

Rubyでデータサイエンスを行う ための取り組み

@kozy2

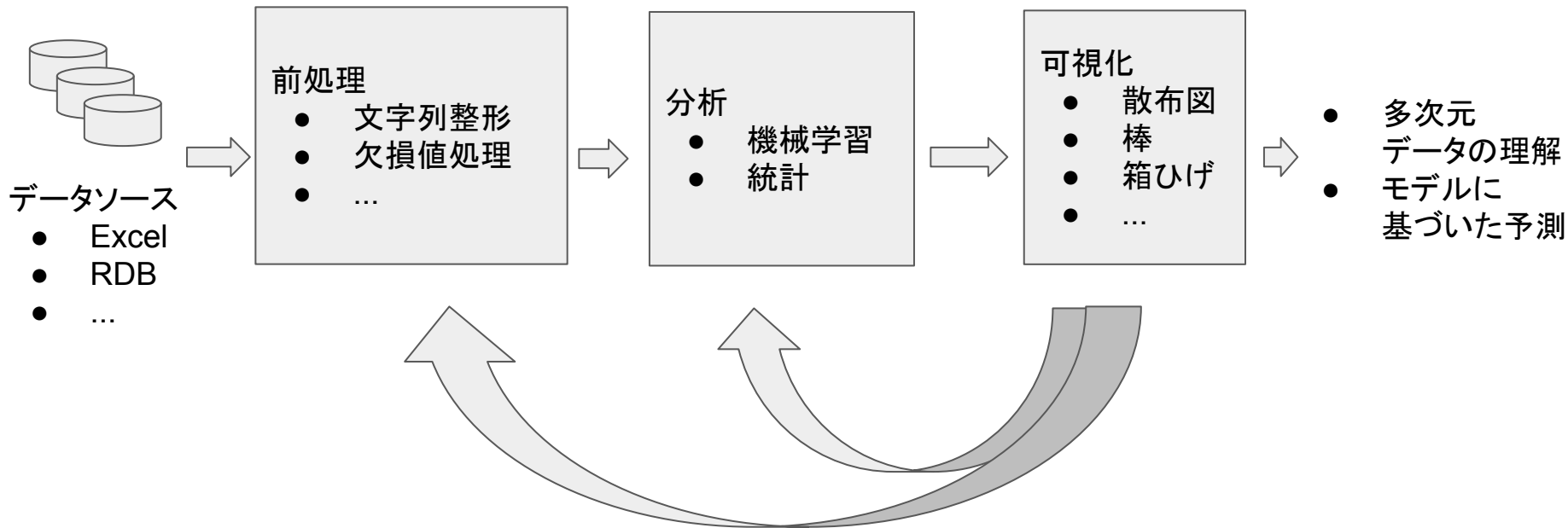
概要

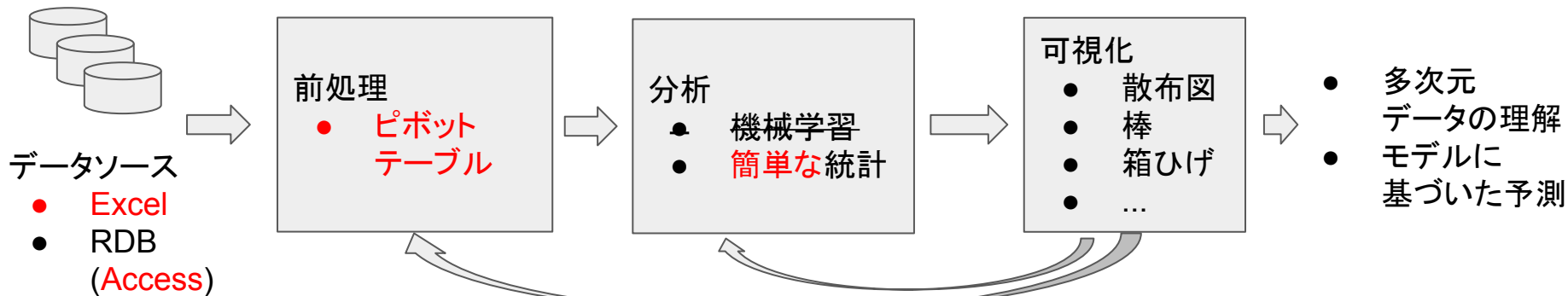
- データサイエンスは技術的にどうおもしろいか
 - Ruby business ?
- Rubyで現状何ができるか
 - Ruby business ?
- Rubyコミュニティに今後何を期待するか
 - Ruby community ?

