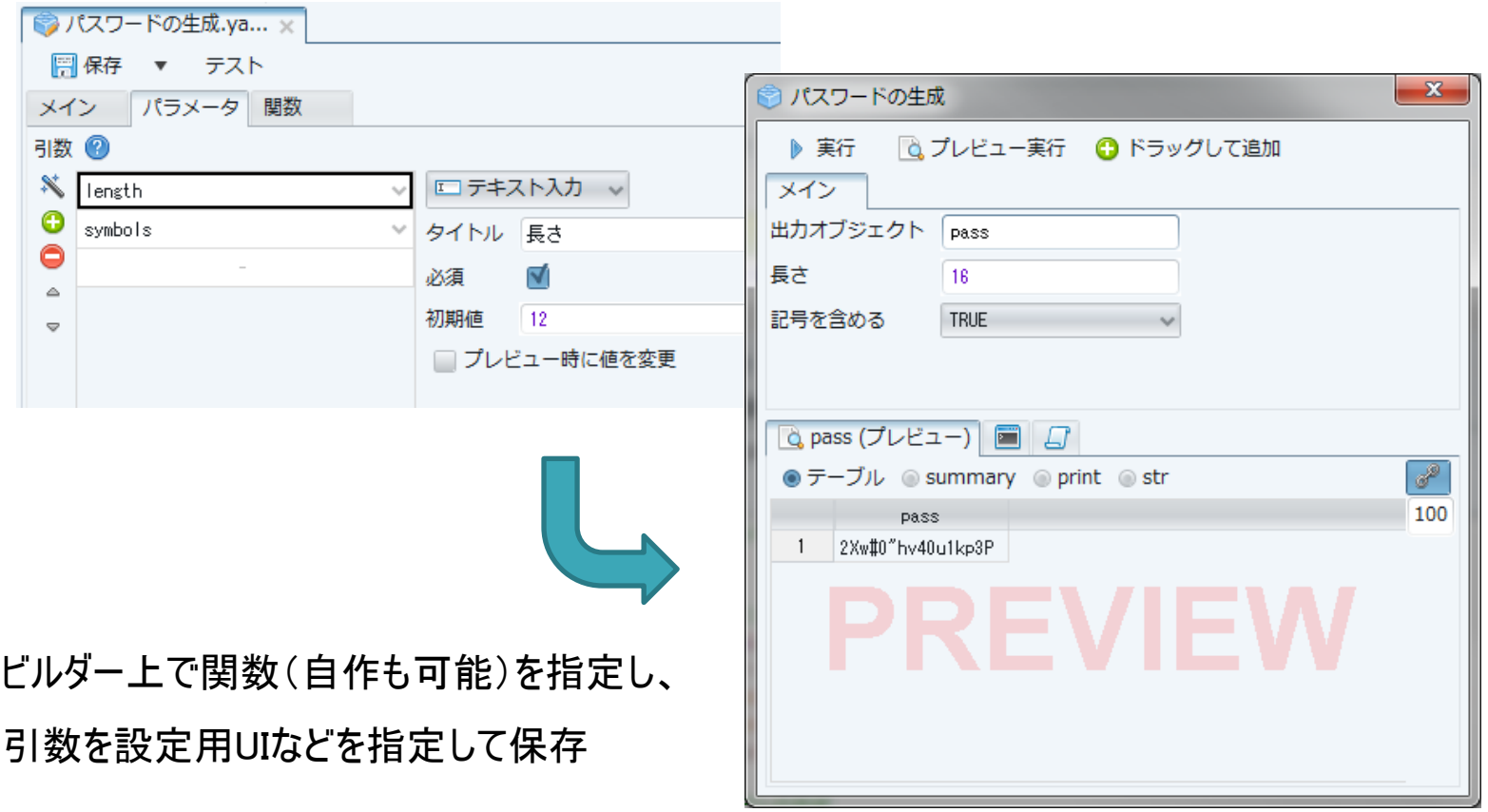
	Accuracy
ニューラル	0.9409449
ロジスティック	0.9251969
決定木	0.9133858

