2007.12.8.SAT

ビジネスデータ分析の現場から

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

ef-prime, inc.

ビジネスデータ分析の現場から

- わたしたちについて
- ビジネスデータ分析
  - ビジネスデータ分析のプロセス
  - 「現場」に理論は必要？
  - どんな能力が必要？
- ビジネスのための分析ツール
  - どんなものが「使える」のか？
  - Rは「使える」のか？
  - わたしたちのアプローチ

ef-prime, inc.

はじめに：わたしたちについて

- 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

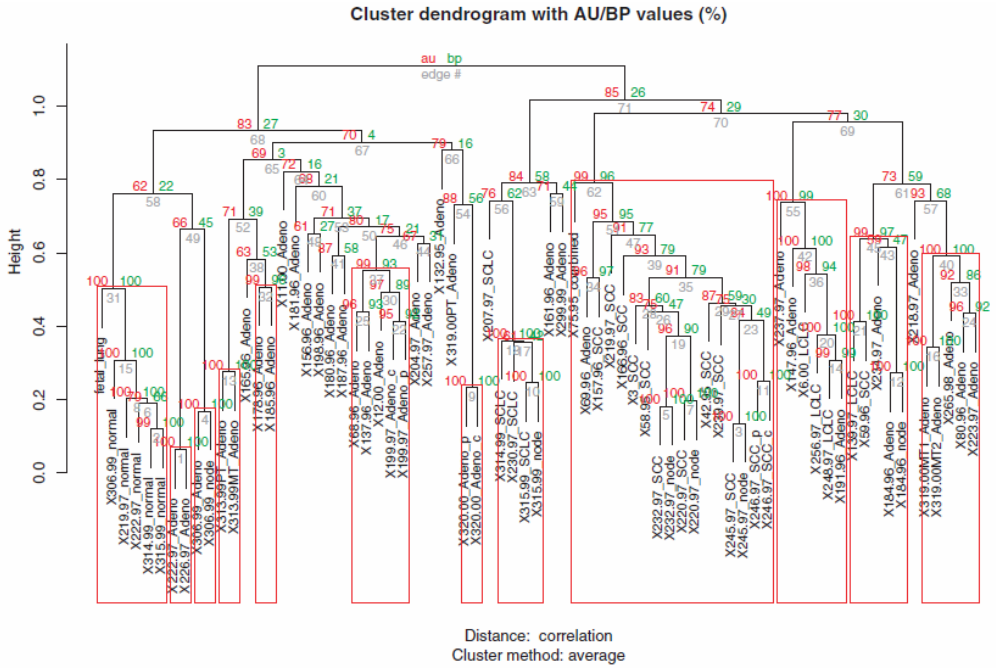
- わたしについて
  - 役職：ef-prime代表取締役
  - 学歴：
    - 2003年、一橋大学商学部 卒業（大上慎吾ゼミナール）
    - 2005年、東京工業大学 情報理工学研究科 数理・計算科学専攻 修了（下平英寿研究室） 修士（理学）
  - 職歴：
    - 在学中：勉強・研究の傍ら、こっそり委託業務を請け負う
    - 2005年4月～：マーケティングデータ分析業務に従事
    - 2006年3月～：株式会社ef-prime設立。代表取締役社長

わたしとR：パッケージ開発

- pvclust: 階層型クラスタリングの信頼性評価
  - 階層型クラスター分析の結果を統計的に評価するパッケージです。
  - 下平英寿先生（東工大）が開発された「マルチスケールブートストラップ法」を実装し、クラスターが存在するかどうかの確率値（*p*–value）を計算することができます。
  - 論文がpublishされ、すでにBioinformaticsの分野でいくつか応用例が出ています。
    - Suzuki and Shimodaira (2006),
      pvclust: an R package for assessing the uncertainty in hierarchical clustering.
      Bioinformatics.

pvclustのご紹介

- 階層型クラスタリングの信頼性評価
  - 変数間の関係を階層型クラスター分析で分析します。
  - 得られた各クラスターについて、「母集団分布においても同じクラスターが得られるかどうか」という仮説を検定し、確率値（*p*-value）を計算します。
  - CRAN公式パッケージのひとつで、Task ViewのClusterにも登録されています。
  - snow, Rmpiパッケージをインストールすれば並列計算も可能です。
    ほぼマシン台数分の速度向上を得られます。