概要

- データサイエンスは技術的にどうおもしろいか
 - Ruby business ?
- Rubyで現状何ができるか
 - Ruby business ?
- Rubyコミュニティに今後何を期待するか
 - Ruby community ?

データサイエンスとは





<https://products.office.com/en/excel>

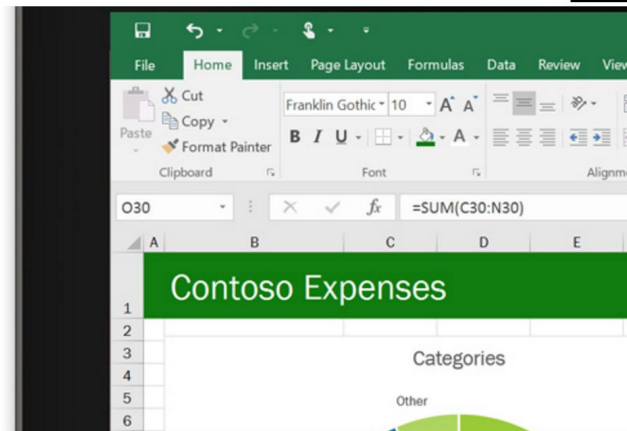
Microsoft Products Support
Office Products Templates Support



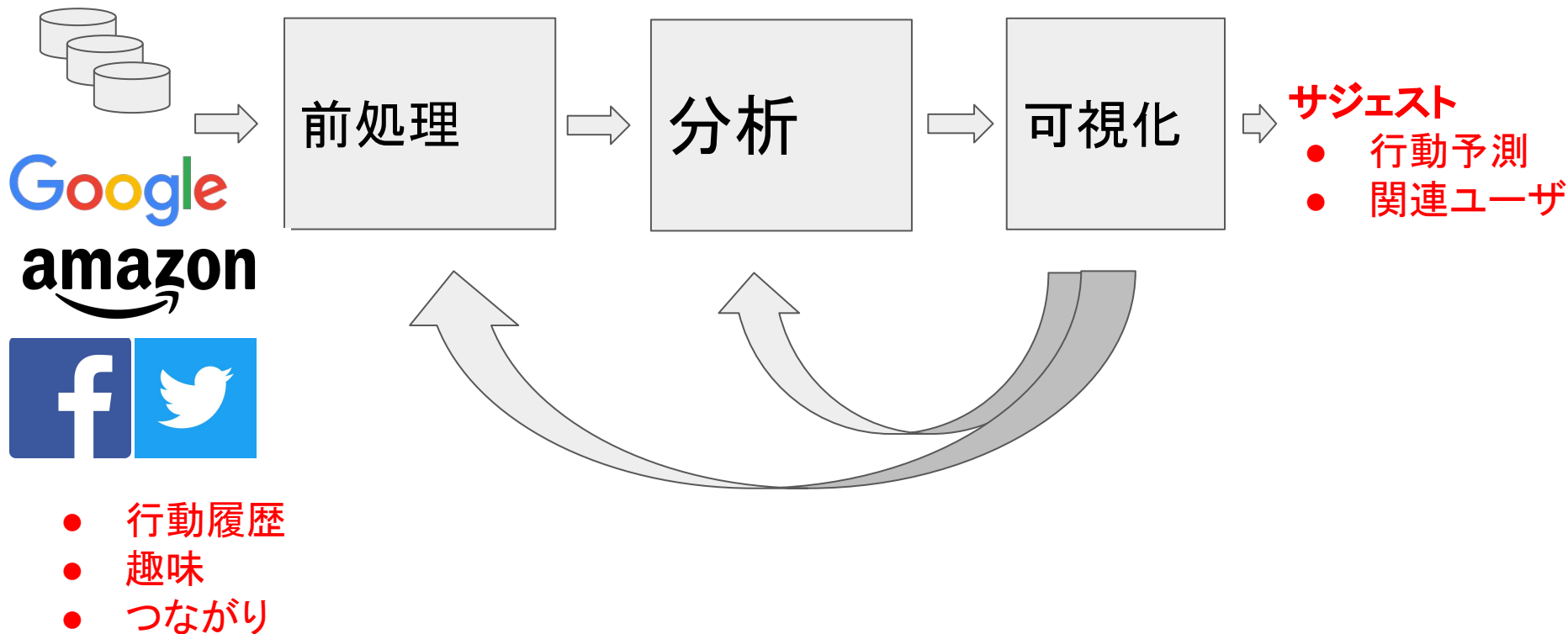
Excel 2016—get it now with an Office 365 subscription

