2009.12.5.SAT

オープンなビジネスデータ分析環境

株式会社 ef-prime
鈴木 了太
suzuki@ef-prime.com

ef-prime, inc.

わたしたちはなにものか？

- 株式会社 ef-prime
  - 2006年3月設立。所在地は東京都中央区
  - 業務内容:
    - 企業向けデータ分析コンサルティング
    - ソフトウェア受託開発
    - データ分析トレーニング
    - その他ソフトウェアの開発、公開

わたしたちはなにものか？

- （ちなみに）わたしはだれか？
  - 代表取締役 / データアナリスト
  - 主な仕事
    - コンサルティング
    - データ分析
    - プロジェクトマネージャー
    - ソフトウェアデザイン
    - 研究開発
    - 営業
  - 他のスタッフにやってもらわないと困ること
    - プログラミング（特にJavaとかC#）
    - もちろんデータ分析もチームでやります
    - ほかにもいろいろ

本日のお題

- オープンなビジネスデータ分析環境
  - Rをはじめとするオープンなツールがもたらす可能性
  - わたしたちのオープンなツールとの関わり方

・ ※ちょっと注意
  - 今回は「オープン」「フリー」といった言葉を厳密に定義しません
  - Open Source Initiative (OSI) が定めるところの「オープンソース」が当てはまればよいですが、そうでなくても構わない場合もあります

ef-prime, inc.

ef-prime, inc.