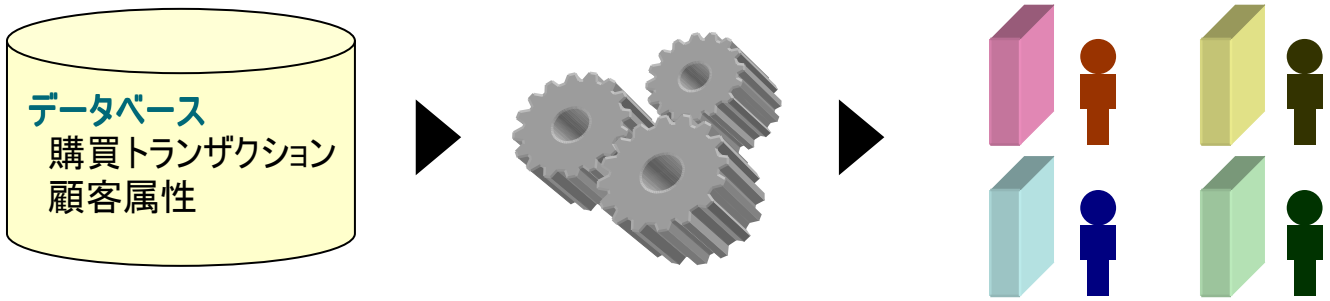


ビジネスデータ分析とは？

- マーケティングのためのデータ分析
  - 目的：企業の市場に対するアプローチを最適化すること
  - 例えば…
    - 誰に、どんな製品を、どうやって売るべきか？
    - 他社との競争に勝つためにはどうすればよいか？
  - 具体的には、
    - 「顧客ごとのおすすめ商品リスト」を作りたい
    - 「ダイレクトメールを送付したときの売上」をシミュレーションしたい
    - 「自社の商品を購入する際に重要な要素」を知りたい

事例紹介

- 広告配信カスタマイズ
  - データに基づき、顧客ごとに各商品の購入確率を予測
  - 顧客グループごとにおすすめ商品リストを作成し、広告として配信



ビジネスデータ分析のプロセス

- 目標達成のためのプロセス
  - ビジネスにおける具体的な目標を実現するための分析
    - 例：顧客のリピート率を向上したい、など
  - 目標達成までの1プロセスとしてのデータ分析
    - どのようなデータからどのような分析ができるのか
    - データ分析の役割をどこに位置づけるか

- プロセスの共有
  - プロジェクト関係者との共有
    - 企画者、データ作成者、分析者、最終的な意思決定者など
    - 多くの人と成果を共有する必要がある
  - 分析チーム内での共有
    - それぞれ知識レベルや得意分野は異なる
    - 分析プロセス自体も分担・共有が必要

「現場」に理論は必要？

- 総合的な理解が必要！
  - 分析の設計
    - ビジネス目的を具体的なデータ分析に落とし込むため
  - 安全性
    - オーバーフィッティングや手法の誤用を避け、安定した結果を得るため
  - 結果の理解
    - 分析の結果として得られた数値を理解し、説明・運用するため
  - 例えば…
    - 「新商品を顧客に売り込みたいが、なるべくコストを抑えたい」という要望
    - ほとんどが0となり、1が少ないダミー変数による予測モデリング
    - モデルは安定していて当てはまりもよいが、係数の符号が直感と異なる
    - サンプリングするたびに決定木モデルの形が全く異なる

どんな能力が必要？（こんな人、探してます）

- データ分析の理解
  - 基礎的な統計学の知識
    - 推定、検定、最尤法、bias-variance tradeoff、予測誤差とオーバーフィッティング、AIC、cross validation…
  - 分析アルゴリズムに関する知識
    - 回帰分析、一般化線形モデル、決定木、クラスター分析…

- ビジネスへの応用力
  - 具体的な問題とデータ分析手法のマッチング
    - 大学院での研究経験が意外と役に立ったり…
  - クライアントの「真のニーズ」を見抜く力
    - この目的なら検定の有意水準は多少粗くてもOK、ただし理解しやすいように結果はクロス集計（分割表）で…
    - でも裏ではロジスティック回帰でしっかり確率推定しておこう！

