2003年度　統計関連学会　連合大会

企画セッション「データとその属性情報の記述」

DandDインスタンス

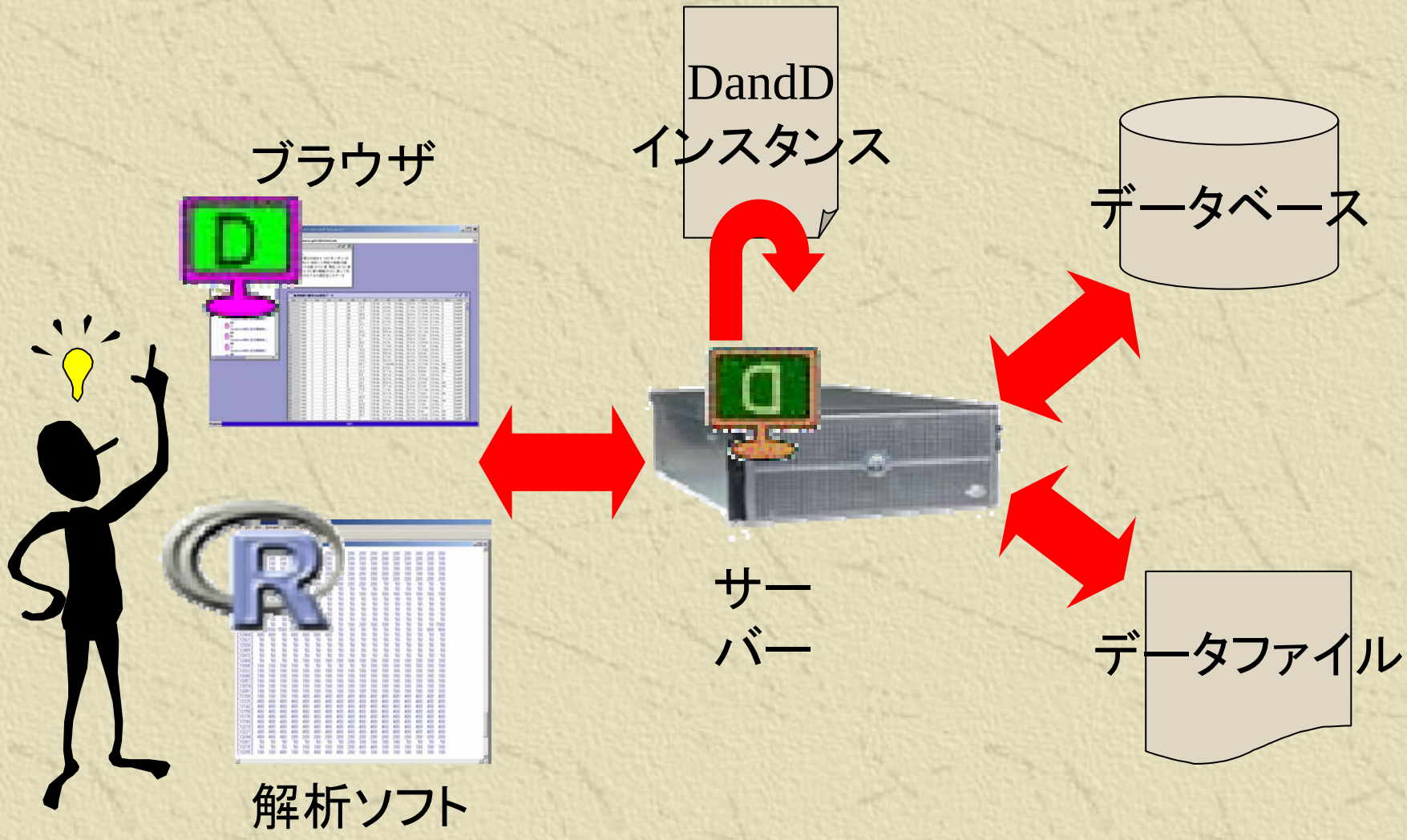
慶應大・理工　島津秀康

DandDインスタンスとは

- XML（eXtensible Markup Language）文書
  - タグを使った記述
- インスタンスは単に「記述」
  - データの所在，データベクトル，構造…

記述を解釈するのは支援システム

DandDの全体像



従来の記述

- Relational（関係形式）
  - ベクトル（変量）を並べたもの
- 正規形
  - 第1正規形
    - すべての要素がアトミック
  - 第2正規形
    - いくつかの項目でレコードを一意に定められる
  - 第3正規形
    - レコードを一意に定める項目以外の項目は互いに独立