と、その前に…

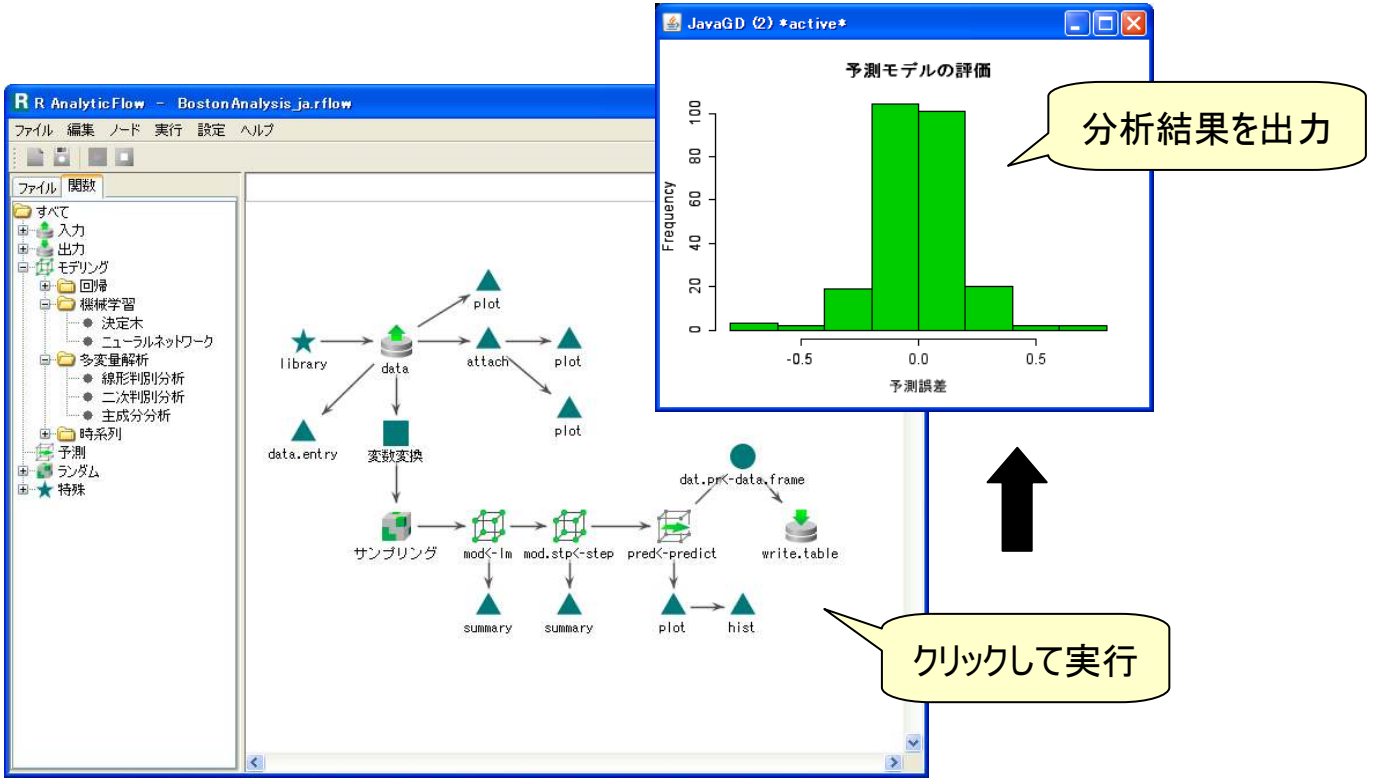
R AnalyticFlow 新作発表会

バージョン1.0 新機能紹介

ef-prime, inc.

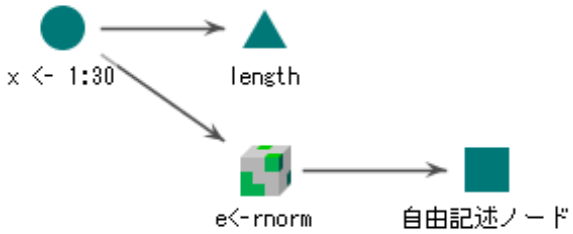
R AnalyticFlow について

- フローチャート形式のJava製R GUIフロントエンド
  - 多くのオープンソースソフトウェアを利用、RとはJRIで接続
  - RAF本体もオープンソース（BSDライセンス）



R AnalyticFlow について

- 特徴
  - 分析過程をフローチャート形式で記述
    - 思考が整理できる
      - 分析の「本流」と「支流」を分離
    - 作業グループでの共有がしやすい
    - 再実行がカンタン
      - Rの使い方を知らなくても、右クリックして「実行」するだけ！



  - 小回りが利く
    - コンソールからの実行も可能
      - ちょっとした確認などはコンソールで
      - コンソールの実行結果からフローを作ることもできる

R AnalyticFlow について

- これまでの歩み
  - 2007年12月公開
    - 初公開はRユーザ会！
  - 累計ダウンロード数は1300以上
    - 2009年9月まで、ソースコードは除く（実行可能形式のみ）
  - 現行バージョンは0.3.1
    - マルチOS対応
      - Windows版、Linux版（MacOSXでも動作確認済み）
    - 多言語対応
      - 現状、日本語と英語が利用可能

