DandD とデータ解析ソフトウェア R

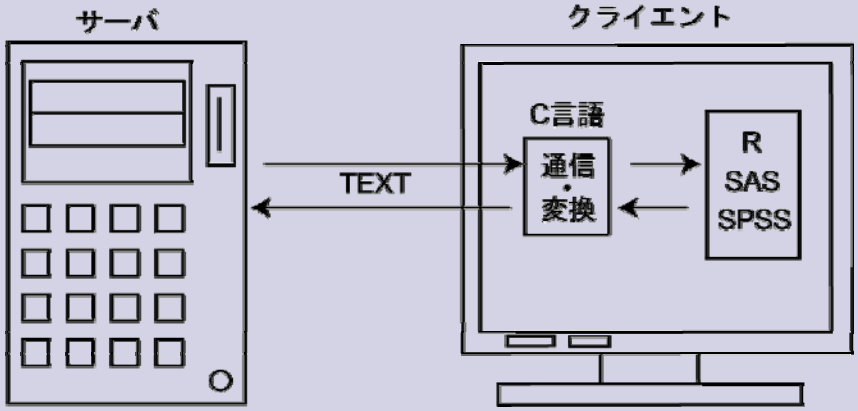
慶應大・理工　熊坂 夏彦
慶應大・理工　横内 大介

データ解析ソフトウェア R

- S/Splus のクローン
- フリーソフトウェア
- 多様な統計手法とグラフィクス環境
- オープンソースにより高度に拡張可能
- DandD クライエント・システムの一例
  →インタフェイスの構築

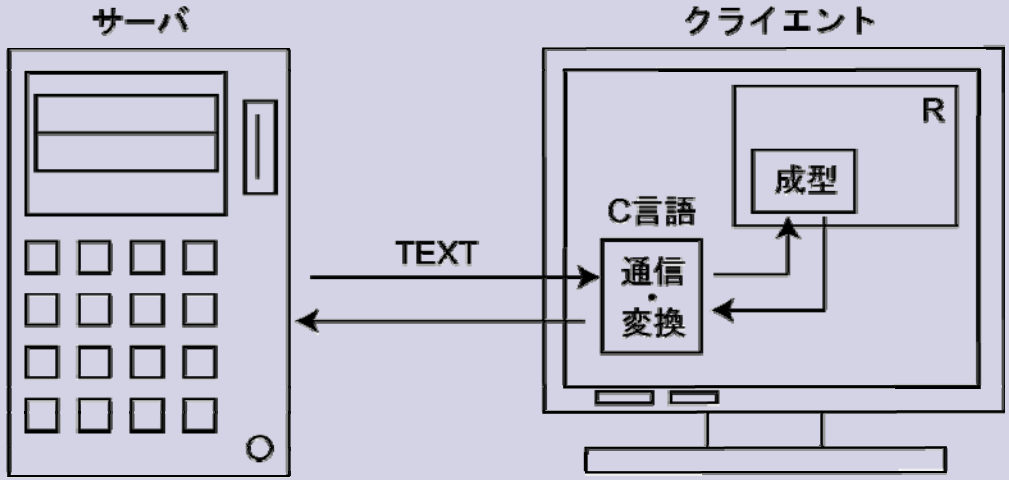
インタフェイス(1)

- サーバとの通信
  - ソケット
    →Berkley Socket API
    →Winsock32 API
  - 文字コード変換
    →GNUの文字コード変換ライブラリ(libiconv)
  - C言語による実装