ベクトル間の関係は不明

DandDでのデータ表現

- Relational（関係形式）

  データベクトルの並び

  - TimeSeries（時系列）
    - 時間，値を示すベクトルの並び
  - PointProcess（点過程）
    - 時刻，場所，値を示すベクトルの並び

- Array（配列形式）

  軸とデータベクトルの並び

- 基数系

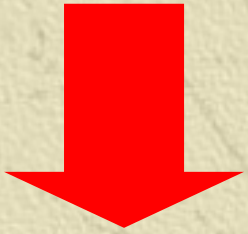
  時間 → 時，分，秒を示すベクトルの並び

データベクトル間の関係でデータを表現

従来の記述との違い

- ベクトル間の関係に基づく組織化
- 組織化に人間の意図が反映される

View



データの解析に直結した表現

解析におけるデータの流れ

- 取得
- クリーニング，記述
- ブラウジング
- 解析
- モデリング
- モデルの評価

試行錯誤を重ねるうち
データの見方（View）が
変化

変化をインスタンスに記述

水泳競技データ

- (財)日本水泳連盟医・科学委員会
- 1996年～2003年
- 13329レコード
- 約100項目
  - 大会名，種目
  - 選手名
  - ストロークタイム
  - ラップタイム　など

実際のデータベース

110項目

13329行

年度	大会名	レース日付	距離	種目名	性別	予選・決勝	組	コース	群	個人識別ID	スタートリストID
1999	日本選手権	1999/6/10	100	BA	M	P	1	2	4	0	1
1999	日本選手権	1999/6/10	100	BA	M	P	1	3	4	0	2
1999	日本選手権	1999/6/10	200	FR	M	P	2	4	4	0	9
1999	日本選手権	1999/6/10	200	FR	M	P	2	5	5	0	10
1999	日本選手権	1999/6/10	200	FR	M	P	2	6	5	0	11
1999	日本選手権	1999/6/10	200	FR	M	P	2	7	4	0	12
1999	日本選手権	1999/6/10	200	FR	M	P	2	8	5	0	13
1998	日本選手権	1998/6/11	400	IM	F	f	0	6	3	0	6
1998	日本選手権	1998/6/11	400	IM	F	f	0	7	3	0	7
1998	日本選手権	1998/6/11	400	IM	F	f	0	8	3	0	8
2002	日本選手権	2002/6/16	400	IM	F	f	0	8	3	450189	8
2002	日本選手権	2002/6/16	400	IM	F	f	0	7	3	135808	7
2003	日本選手権	2003/4/22	400	FR	F	P	5	6	3	808868	34
2003	日本選手権	2003/4/22	400	FR	F	P	5	7	3	1085642	35
2003	日本選手権	2003/4/22	400	FR	F	P	5	8	3	304051	36

ストローク長	ストローク長	ターン1	ターン2	ターン3	ターン4	ターン5	ターン6	ターン7	フィニッシュ	テーブルシ
NA	NA	00:08.1	NA	NA	NA	NA	NA	NA	00:03.0	12215
NA	NA	00:08.3	NA	NA	NA	NA	NA	NA	00:02.8	12216
NA	NA	00:06.5	00:06.7	00:07.1	NA	NA	NA	NA	00:02.8	12901
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	12902
NA	NA	00:06.7	00:06.7	00:07.0	NA	NA	NA	NA	00:03.0	12903
NA	NA	00:06.3	00:06.7	00:06.9	NA	NA	NA	NA	00:02.8	12904
NA	NA	00:06.6	00:06.8	00:06.8	NA	NA	NA	NA	00:02.7	12905
1.74654	1.7321	00:08.7	00:10.5	00:10.6	00:11.6	0:00:12	00:09.1	00:08.0	00:03.2	15770
2.26815	2.28784	00:08.6	00:10.5	00:10.2	00:11.5	00:11.9	00:08.9	00:07.7	00:03.1	15771
2.02404	1.99828	00:08.6	00:10.9	00:10.7	00:12.1	00:12.1	00:09.5	00:07.9	00:03.2	15772
1.591	1.593	00:08.2	00:10.6	00:10.5	00:12.1	00:11.8	00:08.8	00:07.6	00:03.9	32245
2.147	2.133	00:08.0	00:10.2	00:10.2	00:11.6	00:11.7	00:09.0	00:08.0	00:03.2	32244
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	32381
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	32382
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	32383

