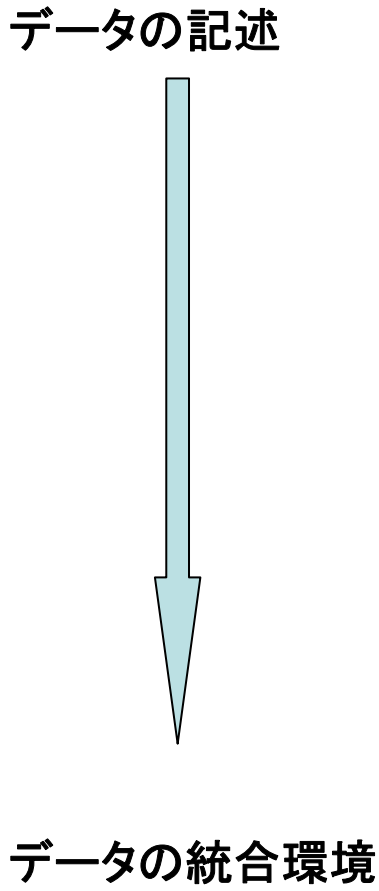
データ統合環境 DandD IV

慶應義塾大学理工学部
横内大介
慶應義塾大学理工学部
柴田里程

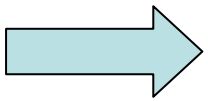
DandD

- D & D
  - データ解析の電子ジャーナル
- DandD
  - XML
  - D&D ルールの見直し
- DandD II
  - DandD クライエント・サーバシステム
  - インターデータベース
- DandD III
  - オブジェクト指向によるデータベクトル表現
  - データの変容
- DandD IV
  - 変量の概念
  - DataBody に対する属性の導入
  - Relational に対する View の抽象化



DandDルール

- 原則
  - データを単なる数値の並びであるデータベクトルの集まりとみなす
  - データベクトルは変量の実現値
- 変量間の関係
  - DandD III までの扱い
    - 構造として記述できるものはData 部に記述
    - 同一変量はデータベクトルの定義域（Domain）を共有していることで表現



DandD III ルールでは記述できない関係が存在

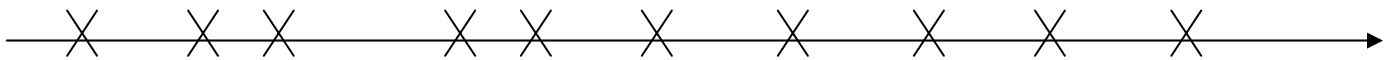
心臓R波・呼吸流量観測データ

- 実験内容
  - 1人の人間に対して行われた
  - 安静時，負荷時，負荷終了後
  - 体調が良好，不良
- 観測項目
  - R波
    - ピーク時点を測定
  - 呼吸流量
    - 0.5秒間隔で計測