ビジネスのための分析ツール

- どんなものが「使える」のか？
  - 簡便性
    - 分析に高度な知識を要求しない
      - 企業に分析の専門家は少ない
    - インターフェースが充実している
      - 直感的なマウス操作
    - アウトプットがわかりやすい
      - 直感的なグラフや見やすい図表があれば、社内で共有しやすい
  - 機能
    - 必要な手法が一通り揃っている
    - 大規模データの分析ができる
  - 価格
    - 導入コスト、運用コストが安い

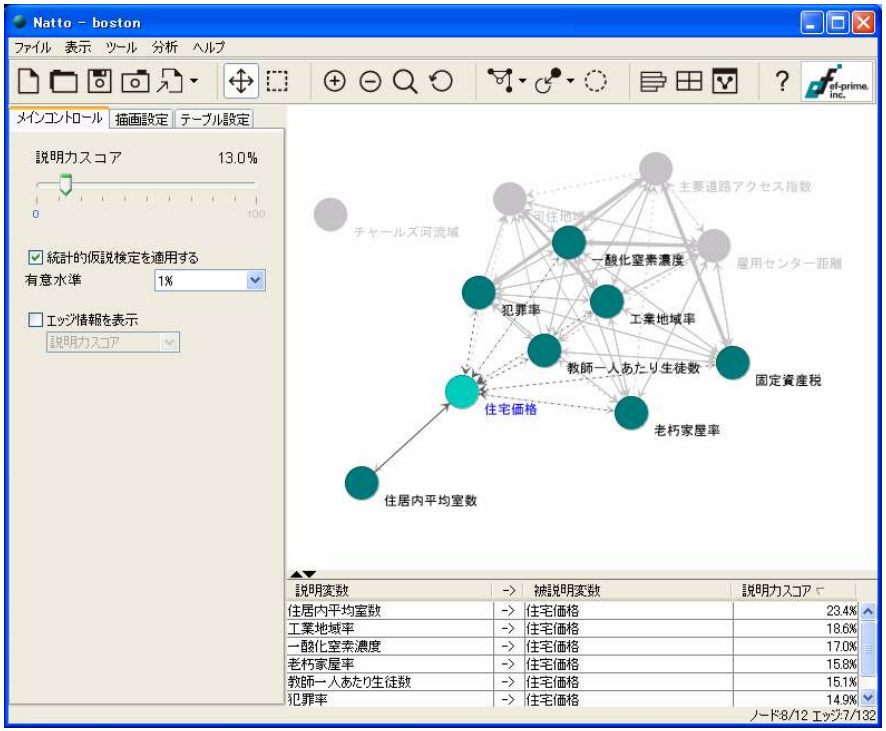
Rは「使える」のか？

- 安くて高機能、でも…
  - 簡便性　　×　難しい、不便
    - 知識なしではマニュアルも読めない
    - コマンド入力は慣れるまで大変
    - アウトプットが「ぶっきらぼう」、そのままでは使えない
  - 機能　　△　高機能、しかし…
    - 大量の追加パッケージで「なんでもできる」
    - 足りなければ自分で新機能を実装可能
    - 大規模データには弱い
  - 価格　　△　導入コストはゼロ、しかし…
    - フリーなので導入コストはゼロ
    - 実際に利用できるだけの知識を身につけるのが大変

ef-prime, inc.

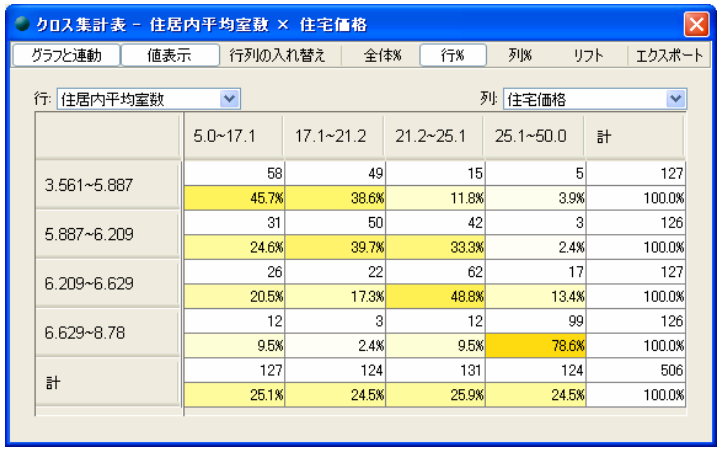
わたしたちのアプローチ①：独自ツール

- Natto：「簡単・便利」にこだわった分析ツール
  - 簡単・便利、でも安心
    - マウスによるグラフィカルで直感的な分析
    - ワンクリックで仮説検定、結果の信頼性を確保