Predict x Actual

	0	1
0	235	9
1	6	4

機能例：カスタムUI

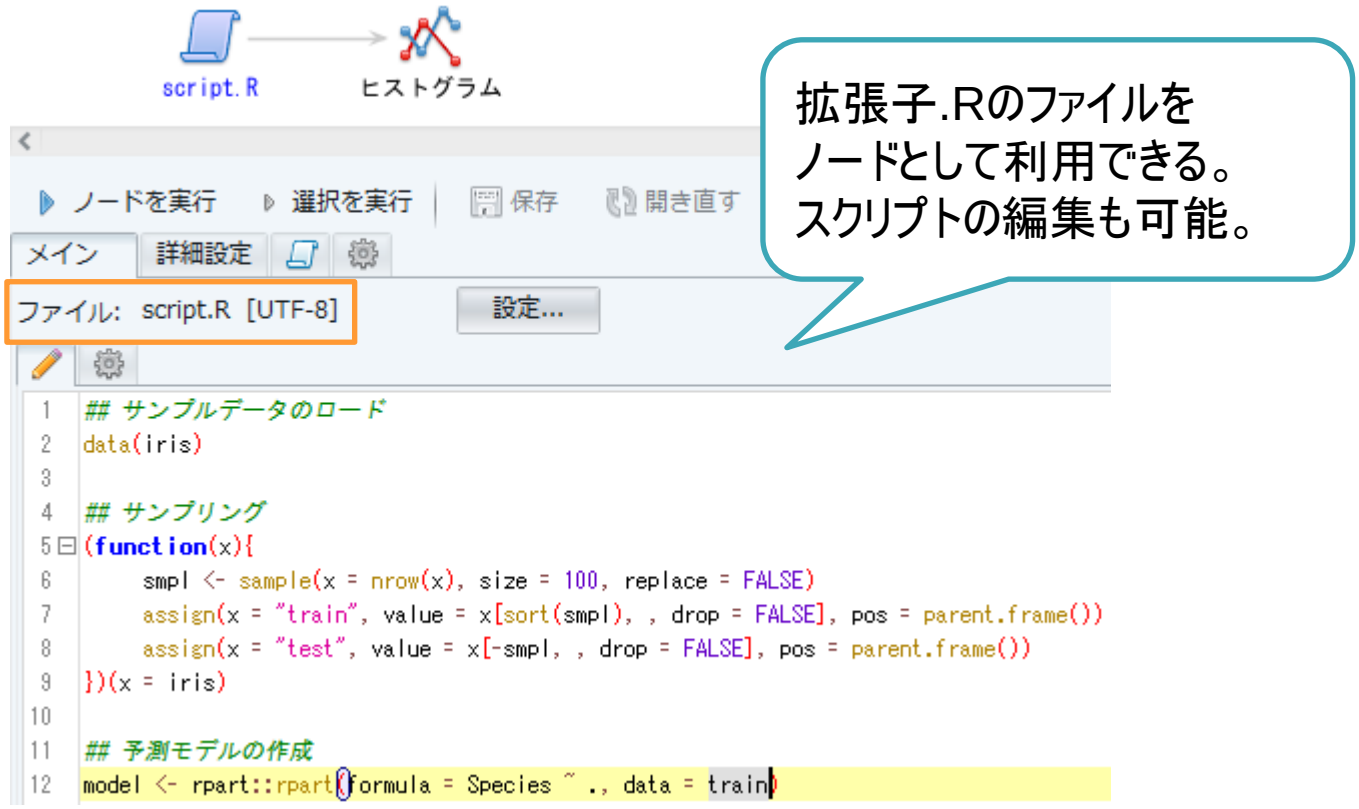
組み込みのGUIに存在しない機能を追加できる。
Javaプログラミングを行わず、ビルダーを使って簡単に作成可能



ビルダー上で関数（自作も可能）を指定し、
引数を設定用UIなどを指定して保存

機能例：外部ツール連携

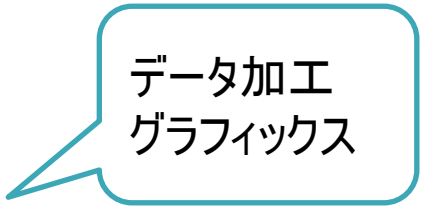
IDEが使用するフォルダをプロジェクトフォルダとして設定し、
Rスクリプトファイルをノードとして利用可能