Buy now

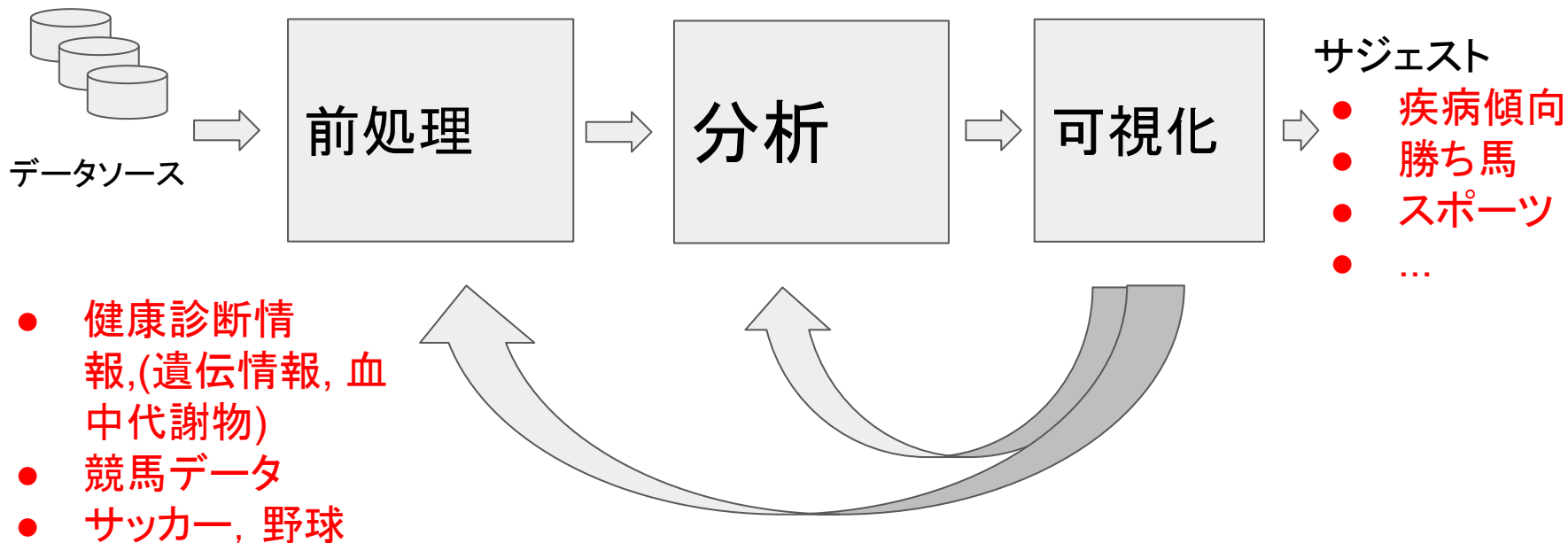
Try for free



Webサービスの具体例



実欲求に従う具体例



近年の状況

- データが増えている
 - 複数のデータソースを組み合わせる
- データ分析環境が進歩している
 - (オープンソース)ソフトウェア
 - より高速な前処理
 - 新しい分析技術の出現
 - ハードウェア
 - GPU, クラウド
- データサイエンスの民主化が進んでいる
 - オープンデータ
 - 情報共有, 切磋琢磨する場 (Kaggle, PyData)

