Groongaの紹介
事例紹介

Naoya (@naoa_y)

全文検索エンジンGroonga勉強会@神戸
2014/06/27

今日の流れ

- ✓ **みんなで自己紹介**
- ✓ Groongaについて紹介
- ✓ 事例紹介
- ✓ Groongaの今後への期待

自己紹介

- Naoya (@naoa_y)
  - 大学は情報系
  - 新卒で3年半ほど金融系のユーザSIでインフラSE
  - 現在は3年半ほどITと無縁の仕事
  - Groonga/Mroonga暦は2年ちょっと

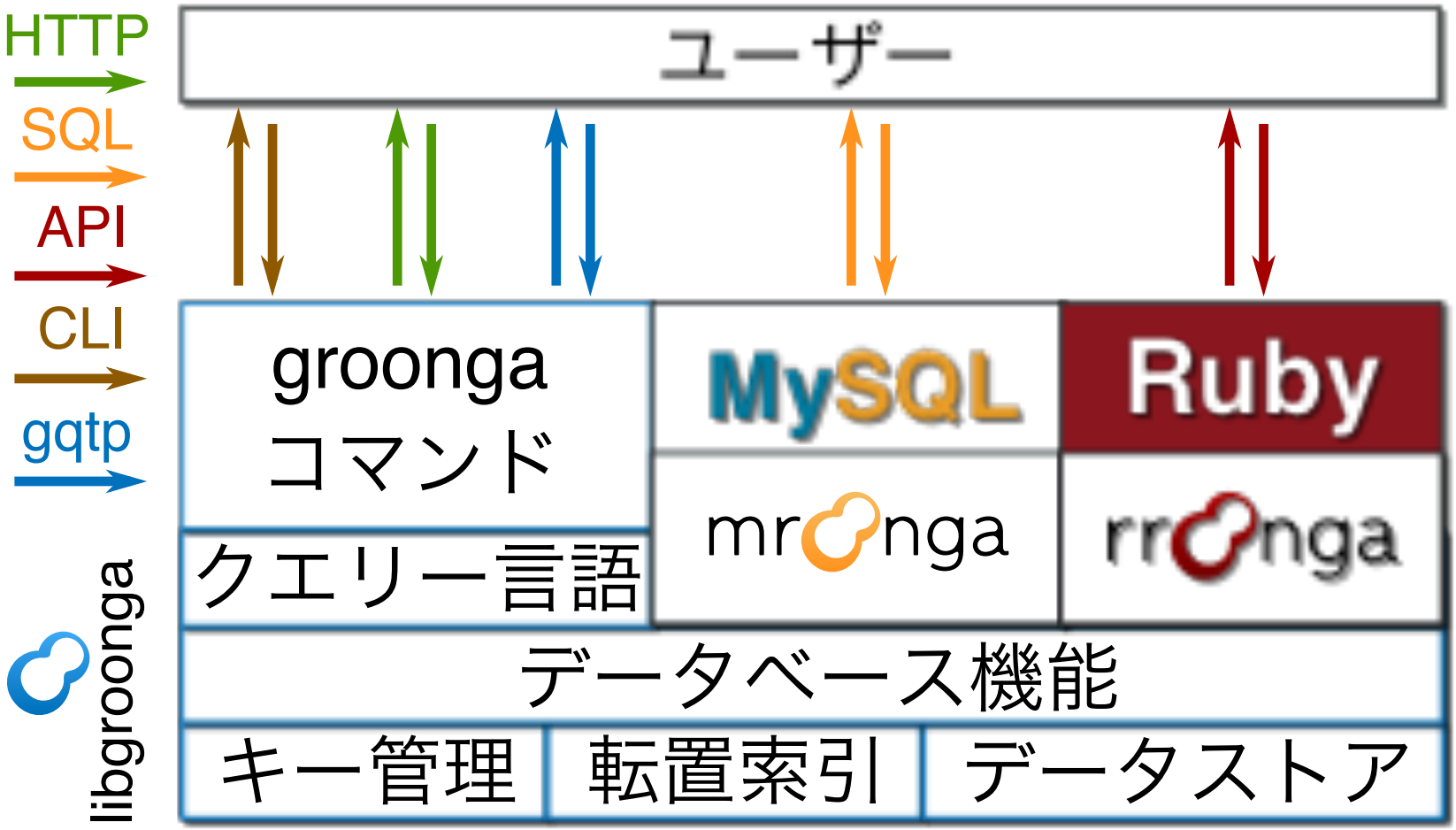
今日の流れ

- ~~みんなで自己紹介~~
- **Groongaについて紹介**
- 事例紹介
- Groongaの今後への期待

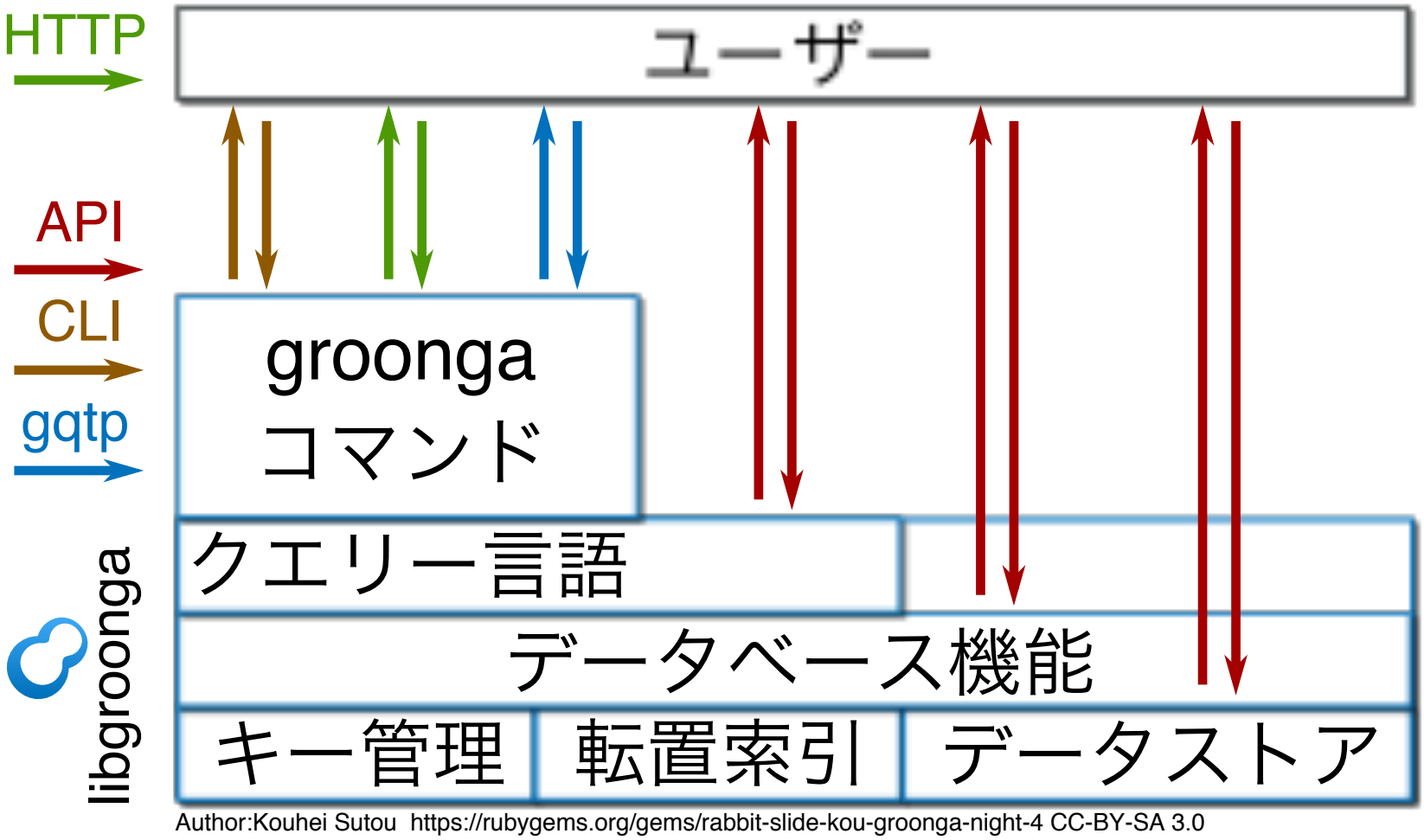
今日の流れ

- できるだけ具体的なコマンド等を見せていきます
- せっかく小規模なので途中で遮ってもいいのでどんどん質問してください。

Groonga族の概要



Groonga



Author:Kouhei Sutou https://rubygems.org/gems/rabbit-slide-kou-groonga-night-4 CC-BY-SA 3.0

Groongaって？

- C言語で書かれた**超高速**な全文検索ライブラリ/サーバ
- カラム指向のデータストア
  - 高速な集計処理
- 即時更新
  - 新鮮な情報をすぐに検索可能に

Groongaって？

- 専用のRDBほど複雑な表現は難しいがほぼRDBに近いイメージでテーブル設計ができる
- 転置索引を使った高速全文検索
  - 文字列をトークンに分割し、トークンが一致する文書IDを検索することにより大幅に演算量を減らす