わたしたちのアプローチ①：独自ツール

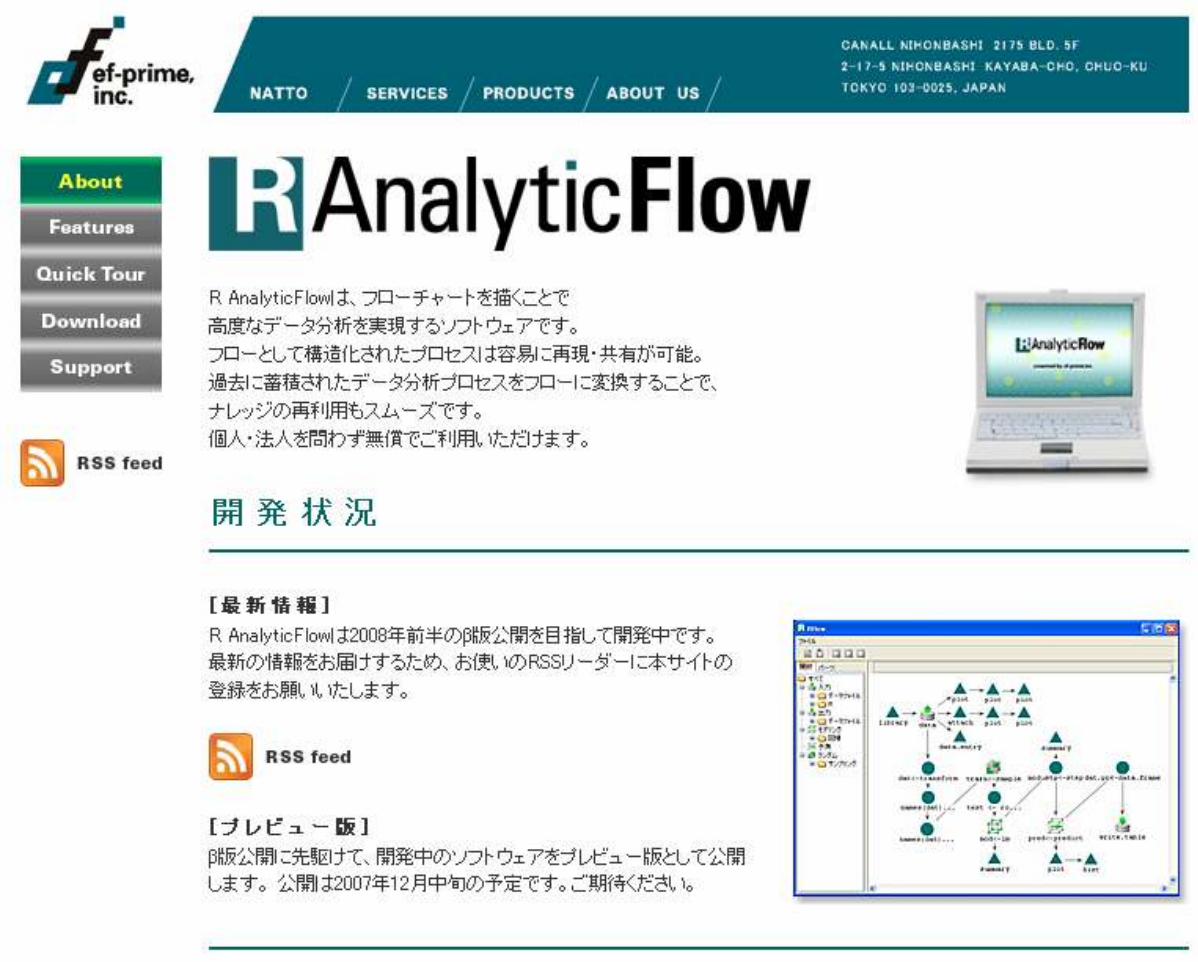
- 分析機能
  - 相互情報量に基づく変数間の相関分析で、非線形の関係も抽出
  - クロス集計と相関ルール探索が可能
  - 大規模データにも対応
    - 300変数、10万件程度のデータならPCレベルでも数分でセットアップ可能
    - 件数が増えると時間はかかるものの、100万件超でも分析可
- 無償で公開
  - 弊社サイト http://www.ef-prime.com/ からダウンロード可能
  - オープンソースライブラリ（JUNGなど）の活用で安価に開発