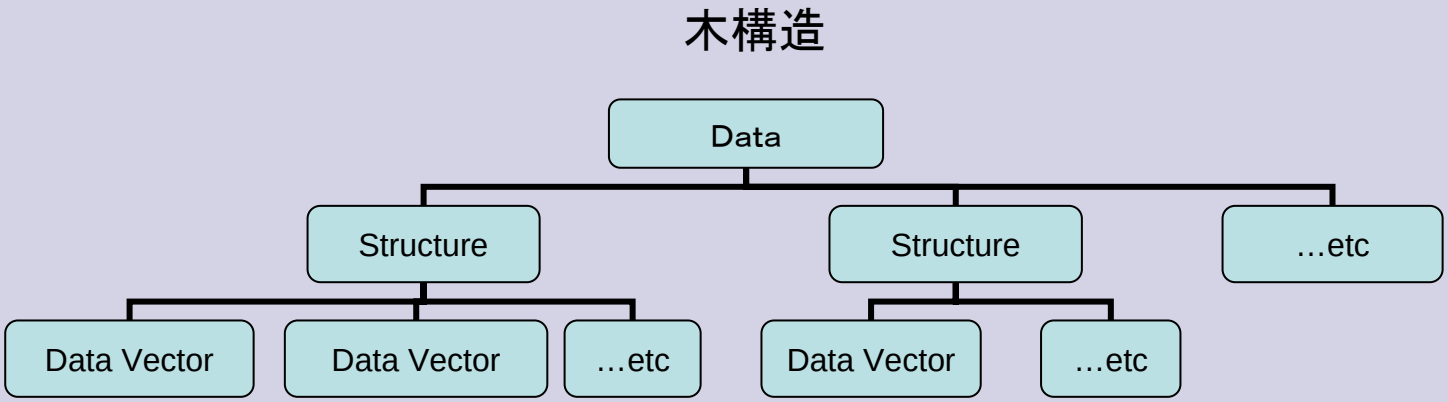
インタフェイス(2)

- Rのオブジェクトに変換
  →Rの内部で成型（Rに依存）
  →他のプログラムからも利用できるよう，将来的にはC言語に移行



クライエントプログラムの条件

- インスタンスを自然な形で取り込むために
  - XMLの木構造を表現できる
    →リスト形式
  - 各ノード（木の節）に属性を付加することができる
- 多国語対応している



ブラウザと解析ソフトウェアの関係

- ブラウザ
  - 豊富な背景情報の記述
  - 画像やTeXによる数式の表示
  - データの外観を知る
- 解析ソフトウェア
  - データを自由に扱える
  - データの視覚化が容易にできる
  - データそのものの性質を知る

デモンストレーション

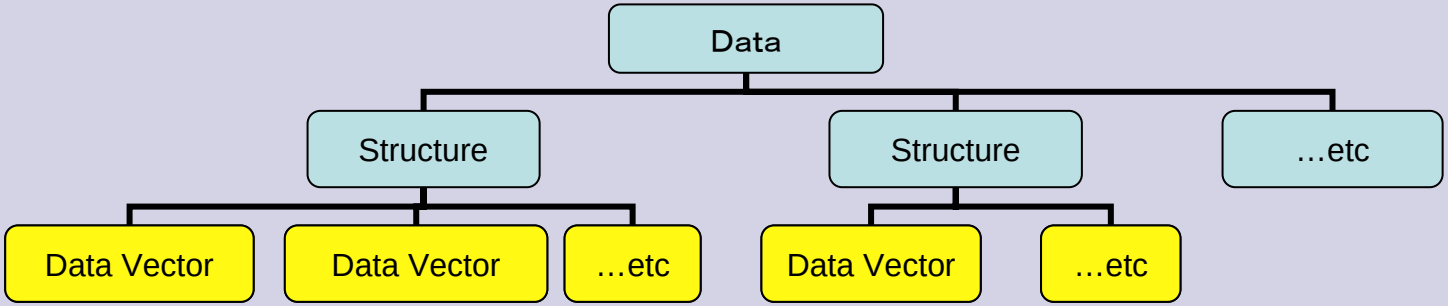
変数名

- 属性にLongName と Id
- 変数名はLongNameを使用
  →Id名は意味のない文字列
- LongNameの特殊文字を除く
  → “+“, “=“, “(“, “)” など
- LongNameを縮約
  →文では変数名として適当でない
  →abbreviate関数

遅延評価（lazy evaluation）

- 最初に構造と属性だけを取得する
- データベクトル部は取得手続きのみを保持
- 大規模データに強い
  →気象データなど
- 予約オブジェクト
  →アクセスされた時点で値が評価される

木構造



基数系（Radix）

- 年月日はユリウス日に変換
  julian関数
- 時分秒は秒に統一
  Seconds = Hour*3600 + Minute*60 + Second
- 度分秒は度に統一
  Degree = Degree + Minute/60 + Second/3600

データの構造

- 配列形式は関係形式に自動変換
  →解析ソフトウェア上で扱いづらい
  →解析する段階で関係形式にすることが多い

まとめ

- インスタンスの様々な属性情報により，データを柔軟に扱うことができる
- サーバ・クライエントシステムにより，ネットワーク上の様々なデータが，必要な段階で取得できる