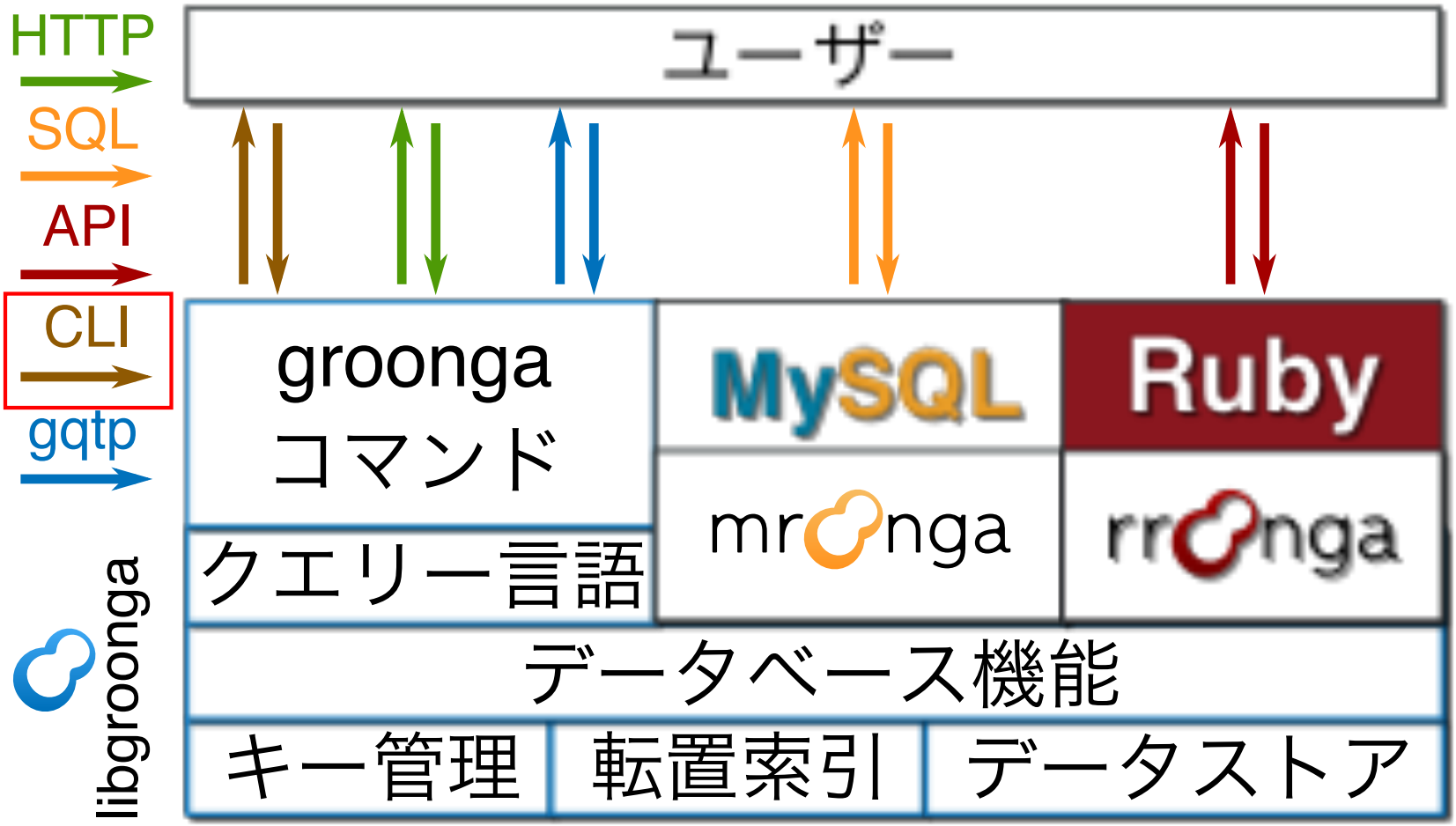
Groongaの全文検索の流れ

- 入力文を正規化
  - 例：This is a pen. → this is a pen.
- 入力文をトークンに分割
  - 例：今日は雨 → 今日/は/雨
- 転置索引を更新/検索
  - 分割したトークンとトークンの文書IDと文書IDにおける出現位置情報を記憶/検索

Groongaの全文検索の流れ

- Groongaではライブラリとして利用してもこれらの作業を勝手にやってくれる
- たぶんApache Luceneはひとつひとつやらないといけない(たぶん)

Groonga CLI

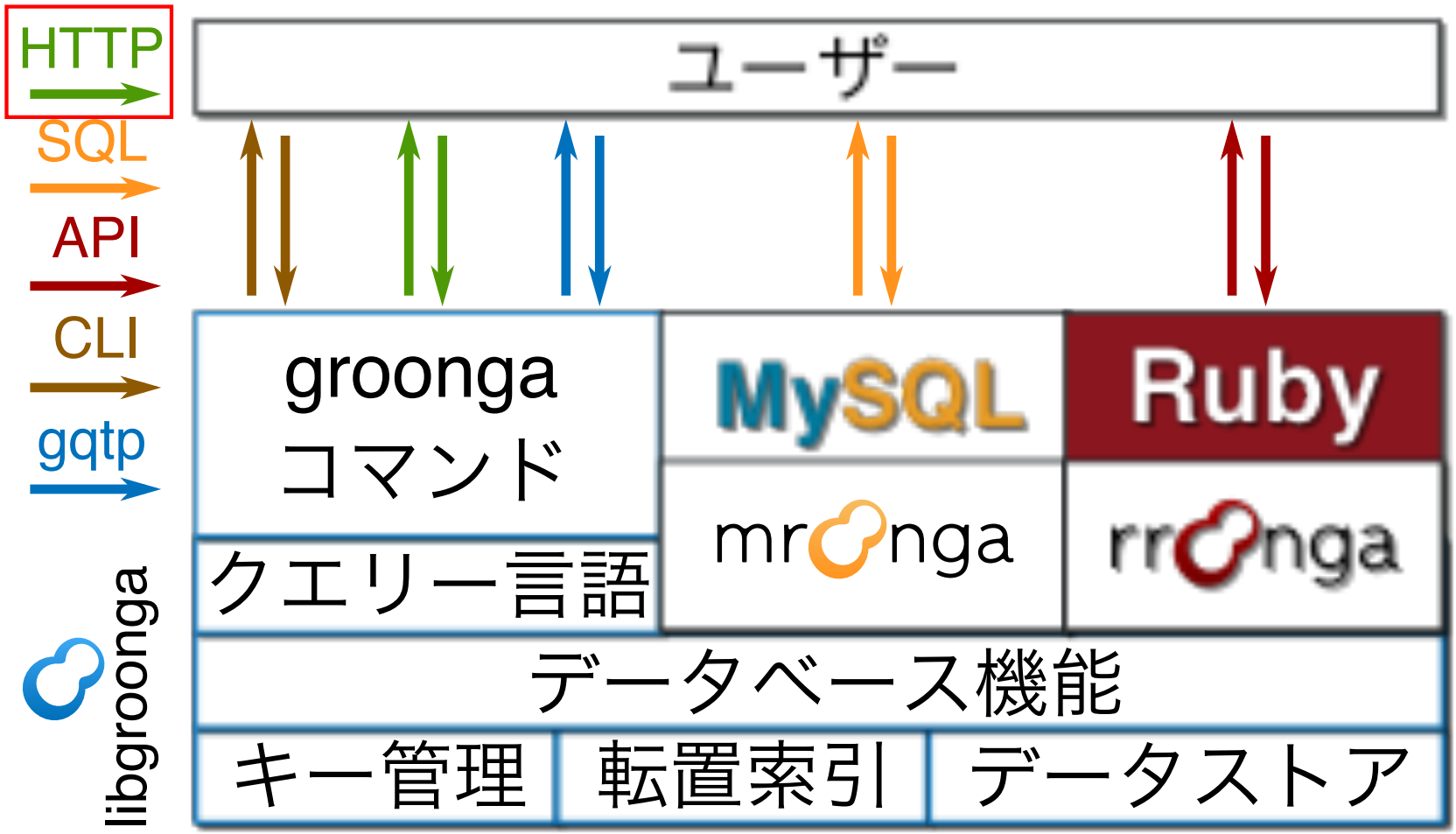


Author:Kouhei Sutou https://rubygems.org/gems/rabbit-slide-kou-groonga-night-4 CC-BY-SA 3.0

Groonga CLI

- CLI コマンドラインインターフェース
  - 対話的にちょっとしたデータを確認したいときやクエリの組み立てに便利
  - 最近のNoSQL系はほとんどJSONインターフェース
  - 息を吐くようにスクリプトでJSONかける人なら無くてもいいだろうけど、そうではないのでCLIがあるのは凄い便利