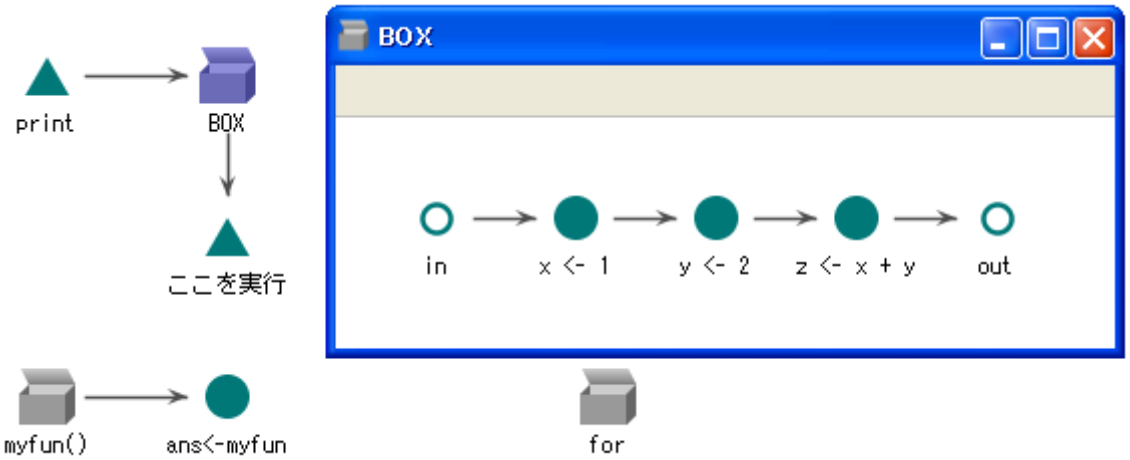
2008年3月以来、久々のアップデート！

新バージョンについて

- バージョン1.0の特徴
  - 新機能
    - ボックス
    - キャッシュ
    - オブジェクトブラウザ
    - 他にもいろいろ
  - 各種OSに対応
    - Windows 7, Vista, MacOSX
      - MacOSXは.app形式で配布の予定
  - 最新のRに対応
    - 2.10.xまで対応
      - Windowsでは従来必要だった「ライブラリパック」を廃止し、よりシンプルに