設計方針と問題点

- 方針：生成されるコードが「どこでも確実に動く」
  - 外部パッケージに依存せず、標準パッケージ群（base, recommended）のみで動くスクリプトを生成
  - パッケージを使用するスクリプトをユーザーが書くことはできるが、GUIで生成するコードには含めない

- 問題点
  - モダンなRのデータ分析はパッケージに強く依存
    - tidyverse（dplyr, ggplot2, tidyr, ...）
    - 予測モデル（glmnet, ranger, xgboost, ...）
  - かなりメジャーなパッケージでも、標準配布される様子もない



新しい方針

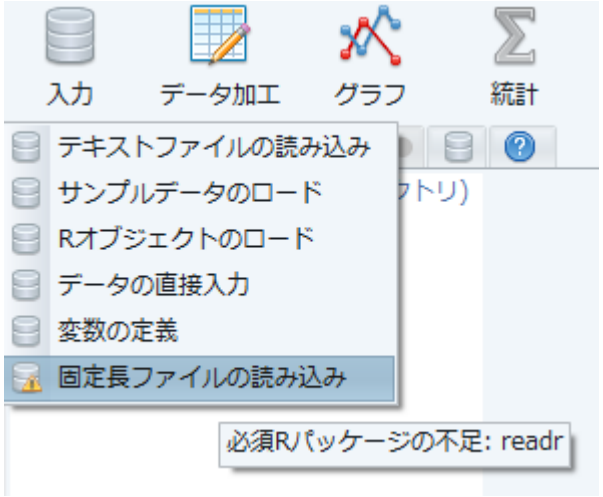
基本方針は
変えない！

- 方針：生成されるコードが「どこでも確実に動く」

- 実現手段
  - 必要なパッケージを検出して通知
  - パッケージのインストールと管理をサポート
  - 必要であれば事前にパッケージをロード

必要なパッケージを検出して通知

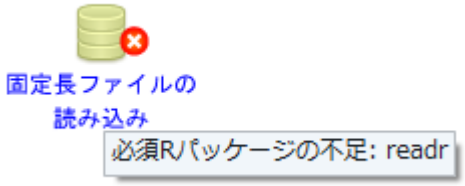
- メニューでの表示
  - 分析機能を呼び出す前にパッケージの有無をチェック



ef-prime, inc.

必要なパッケージを検出して通知

- 分析フロー上での表示
  - ノードごとに必要なパッケージを管理し、実行前に通知
  - ユーザーの記述したスクリプトからもパッケージを検出



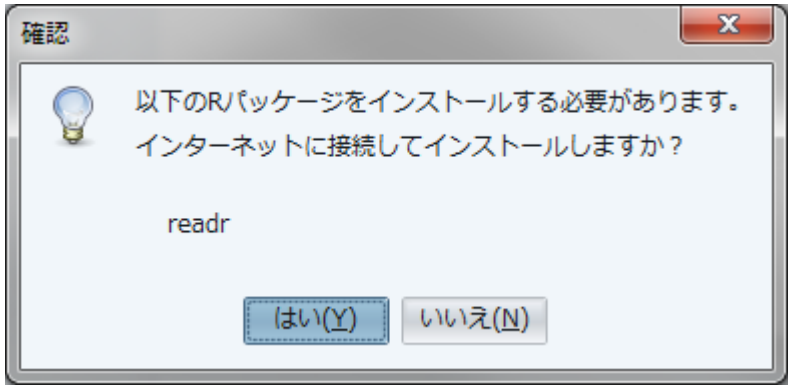

```
base::parse
utils::getParseData
```

- スクリプトのエクスポート
  - 分析フローをRスクリプトに出力する際、パッケージをチェックするコードを含めるオプションを作成。
  - 必須パッケージがない場合、実行後すぐにエラーで停止

```
Error: Missing required package: readr. Try:
   install.packages("readr")
```