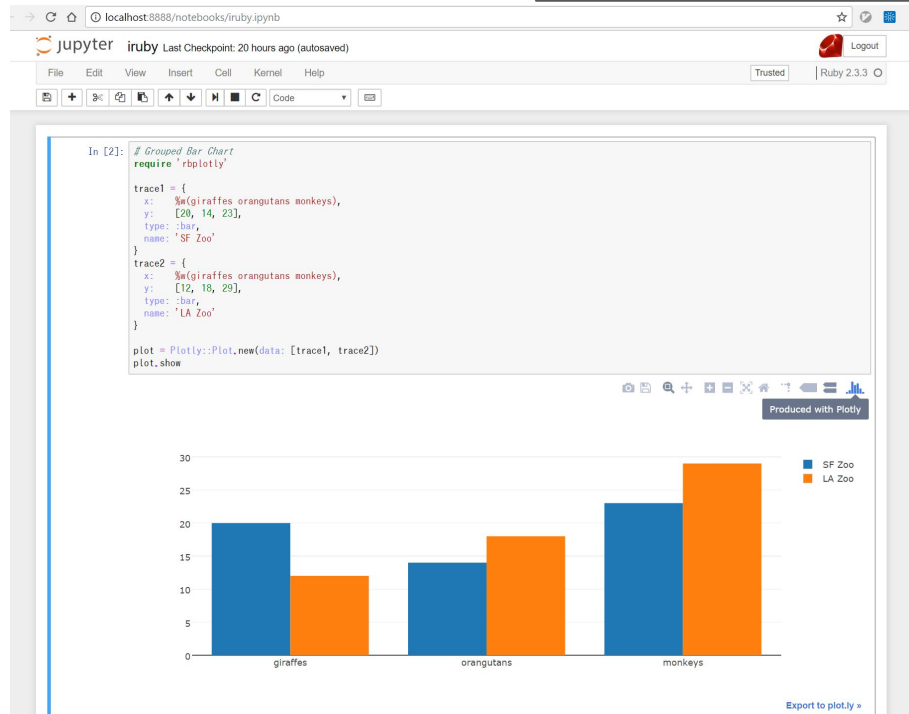
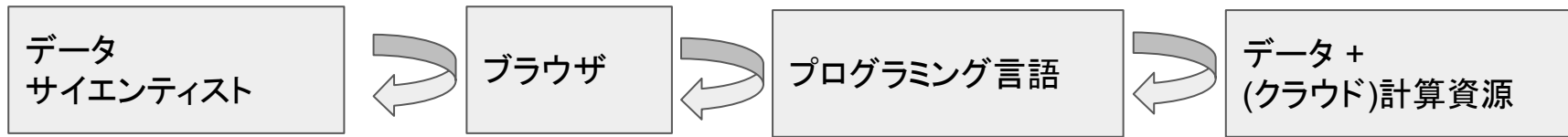
概要

- データサイエンスは技術的にどうおもしろいか
 - Ruby business ?
- Rubyで現状何ができるか
 - Ruby business ?
- Rubyコミュニティに今後何を期待するか
 - Ruby community ?