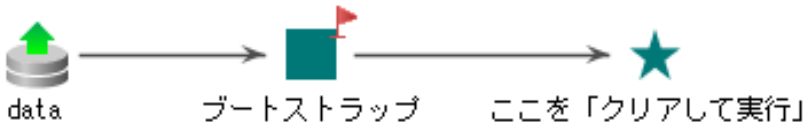
新機能の紹介

- ボックス
  - 部分フローを格納できる特殊なノード
  - 入れ子もOK
  - 関数やループを表現可能な「特殊ボックス」



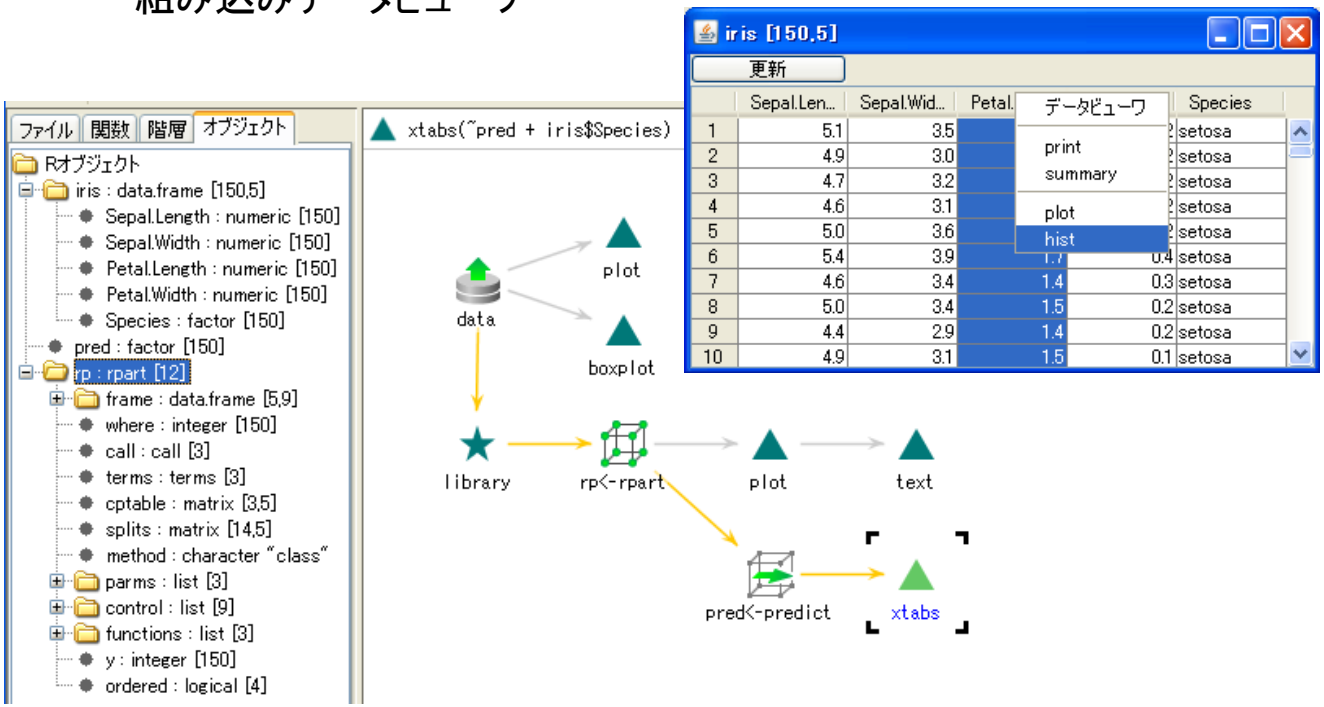
新機能の紹介

- キャッシュ
  - 実行結果を保存し、次回以降は自動的に結果を呼び出し
    - 時間のかかる計算を何度も行わなくてよい
  - 右クリックして設定するだけの簡単操作
    - 上流のコードが変更されたり、実行結果ファイルが書き換えられるとキャッシュは自動的にリセットされる
    - 「この.Rdata、いつ出力したんだっけ？」と気にしなくてよい



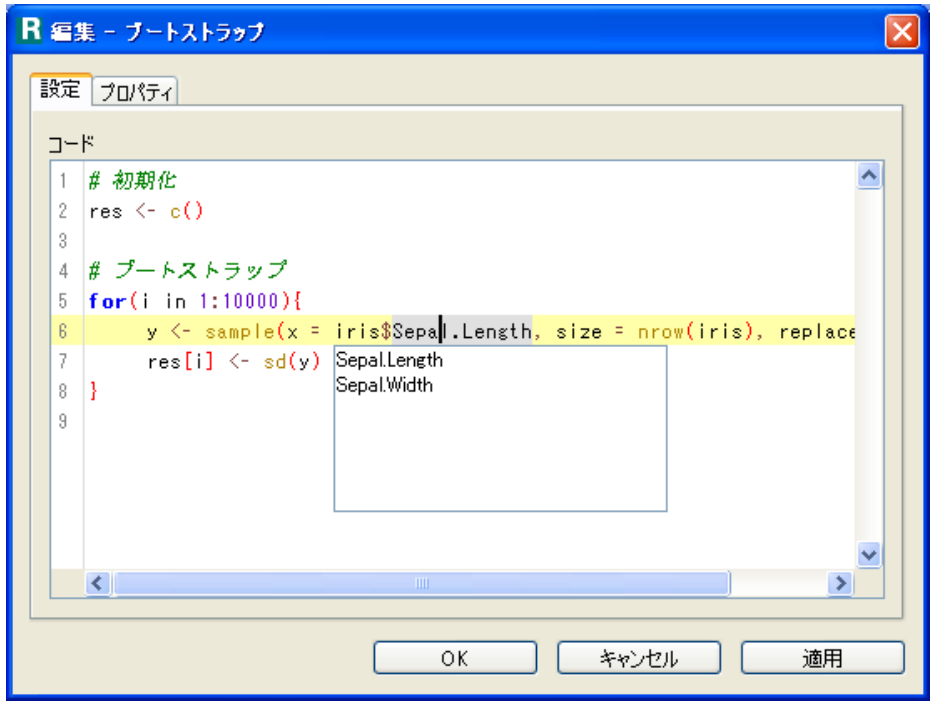
新機能の紹介

- オブジェクトブラウザ
  - ワークスペース内のオブジェクトをツリー形式で表示
    - オブジェクト内の階層もツリー形式
    - マウス操作でプロットなども可能
    - 組み込みデータビューワ



新機能の紹介

- コードエディタ
  - コードの色分け、ハイライトなどが可能
    - オープンソースのJavaライブラリ「RSyntaxTextArea」により実現
    - 自由記述ノード（複数行のコードを記述できるノード）で利用可能



新バージョンの公開

- 年内に公開予定
  - プログラム、ドキュメントともに完成
    - ウェブサイト、ライセンス文書などの更新作業が残っています
  - 12月中旬あたりを目指しています

- みなさまへのお願い
  - バグ報告、ご要望などをお寄せください！
    - RjpWikiにR AnalyticFlowページがあります
      - 岡田先生、いつもありがとうございます！
    - 特にLinux、MacOSX、Windows 7/Vistaは社内ユーザーがいないため、手薄になりやすい部分です
      - Javaなので基本的な動作には問題ないはずです
    - 「動かない」「こうしたら動いた」などご報告ください

更新情報をチェック！

http://www.ef-prime.com/

または

「**R AnalyticFlow**」「**ef-prime**」で検索

RSSで更新情報を配信しています。
RjpWikiでも情報公開させていただきます。

ef-prime, inc.