Groonga HTTP

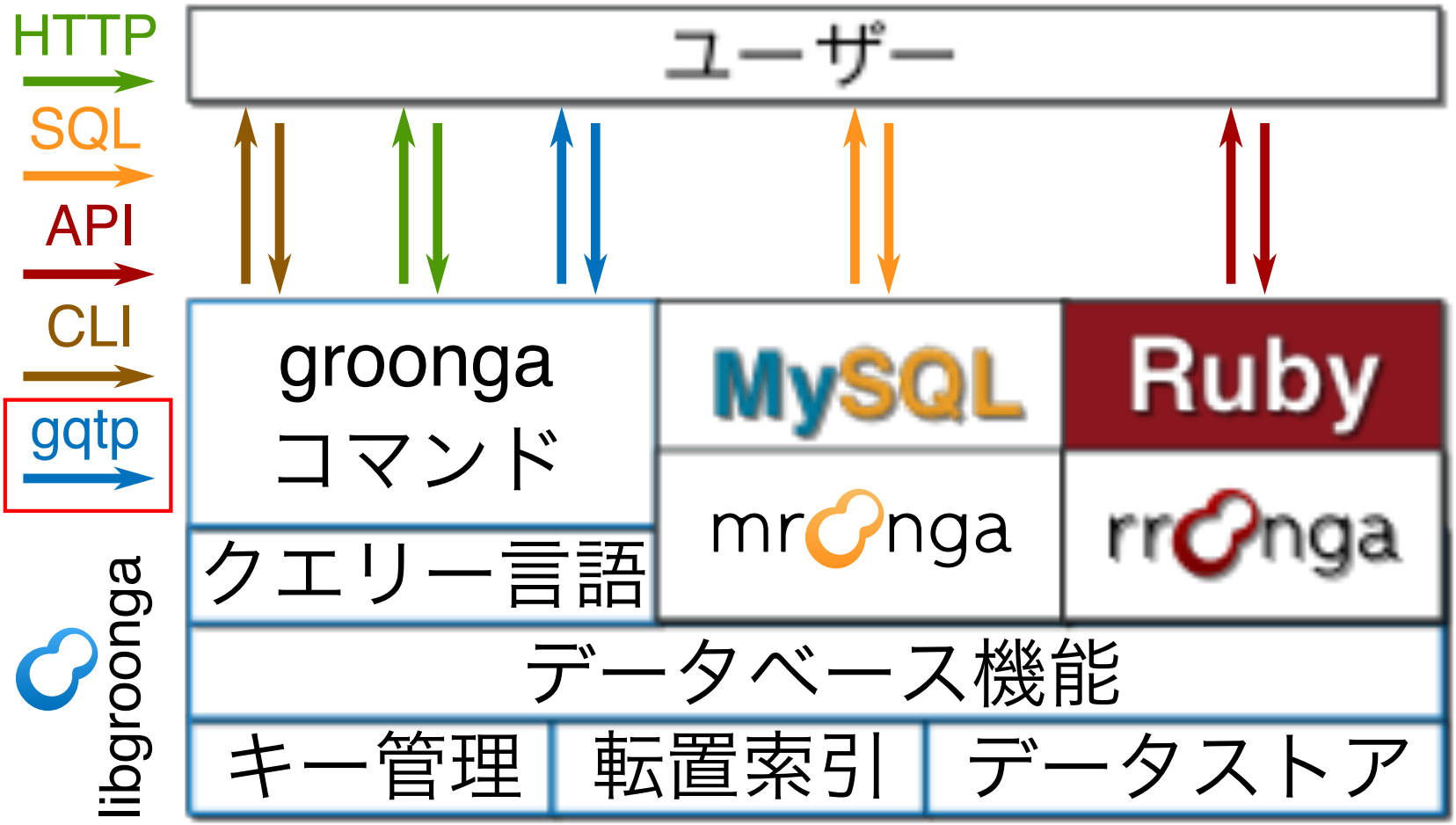


Author:Kouhei Sutou https://rubygems.org/gems/rabbit-slide-kou-groonga-night-4 CC-BY-SA 3.0

Groonga HTTP

- GroongaのHTTPはnginx組み込みと独自実装の2つ
- nginxの豊富な機能を使う場合はgroonga-httpd
- 単純に使うだけならgroongaコマンドをhttpモードで起動
  - サービスとして導入したい場合はgroonga-server-httpをインストール