Ruby のデータサイエンスにおける現状

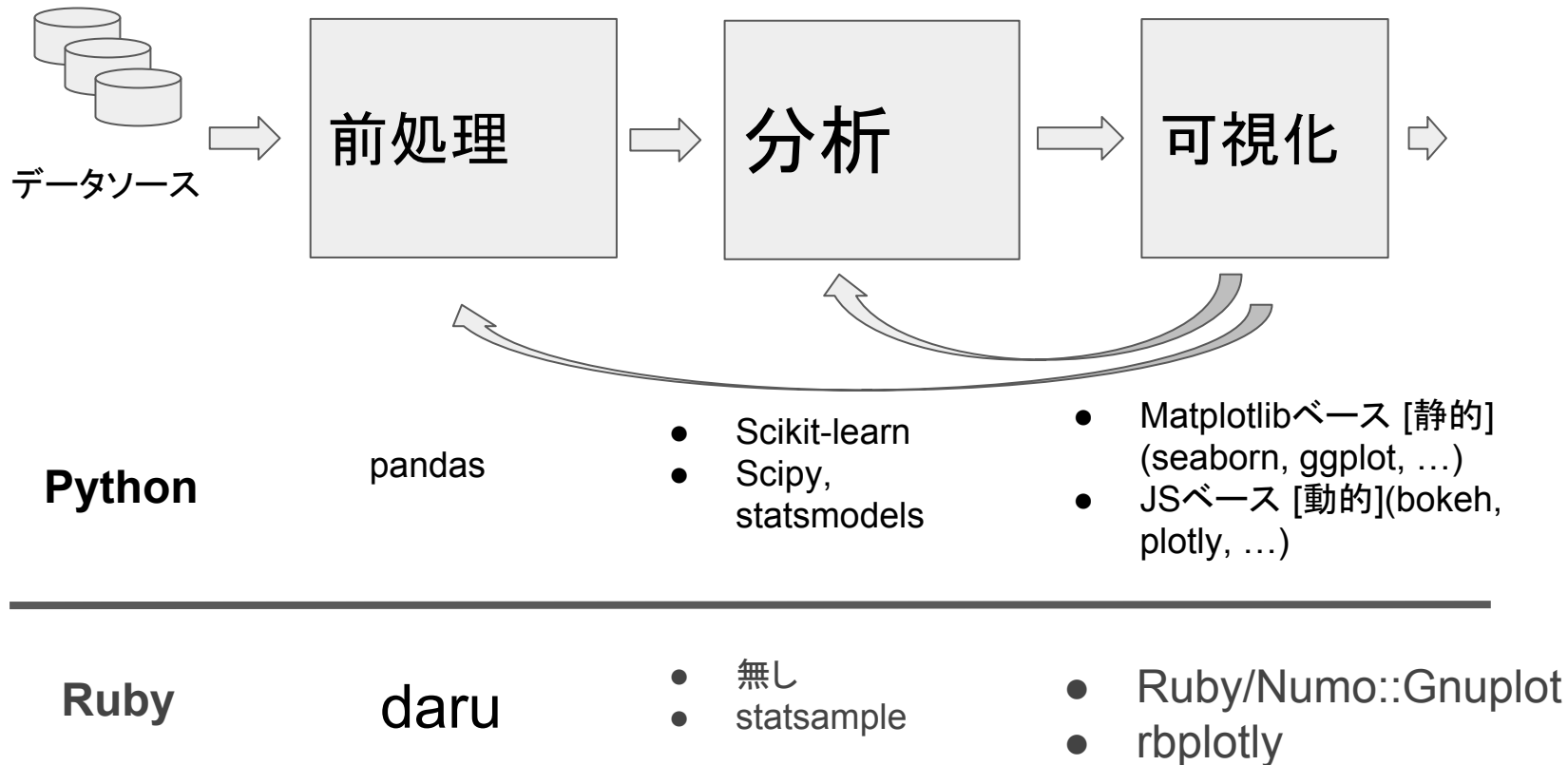
- 基本的な機能を備えたRubyのgem群
 - Daru (前処理)
 - [ruby-numo project のgem群] (分析)
 - Jupyter notebook + Ruby + plotly (可視化)
- Ruby外の言語との連携を推進するgem群
 - PyCall (PyData ソフトウェア群 との連携)
 - Red Data Tools (Apache Arrow を活用)