	5.0~17.1	17.1~21.2	21.2~25.1	25.1~50.0	計
3.561~5.887	58	49	15	5	127
	45.7%	38.6%	11.8%	3.9%	100.0%
5.887~6.209	31	50	42	3	126
	24.6%	39.7%	33.3%	2.4%	100.0%
6.209~6.629	26	22	62	17	127
	20.5%	17.3%	48.8%	13.4%	100.0%
6.629~8.78	12	3	12	99	126
	9.5%	2.4%	9.5%	78.6%	100.0%
計	127	124	131	124	506
	25.1%	24.5%	25.9%	24.5%	100.0%

A	->	B	確信度	支持度	リフト
犯罪率=3.693~88.976	->	固定資産税=666~711	99.2%	24.9%	3.66
固定資産税=666~711	->	犯罪率=3.693~88.976	91.2%	27.1%	3.66
固定資産税=666~711	->	工業地域率=18.1~27.74	100.0%	27.1%	2.68
犯罪率=3.693~88.976	->	工業地域率=18.1~27.74	100.0%	24.9%	2.68
犯罪率=3.693~88.976	->	教師一人あたり生徒数...	99.2%	24.9%	2.56
固定資産税=666~711	->	教師一人あたり生徒数...	96.4%	27.1%	2.49
工業地域率=18.1~27....	->	固定資産税=666~711	72.5%	37.4%	2.68
住宅価格=25.1~50.0	->	住居内平均室数=6.62...	79.8%	24.5%	3.21
一酸化窒素濃度=0.6...	->	工業地域率=18.1~27.74	91.9%	24.5%	2.46
住居内平均室数=6.6...	->	住宅価格=25.1~50.0	78.6%	24.9%	3.21
一酸化窒素濃度=0.6...	->	犯罪率=3.693~88.976	77.4%	24.5%	3.11
犯罪率=3.693~88.976	->	一酸化窒素濃度=0.63...	76.2%	24.9%	3.11
工業地域率=18.1~27....	->	犯罪率=3.693~88.976	66.7%	37.4%	2.68
一酸化窒素濃度=0.6...	->	固定資産税=666~711	79.0%	24.5%	2.92
教師一人あたり生徒...	->	固定資産税=666~711	67.3%	38.7%	2.49
固定資産税=666~711	->	一酸化窒素濃度=0.63...	71.5%	27.1%	2.92

わたしたちのアプローチ②：Rを機能拡張

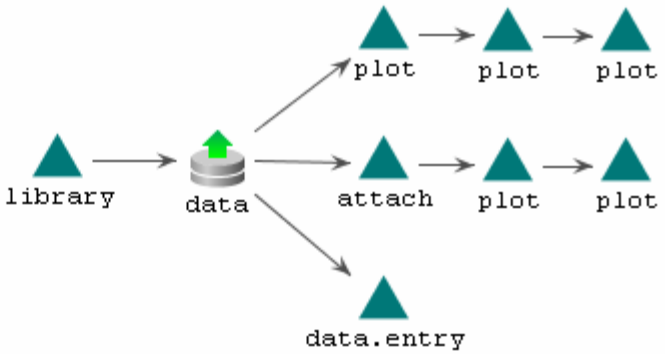


http://www.ef-prime.com/

R AnalyticFlowの特徴

[分析フローの記述]

R AnalyticFlowでは、データ分析のプロセスをフローチャートによって記述します。すべての分析はフローチャートに従って実行されるため、一度作成したプロセスは誰でも簡単に、かつ正確に再現することができます。



[R: 最高のデータ分析エンジン]

Rはオープンソースの統計的データ解析システムとして、世界中の専門家によって開発されています。有志によって提供された追加パッケージの数は1200を超え(2007/12現在)、あらゆる分野における最先端のデータ分析手法が実装されています。

R AnalyticFlowは、Rで利用できるほとんどの機能をフローチャートを通じて実行できます。最高のデータ分析エンジンを、より便利に扱いやすく。そのパフォーマンスを最大限に引き出します。

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

R AnalyticFlowの特徴

[蓄積されたナレッジの活用]