Groonga GQTP

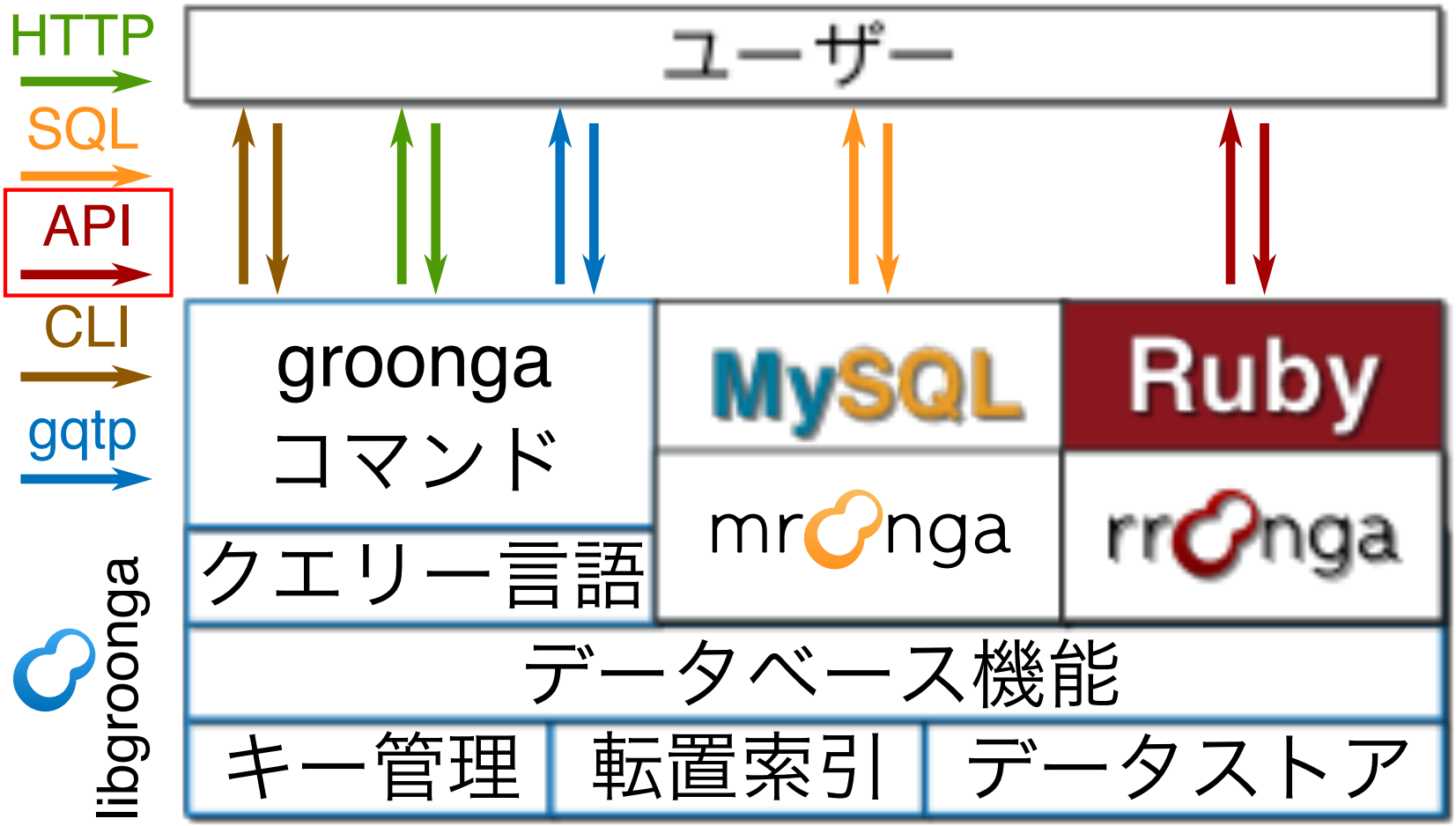


Author:Kouhei Sutou https://rubygems.org/gems/rabbit-slide-kou-groonga-night-4 CC-BY-SA 3.0

Groonga GQTP

- コマンドと同じようにしてリモートのサーバと直接会話できる
- groongaコマンドをgqtpモード起動
  - サービスとして導入したい場合はgroonga-server-gqtpをインストール
- 今後非推奨になる

Groongaライブラリ



Author:Kouhei Sutou https://rubygems.org/gems/rabbit-slide-kou-groonga-night-4 CC-BY-SA 3.0

Groongaライブラリ

- サーバを立てずにGroongaを組み込んでアプリを作る場合はこれ
- 結構難しい。が、コマンドを直接投げたり受けたりするモードがあるのでそれを使うだけなら簡単
- 難しいことしないならLuceneよりはるかに簡単？かも