ウェブブラウザからのRubyの実行



- JS を用いた可視化
- インタラクティブ
- リモート計算資源

データサイエンスに関わるRubyのgem群



Docker を用いた試用方法

daruの使い方 や 基本的な統計ワークフロー の実例集

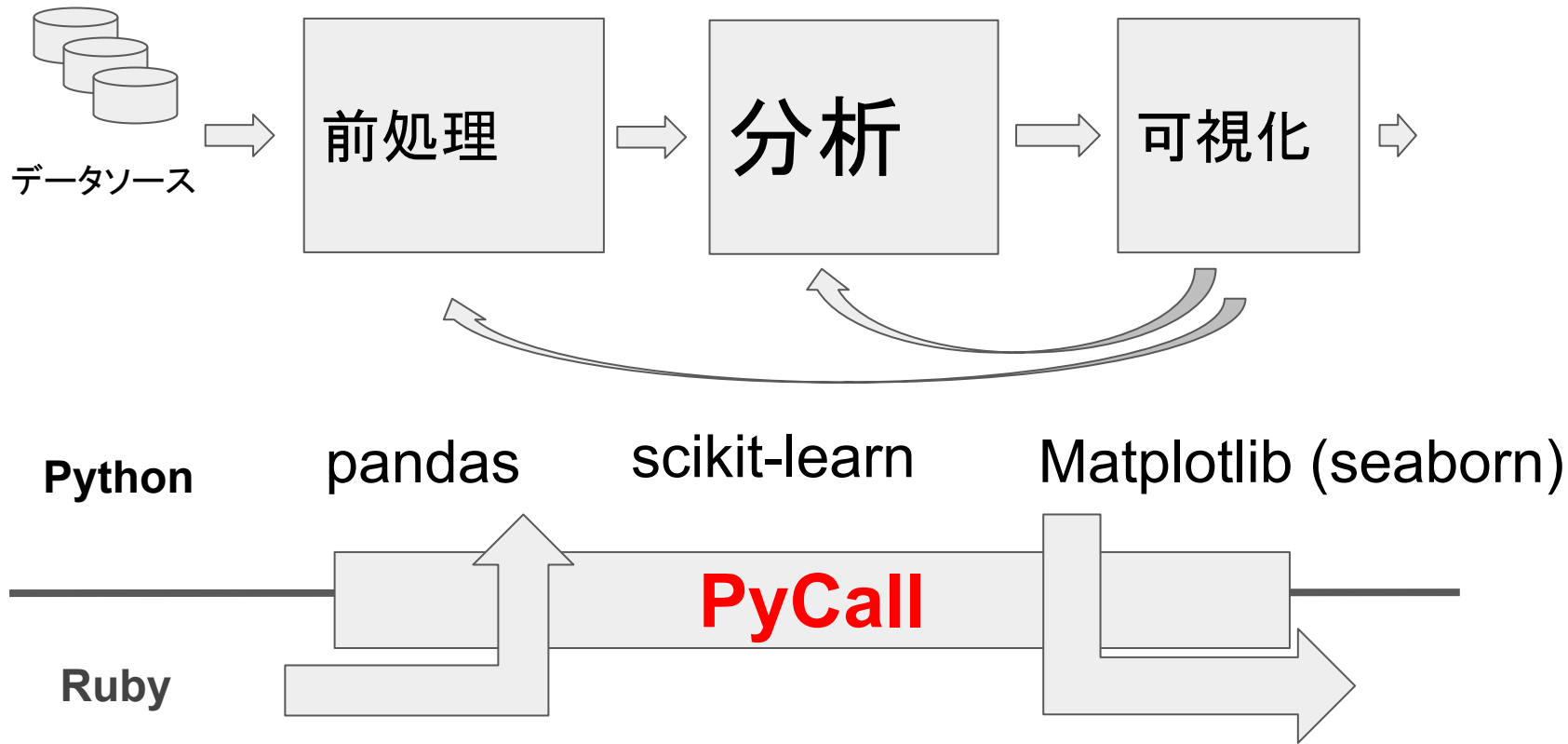
```
docker pull sciruby/ruby-datascience-examples
docker run -d -p 8888:8888 sciruby/ruby-datascience-examples start-notebook.sh --NotebookApp.token=''
```