既にRをお使いですか？ R AnalyticFlowは、既存のRコードからフローチャートを生成することもできます。過去に記述したソースを読み返すのは意外と骨の折れる作業です。フローとして視覚化することで、当時の思考プロセスを取り戻しましょう。

作成したフローをソースに変換することも可能（※1）です。Rの優れた互換性を失うことなく、蓄積されたナレッジを再利用することができます。

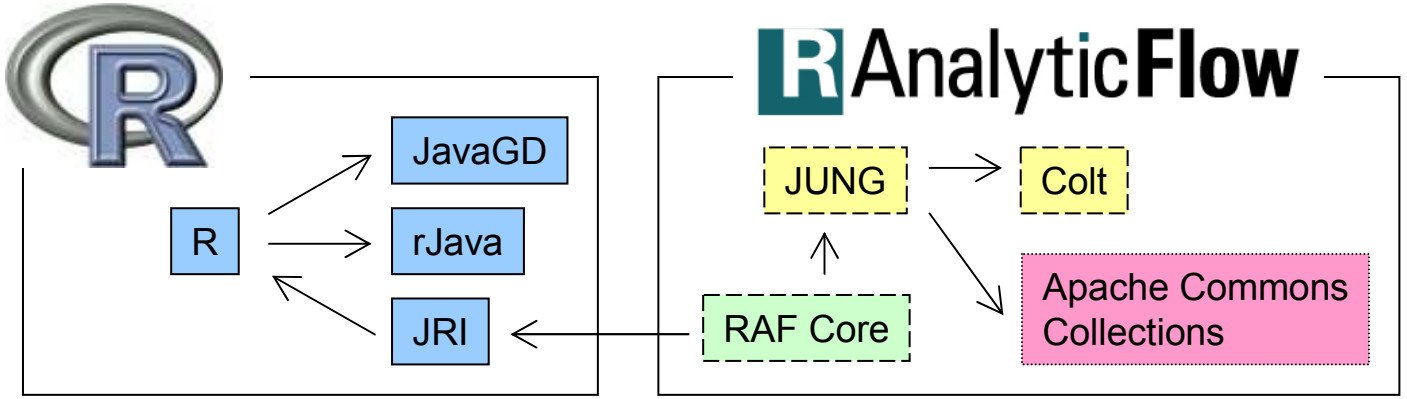
※1: プレビュー版では連続したフローのみ変換可能。将来的には全体の抽出が可能となる予定です。

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

システム構成

- Java
  - JRI（Java-R Interface）経由でRを起動
  - 関連パッケージはマルチバイトに対応できるよう修正（現状R-2.5.1のみ）

- オープンソース
  - 各種オープンソースライブラリの活用で効率的に開発
  - 弊社開発分のソースも公開予定



R AnalyticFlowが可能にすること

- 小回りの効くデータ分析
  - 「ここだけ実行」「ちょっと条件を追加」が可能
  - フローチャート形式を採用しながら、R特有のフットワークも両立
- 思考の整理
  - フローチャートに表すことで思考が整理される
  - 探索的分析と本筋のモデリングを分離できる
- 分析プロセスの共有
  - フローチャートを使って分析プロセスを説明できる
  - ソースを読むのに比べて、素早く全体を把握できる
  - 一度作成したフローはRを知らなくても実行可能

R AnalyticFlow リリース予定

- プレビュー版
  - 2007年12月中旬
  - Windows XP対応、要R-2.5.1
- β版
  - 2008年前半
  - Windows / Linux / Mac版

ウェブサイト:

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

私たちが目指すもの

- 手法
  - 「ビジネスのための」データ分析方法の開発
  - 理解しやすく、現象をよく表し、実用的な方法
- ツール
  - 「ビジネスで使える」データ分析ツールの開発
  - ビジネスのニーズに合った、「簡単・便利・安全」なツール
  - Rの機能性と拡張性をより使いやすく
- 人
  - 個人、そしてチームにより高い能力を
  - データ分析という「仕事」の確立

ef-prime, inc.

むすびに

Rをはじめデータ分析の発展に寄与していただいた全ての方々、
エキサイティングな仕事をいただいているクライアントの皆様、
オープンソースソフトウェアの開発者の方々、
本日お越しいただいた皆様、
諸先生方…

私たちにチャンスを下さった全ての方々に感謝いたします。

ご清聴ありがとうございました。

suzuki@ ef-prime.com