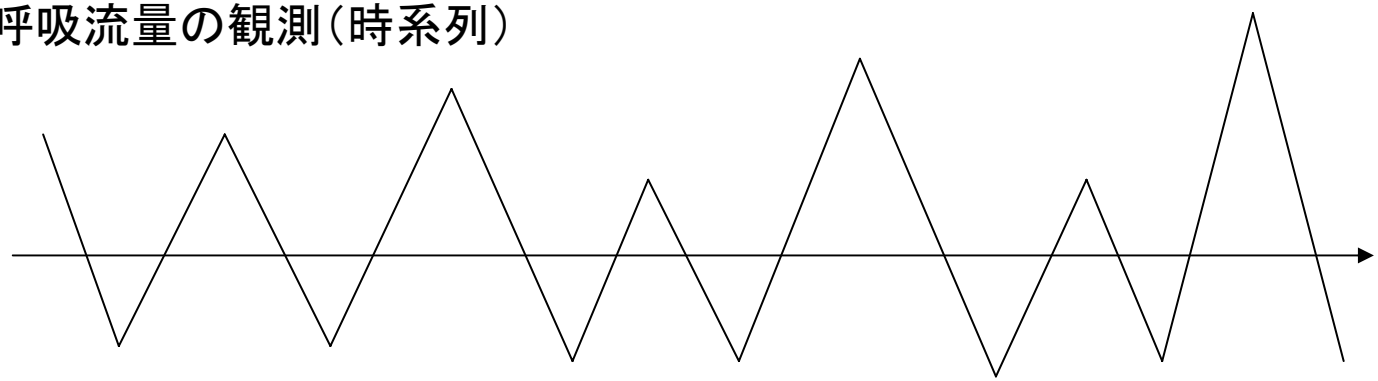
同時に計測

心臓R波・呼吸流量観測データ（２）

R波の観測（点過程）



呼吸流量の観測（時系列）



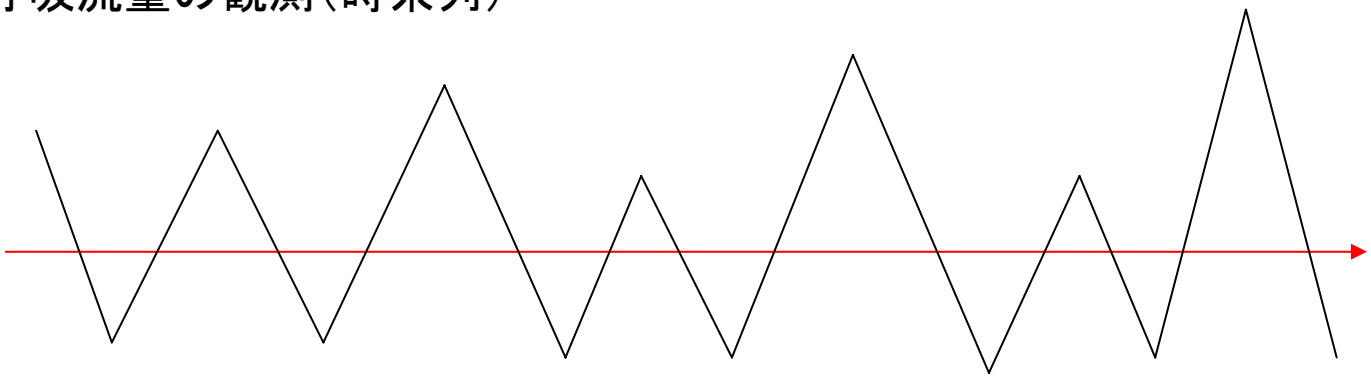
時間の経過

心臓R波・呼吸流量観測データ(2)

R波の観測(点過程)

呼吸流量の観測(時系列)

オブジェクトの共有



R波の生起時間と呼吸流量の観測時間は変量としては異なるが，同時に観測を行っているので，どちらも同じ時間というオブジェクトを共有していることになる.

DataBodyの属性

- DataBody
  - 複数のデータベクトルを記述しておくための要素
- DandD IV
  - CommonObject
    - オブジェクトの共有を示す属性
  - CommonVariable
    - 同一変量であることを示す属性
  - FunctionalDependency
    - 変量間の関数従属関係を記述

	ベクトルA	ベクトルB	ベクトルC
名前	商品A	商品B	商品C
太郎	1	3	2
次郎	2	1	3
花子	3	2	1
明子	2	3	1

$$A + B + C = 6$$

構造化によって現れる属性

- DandDによるデータベクトルの構造化
  - 関係形式
  - 配列形式
- 構造化によって現れる属性
  - 関係形式
    - キーに関する属性
    - 系を示す属性
    - etc

キーに関する属性

- PrimaryKey
  - 主キー
  - 1つの関係形式を構成するデータベクトルのうち，記録を一意に定めることができるデータベクトルを示す属性
  - 該当するデータベクトルのIDの並びを記述
- ForeighKey
  - 外部キー
  - 正規化の過程を経て複数の関係形式に分解されたデータを記述するための属性
  - 関係形式の間に存在する従属関係を示す鍵となるデータベクトルのIDの並びを示す

系

- Coordinate
  - 座標系
  - 直交座標系への変換方法を属性 Normalization に記述
- Radix
  - 基数系
  - 基数系を構成する最小の位にあわせるための変換方法を属性 Reduction に記述
- CauseEffect
  - 因果系
  - データが示す現象において，原因となるデータベクトルと結果を表すデータベクトルを示すための属性
  - 要因の種類については属性 FactorType に記述