Ruby/Numo::Gnuplot の実例集

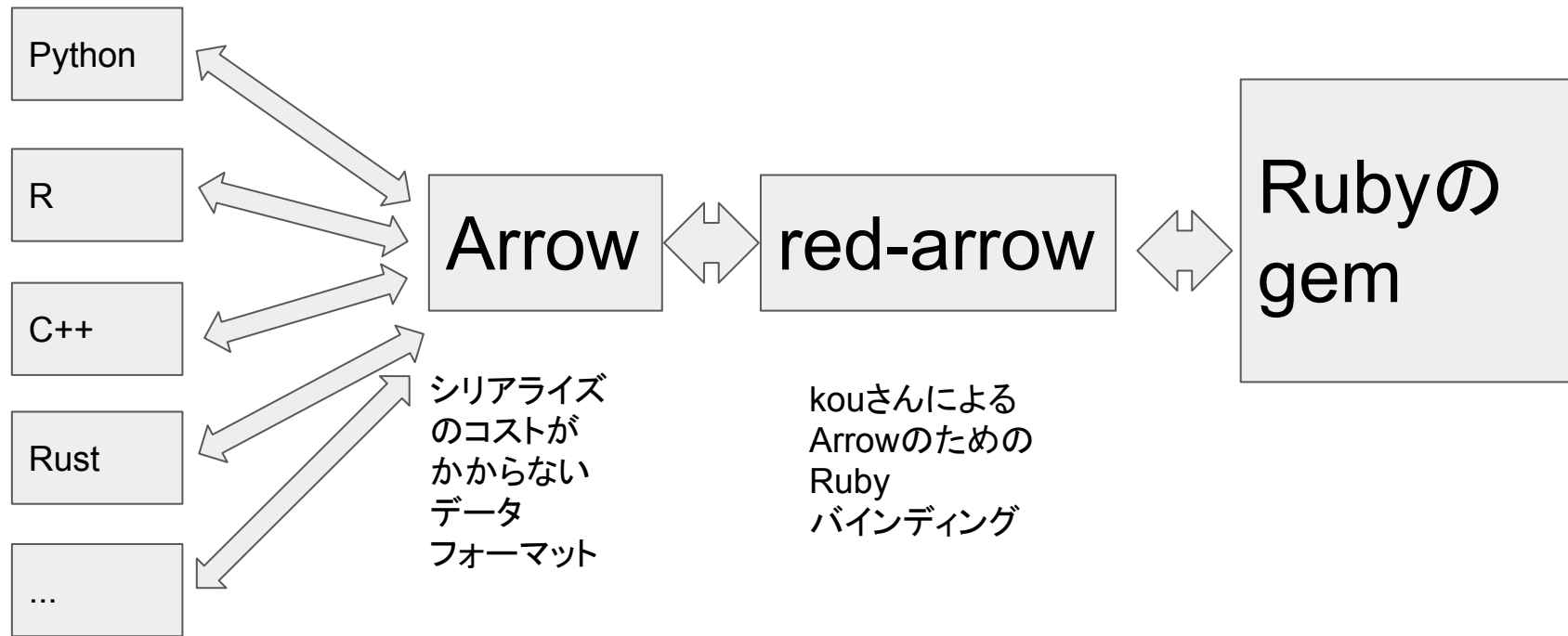
```
docker pull rubynum/gnuplot-demo
docker run -d -p 8888:8888 rubynum/gnuplot-demo start-notebook.sh --NotebookApp.token=''
```

上記のDocker コマンドを実行後 localhost:8888 をブラウザで開く

PyCall について



Arrow を介したRuby外言語との連携



概要

- データサイエンスは技術的にどうおもしろいか
 - Ruby business ?
- Rubyで現状何ができるか
 - Ruby business ?
- Rubyコミュニティに今後何を期待するか
 - Ruby community ?