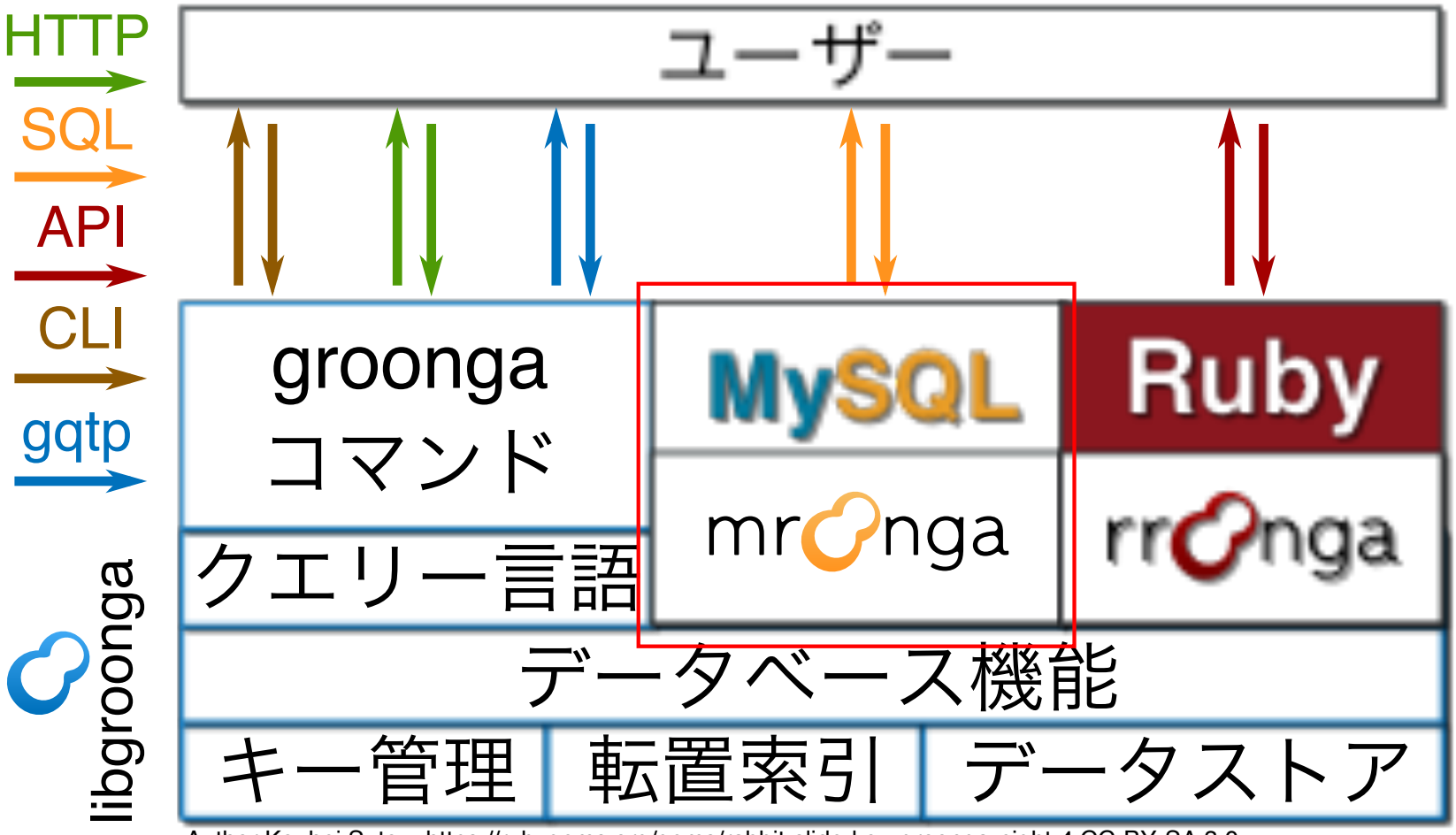
C言語でのクエリAPI

```
# https://github.com/naoa/groonga-api-sample/blob/master/ctx_sample.c
# complie: gcc -Wall -lgroonga -I/usr/include/groonga ctx_sample.c -o ctx_sample.o

#include <groonga.h>

grn_ctx ctx;
grn_obj *db;
const char *path = "test.grn";
grn_init();
grn_ctx_init(&ctx, 0);
db = grn_db_open(&ctx, path);
db = grn_db_create(&ctx, path, NULL);
grn_ctx_send(&ctx, "status", strlen("status"), GRN_CTX_QUIET);
grn_ctx_recv(&ctx, &result, &result_size, &recv_flags);
grn_ctx_fin(&ctx);
grn_fin();
```