データ統合環境の必要性

- インターネットの発展
  - さまざまなデータがネットワーク上で公開
  - データの詳細な背景情報の重要性
  - データの有機的な結合の必要性

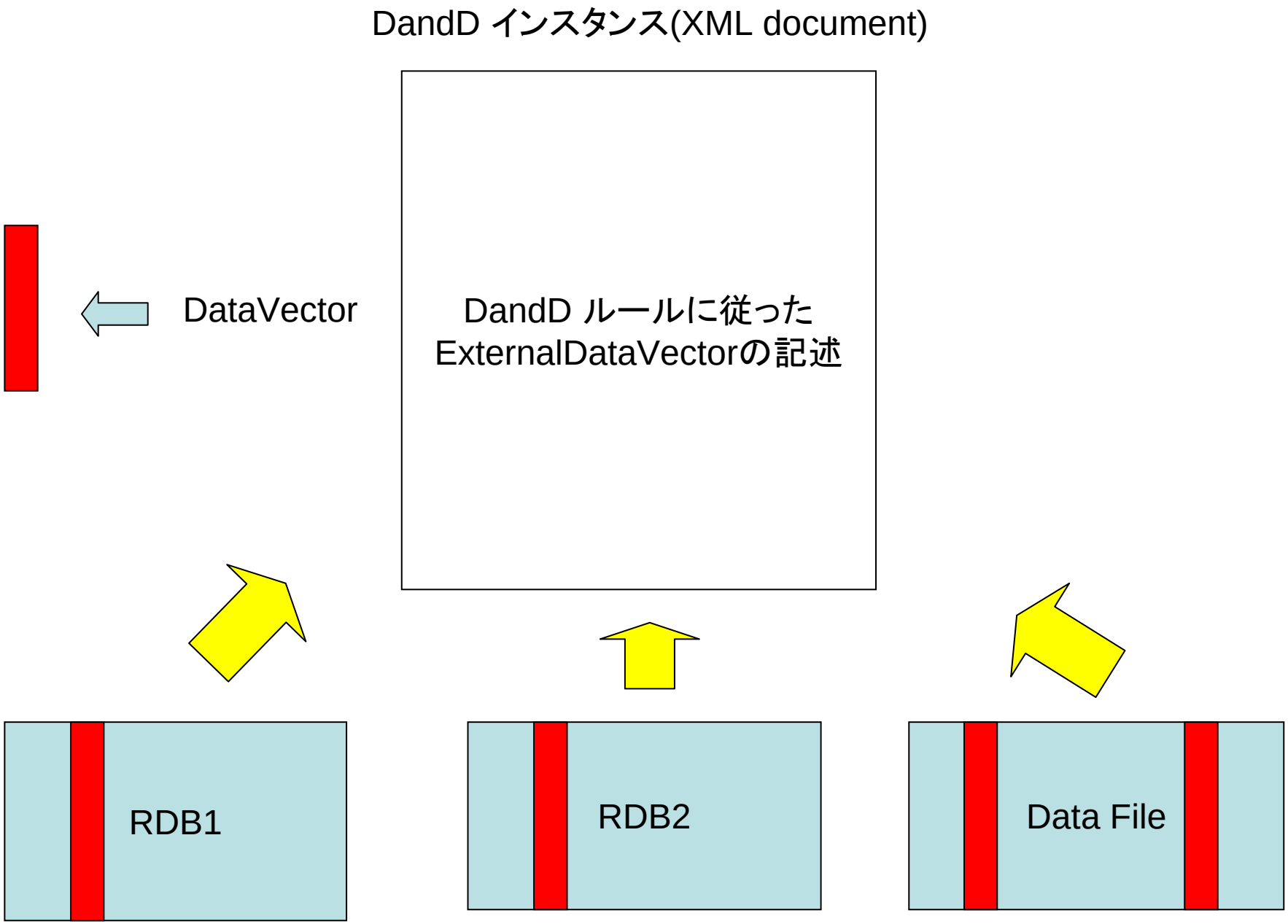
インターデータベース