さて、本題です。

なぜ「オープン」がいいのか？

- やりたいことができる
  - すでにあるものを組み合わせて、素早くやりたいことに辿りつける
    - Rの場合
      - とりあえずCRANのパッケージを探してみる
      - なければ自分で作る。その基礎となる機能は大体すでにある
    - R AnalyticFlowの場合
      - 基本となる統計エンジンはR
      - グラフ表示にJUNG、コードエディタにRSyntaxTextArea…
      - コードはEclipse上でJavaで書いて、インストーラはNSISで…
  - 「ちょっとここだけ変えたい」ならもっとカンタン
    - Rはここが特に優れている

ef-prime, inc.

「クローズド」はだめなのか？

- うまく併用すればよい
  - 商用ツールにも素晴らしいものがたくさんある
    - 売れるものには売れるだけの理由がある
    - わたしたちも商用ツールを使います
      - 費用対効果で見合うものであれば、買ったほうがよいことも多い
  - ただし、自由度は限られる
    - 機能追加や改良が難しい
      - 高価なアドインや上位バージョンが必要、そもそもできない、など
    - ライセンスに縛られる
      - もう1ライセンスあれば便利だけど、高すぎて1本しか買えない
      - コンサルティングの場合、お客さまにもソフトを買ってもらわないといけない
  - 要するに適材適所
    - 世の中に選択肢が多いことはよいこと

「オープン」であることの安心

- 「使えなくなる」ことがない（起こりにくい）
  - いま動いているものはそのまま使える
    - ベンダーの都合で手に入らなくなる心配がない
    - もうひとつ必要なら複製もできる
  - 同じ仕組みで動くものを作ることができる
    - 動作原理はソースやドキュメントを見ればわかる
      - Rの場合、ヘルプを見れば実装の元になった論文までわかる
    - たとえ将来コンピュータ環境が大きく変わっても、対策をとることができる

▶ 重要なものだからこそ「オープン」がよい場合もある！

わたしたちと「オープン」の関わり方

- 使う、作る、支援する
  - 使う
    - Rをはじめ、多くのソフトウェアを使わせていただいています
    - Rコアメンバーはもちろん、多くの方の努力に支えられています。
      この場をお借りしてお礼申し上げます。ありがとうございます
  - 作る
    - R AnalyticFlowをオープンソースで公開しています
    - R AnalyticFlowのほかに、Nattoという分析ツールも公開しています
      - こちらはオープンソースではありませんが、無償でご利用いただけます
  - 支援する
    - ささやかながらR Foundationに寄付しています
      - Supporting Institutionsとして登録されています

むすびに

私たちの人生の多くが、Rをはじめとする
オープンなソフトウェアに支えられています。

素晴らしいソフトウェアを開発、公開、
そしてサポートしてくださっているすべての方々に
改めて心から感謝いたします。

ありがとうございます！

ef-prime, inc.

おまけ

新しい試み

- 大規模データ加工ツール
  - Rに限らず、大規模データを扱うことは悩みのタネ
  - 加工・集計してしまえばサイズが小さくなる
    （したがってRでサクサク分析できる）場合も多い
    - 例：ウェブのアクセスログデータ
      - 1ユーザ単位、1ユーザグループ単位などに集計して
        相関関係などを見る分析では、数千万、数億行のデータでも
        せいぜいユーザ数ぶんの行数にしかならない
  - オープンな技術に基づいて、
    大規模データをサクサク処理できるツールを開発中！

ef-prime, inc.