Mroonga



Author:Kouhei Sutou https://rubygems.org/gems/rabbit-slide-kou-groonga-night-4 CC-BY-SA 3.0

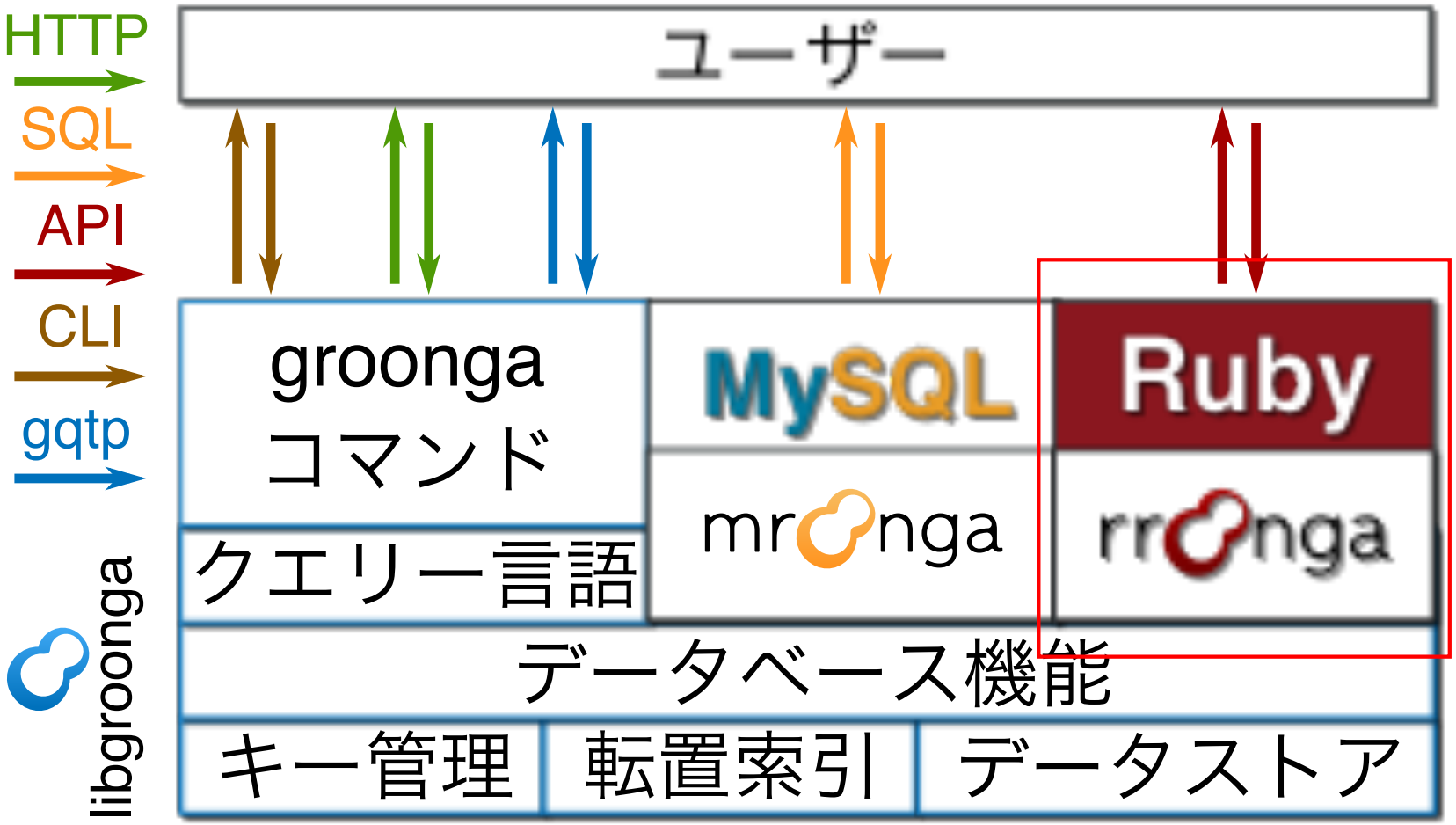
Mroongaって？

- Groongaをライブラリとして利用して全文検索機能が組み込まれたMySQLのストレージエンジン
- SQLで簡単に全文検索が可能
- MariaDBにもバンドル予定　？

Mroongaって？

- SQLでGroongaのコマンドよりも柔軟なデータ操作ができる
  - ※全文検索以外はMySQLの層でデータ操作するので同じ操作でもGroongaよりも遅くなることも
- 関数やツールなどMySQLの豊富な資産が利用できる
  - phpMyAdmin、レプリケーション、バックアップ etc

Rroonga



Rroongaって？

- Rubyらしい記法でGroongaライブラリを利用できる
- gemで一発インストール
  - ※Windows以外は数十分必要
- RubyとGroongaが動けば使える
  - クロスプラットフォーム(Win/Linux/Mac)

Rroongaって？

- ライブラリとして利用できるので他にサーバが不要(≒SQLite)
- Rubyなのでメモリ管理や文字列操作、データの操作が抜群に楽
- データベースAPIレベルが使えるので組み込みコマンド以上の複雑な操作をRubyで実現可能

どれを使うべき？

- サーバ経由でGroongaを使って高速な全文検索をしたい
  - Groonga HTTP/GQTP Mroonga
- 既存のMySQLのツールや複雑なSQLを扱いたい(速度差に注意)
  - Mroonga

どれを使うべき？

- Rubyで全文検索アプリケーションを作りたい
  - Rroonga
- 他の言語で全文検索アプリケーションを作りたい
  - Groonga GObject
    http://qiita.com/groonga/items/71b145b37d77bd160bf2