コミュニティの重要性

- 「ユーザのフィードバック <-> ソフトウェア更新」のサイクルの促進 が重要
- すべてを(主)開発者だけに求めるのはきびしい
 - さまざまな環境での動作確認
 - (データを伴った実例指向の)ドキュメント (== Jupyter notebook)
 - 継続的インテグレーション
 - ポータビリティの高いgemパッケージング
- 逆に上記があると非常に助かる
 - gem 自体の開発だけが全てではない

Ruby のデータサイエンスにおける現状 (再掲)

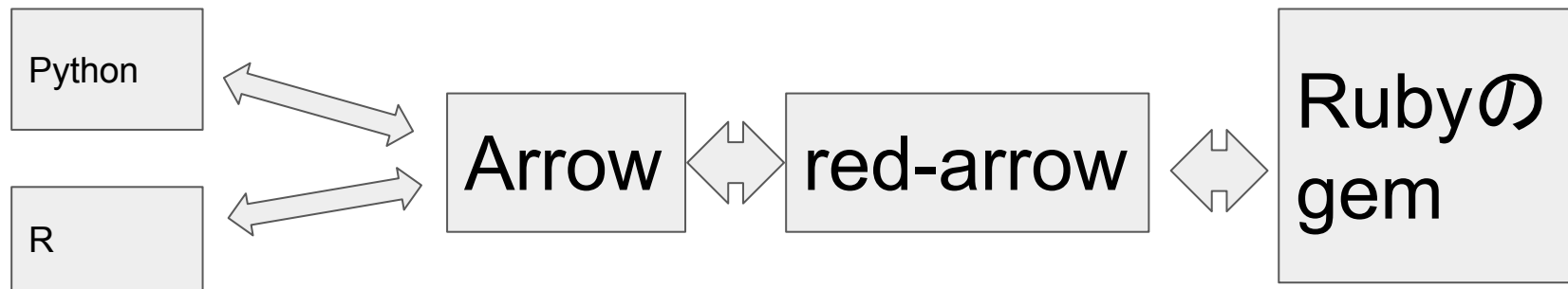
- 基本的な機能を備えたRubyのgem群
 - Daru (前処理)
 - [ruby-numo project のgem群] (分析)
 - Jupyter notebook + Ruby + plotly (可視化)
- Ruby外の言語との連携を推進するgem群
 - PyCall (PyData ソフトウェア群 との連携)
 - Red Data Tools (Apache Arrow を活用)

コミュニティに求める貢献の一案

- 基本的な機能を備えたRubyのgem群
 - 開発に加わる
 - 実例豊富なドキュメントを充実させる
 - 不足機能を開発者に求める
 - ビルドプロセスやパッケージングの改善
- Ruby外の言語との連携を推進するgem群
 - Binding gemの追加
 - そのgemにしか独自の機能や優位性を持つgemの発見, 創出

(red-)arrowの出現に伴う変化？

- 複数言語の連携が容易になる
=> 他言語には無い優位性を持つgemや独自機能を持つgemが求められる？
- そのようなgemがあるか，作れるか



コミュニティ

Slack, Gitter

- <https://sciruby-slack.herokuapp.com/>
- <https://gitter.im/red-data-tools/>

GitHub

- <https://github.com/sciruby-jp>
- <https://github.com/red-data-tools>

謝辞

- mrknさん (PyCallなど)
- 笹田耕一さん
- ITOC (しまねソフト研究開発センター) のみなさん
- 田中昌宏さん (ruby-numo プロジェクト)
- kouさん (cztop, iruby, red-data-tools)
- Rubyアソシエーション
- Sameer Deshmukhさん (daru, iruby)