取得

- 会場に設置されたビデオカメラより取得
- 手入力

＜理由＞

- 1選手1競技の結果を1行で管理
- 様々な競技を1つのテーブルで管理
- 更新が容易

クリーニング

- 2次データの混在
  - ストローク頻度＝60/ストロークタイム

    ➡ 1次データのみを記述
- 個人識別

  ➡ 新たなIDベクトルを作成

記述

- 意味を捉えた分かりやすい形へ（テーブルの分割）
  - 競技情報
    - 競技ID，年，月，日，大会名，性別，種目，距離…
  - 個人情報
    - 個人ID，氏名
  - 時間情報
    - 競技ID，個人ID，観測ポイント，ラップタイム
  - ストローク情報
    - 競技ID，個人ID，観測ポイント，ストロークタイム

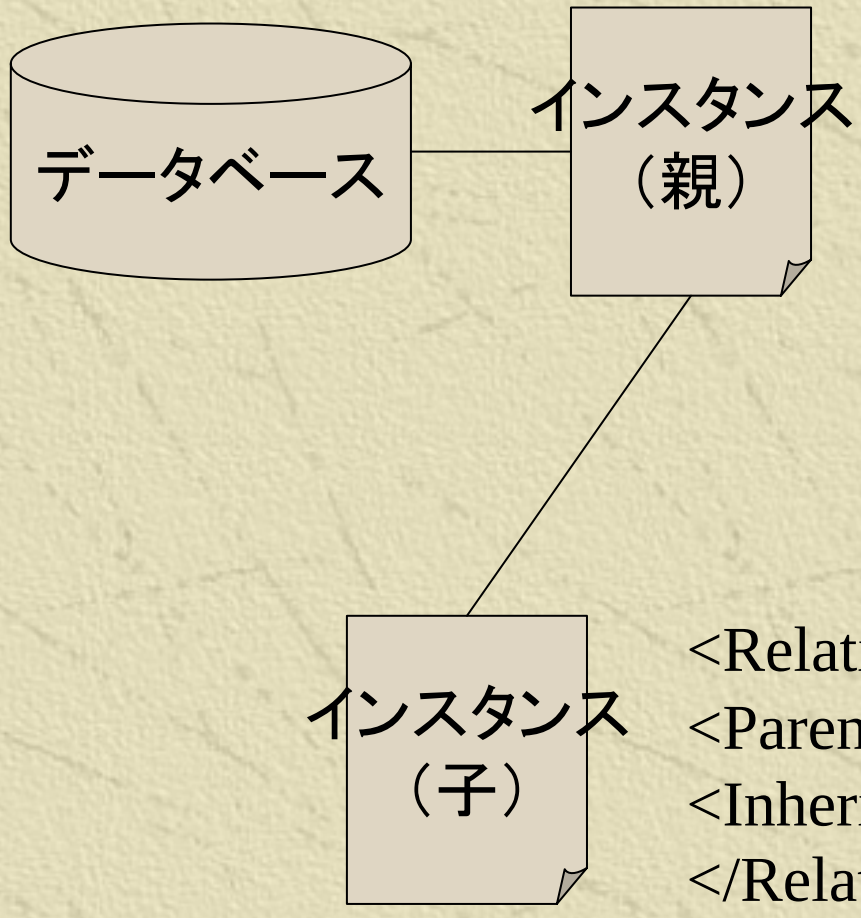
解析へ向けて

- 注目する変量
  - ストロークタイム
  - ラップタイム
- 注目する水準
  - 男子，200m，自由形

変化の記述（Relatives）

- 変化するのはデータの見方（**View**）
- データベクトルの継承の関係を記述

インスタンスのネットワーク

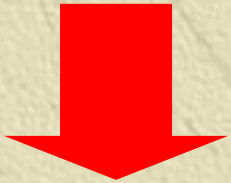



```
<Relatives>
<Child Id=〃〃 URL=〃〃>
<Inherit Id=〃〃>
  継承させるベクトルの作成手順
</Inherit>
</Relatives>
```

```
<Relatives>
<Parent Id=〃〃 URL=〃〃>
<Inherit Id=〃継承するベクトル〃/>
</Relatives>
```

まとめ

- ベクトル間の関係に注目する組織化により，多様なデータを表現
  - 元のデータを書き換えることなく，インスタンスの書き換だけでデータの見方（**View**）を自由に変えられる
- 特定のソフトに依存しない利用環境



高い汎用性