どれを使うべき？

- サーバ経由もしくはクライアントでGroongaの力を100%使いたい
  - Groonga
- 超大規模なデータベースで全文検索したい
  - Droonga、Mroonga & Spider

Groongaってどのくらい速い？

- 検索性能は文書やトークナイズの仕方にかなり左右される
- 日本語の文書でBigram(2文字ごと)なら数十GiB超でも余裕
- 日本語の文書でTrigram(3文字ごと)なら100GiB超でも余裕
- MeCabは単語間の検索性能に偏りが出やすい

Groongaってどのくらい速い？

- 英語のみの文書でBigramにすると10GiB、20GiBでかなり苦しい
- 英語の文書をNgramにするならTokenTrigramのSplitBy系が欲しいところ

Groongaの拡張性

- 以下をC言語のプラグインで拡張することができる
  - トークナイザ
  - ノーマライザ
  - コマンド
  - 関数
- ただし情報は少なくC言語なので単なる文字列操作も大変

Groongaの紹介

- ✓ 終わり
- ✓ 質問があれば受け付ける
- ✓ ここで休憩 & 雑談

今日の流れ

- ~~みんなで自己紹介~~
- ~~Groongaについて紹介~~
- **事例紹介**
- Groongaの今後への期待

事例紹介

- 特許の全文検索サービスを個人で作りました。
  - http://patentfield.com
  - まだデザイン面とか使い方とか力をいれてないのでその辺は気にしない

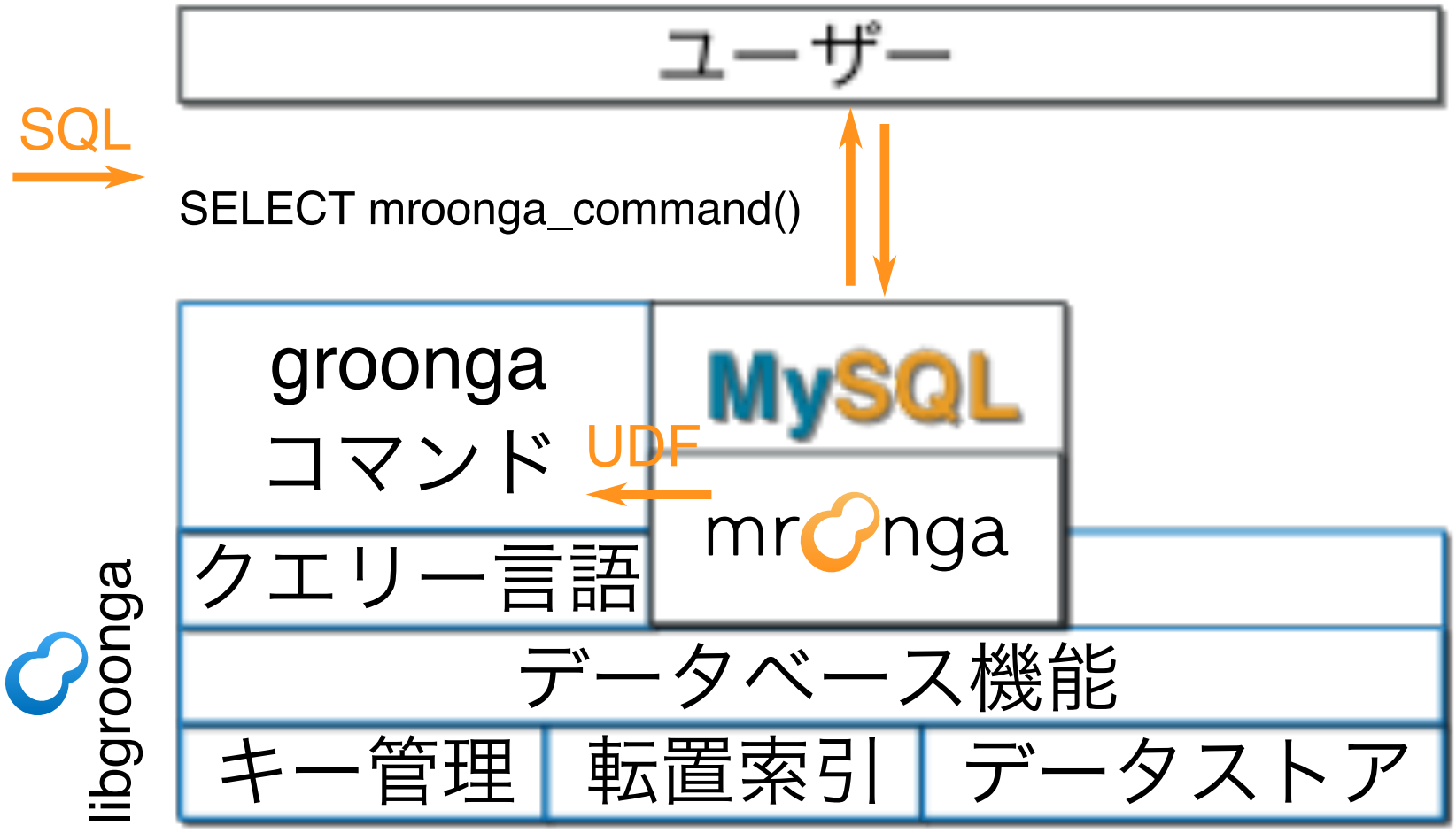
現状の公的サービス

- 700万件以上蓄積されているのに1000件以下に絞り込まないと表示できない
- ソートできない、検索結果から絞込みできない
- 検索結果で番号とタイトルと出願人以外表示できない
- DBがいくつも分かれていて検索対象にできない項目が多数