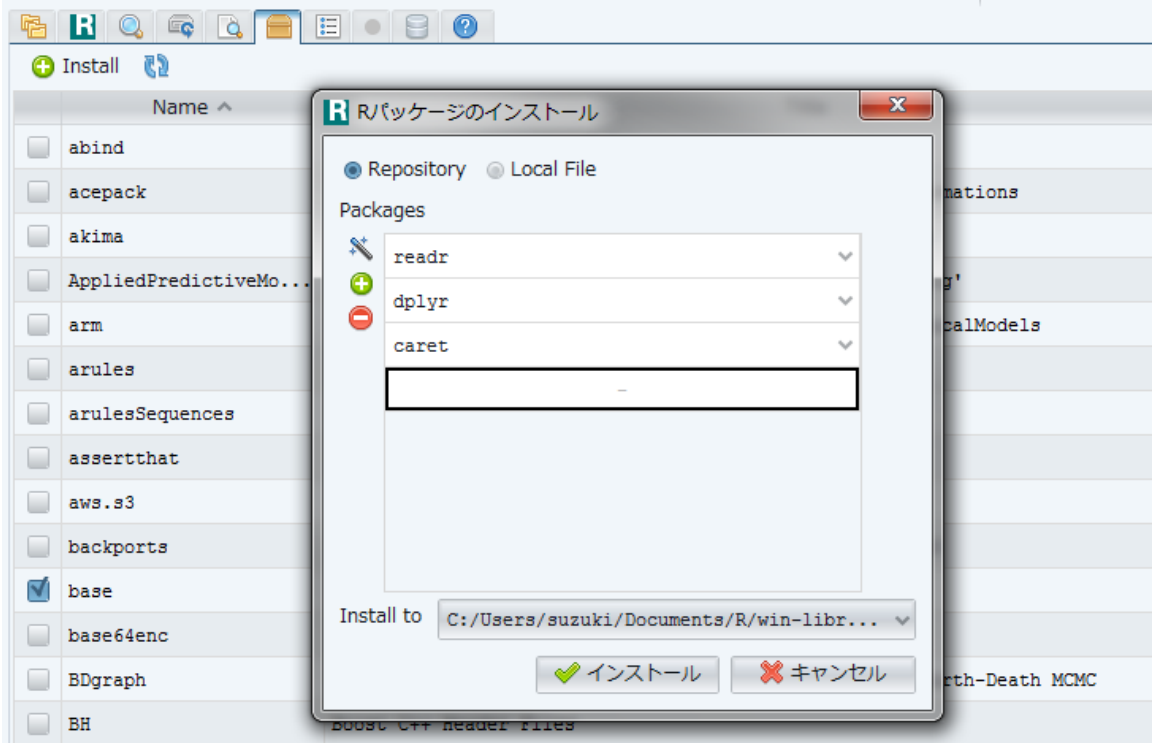
パッケージのインストールと管理をサポート

- インストールの提案
  - 不足するパッケージがあればダイアログを表示し、ワンクリックでインストール
    - 標準のCRANミラーを設定してあり、初心者にとっては選択に困るミラーサイトの選択も省略



パッケージのインストールと管理をサポート

- パッケージマネージャ
  - Rスクリプトを使わずにパッケージのインストールなどの管理を行うための仕組みを提供



ef-prime, inc.

必要であれば事前にパッケージをロード

- パッケージの自動ロード
  - 必要がある分析機能に限り、実行時に自動的にパッケージをロードするコードを生成
    - `library()` または `require()`
  - 関数をマスクするなどの意図しない副作用を起こす可能性があるため、できる限りアタッチせず最低限の利用とする方向で検討
    - 基本的には `package::function()` の形
    - 読みづらくなるtidyverse系やmagrittrなどのパイプの扱いは検討中

今後について

- 開発状況
  - パッケージの通知、管理の仕組みを開発中
  - パッケージを利用した分析機能を開発中
    - プロット（ggplot2）
    - データの加工と集計（dplyr + tidyr）
    - モデリング（glmnet, ranger, xgboost）

Java + R な開発者の方、ご興味ありましたらご一報を！

ご清聴ありがとうございました

ef-prime, inc.

R AnalyticFlow

https://r.analyticflow.com

@ef-prime_jp　R AnalyticFlow