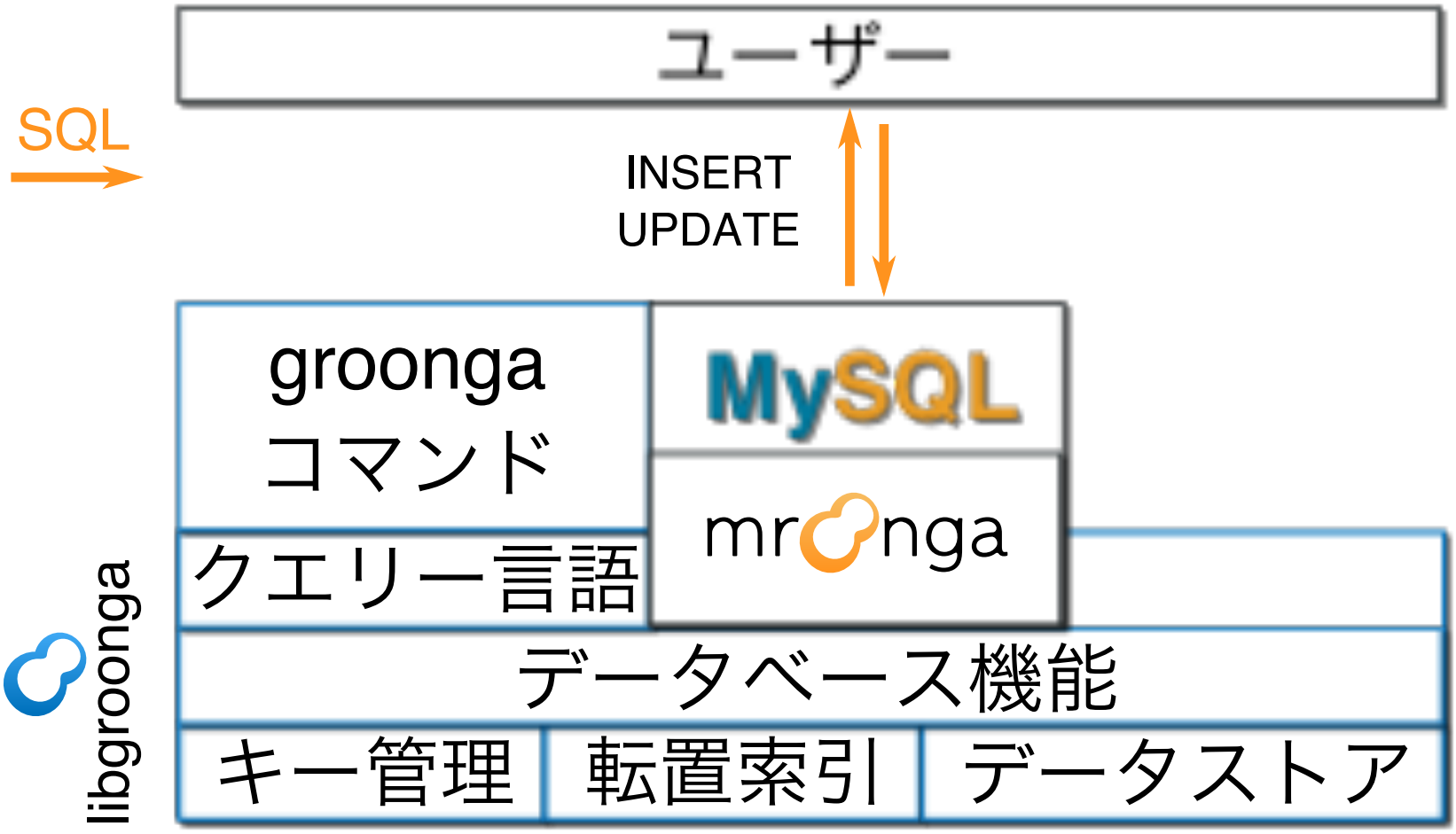
システム概要

- 最大で1000万件、データサイズ400GiB超 トータルで1.5TBぐらい
- カラム数は100ぐらい
- Mroongaを利用
- 複雑な絞込みやドリルダウンを使いたいため全文検索はGroongaを利用

全文検索時


ユーザー
SQL
SELECT mroonga_command()
groonga
コマンド
UDF
MySQL
mroonga
クエリー言語
データベース機能
キー管理
転置索引
データストア
libgroonga

更新・メンテナンス時


ユーザー
SQL
INSERT
UPDATE
groonga
コマンド
MySQL
mroonga
クエリー言語
データベース機能
キー管理
転置索引
データストア
libgroonga

Groongaで使った機能

- ✓ 全文検索
- ✓ 重み付け
- ✓ ドリルダウン
- ✓ サジェスト
- ✓ 近傍検索
- ✓ 類似検索
- ✓ スニペット

トークナイザによる検索性能差

- 400GiBのデータベースを標準のTokenBigramすると検索速度がかなり劣化
- TokenMecabを採用 しかし、Mecabでは助詞等の頻出用語を含むと性能が顕著に劣化
- トークナイザプラグインで頻出用語を除去することで回避

英語の複数形・過去形の取扱

- 英語でTokenBigramを使うと単語ごとにトークナイズされる
- これでは複数形や過去形を検索できない
- トークナイザプラグインでステミングを追加することで回避
  - 一定のルールで複数形や過去形っぽい末尾を削除してくれる

ノーマライザのカスタマイズ

- カタカナの小文字、大文字の扱いがばらばら
  - 例：フィルム⇔フイルム
- 古いデータでは長音記号がハイフンだったり
  - 例：データベース⇔データベース
- ノーマライザプラグインでこれらを同一視

大量レコードのドリルダウンによる性能劣化

- 全文検索結果で大量にレコードがあるとドリルダウンはコストが高い
  - 100万件以上はドリルダウンしないように組み込みselectコマンドをいじる
  - C言語やRroongaでDB-APIを使うことでも解消可能

複数カラム絞込みスニペット

- Groonga組み込みのsnipet_html関数はカラムを指定するだけで便利
- しかし、対象のカラム以外の検索クエリも対象としてしまう
- ノーマライザの指定ができない
- Mroongaと同等のスニペット関数プラグインで作る

感想

- 400GiB超のデータベースが1台でさばけるとは思わなかった
- このサイズを1台で高速に検索できるというのはすごく夢が広がる

感想

- 個人や少人数でも少し昔では考えられなかった規模の全文検索システムを作ることができる
- メーリングリストが親切なのでプログラミング経験がほとんどなくても作れた
- 今後はRubyで全文検索デスクトップアプリを作りたい！

事例紹介

- ✓ 終わり
- ✓ 質問があれば受け付ける
- ✓ ここで休憩 & 雑談

お疲れ様でした