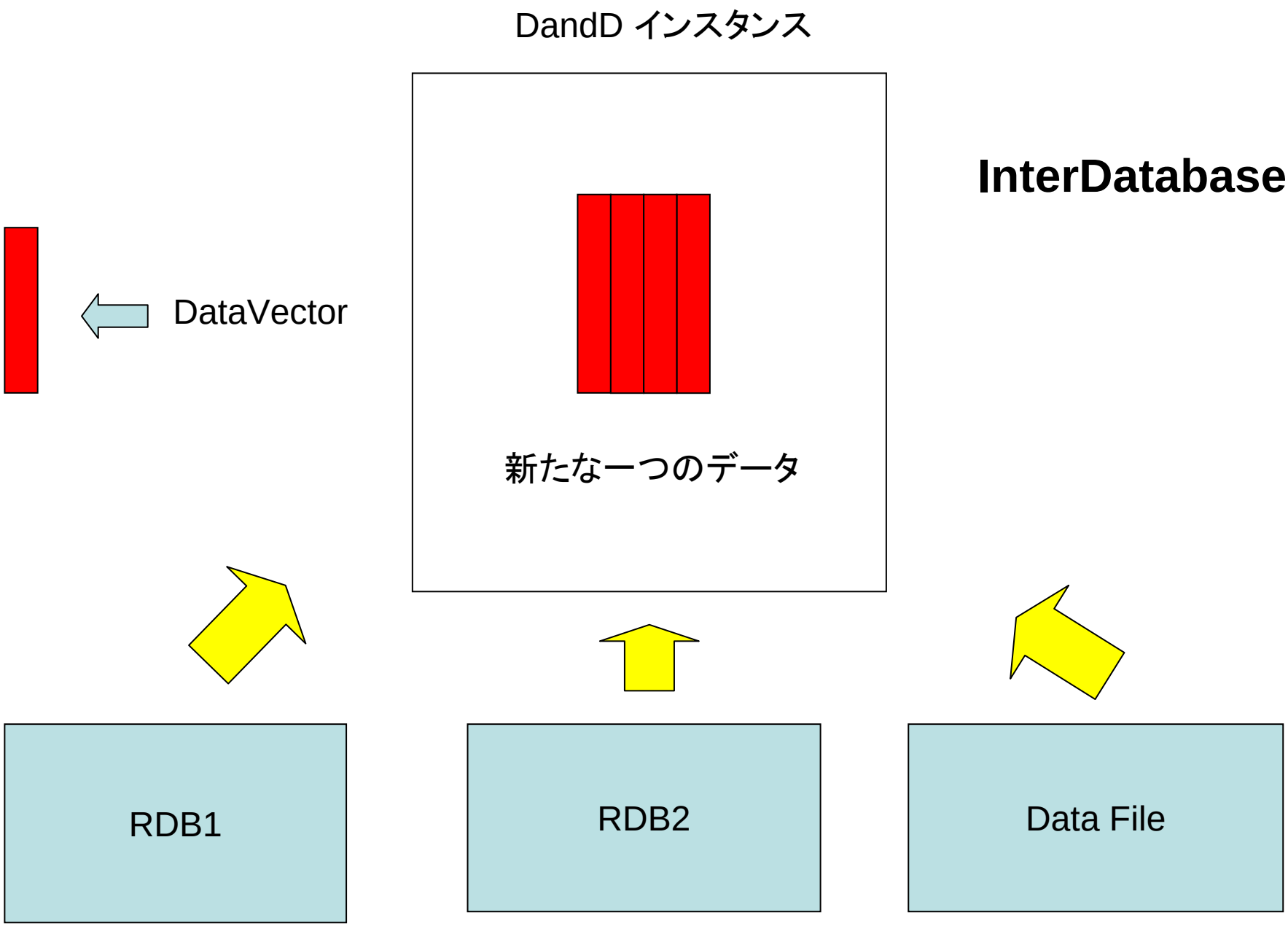
- **DandD ルールに基づいた十分な背景情報の記述**
- **既存のデータベースをそのまま活用できるフレームワーク**
- **インターネット上に散在する異種データの統合**

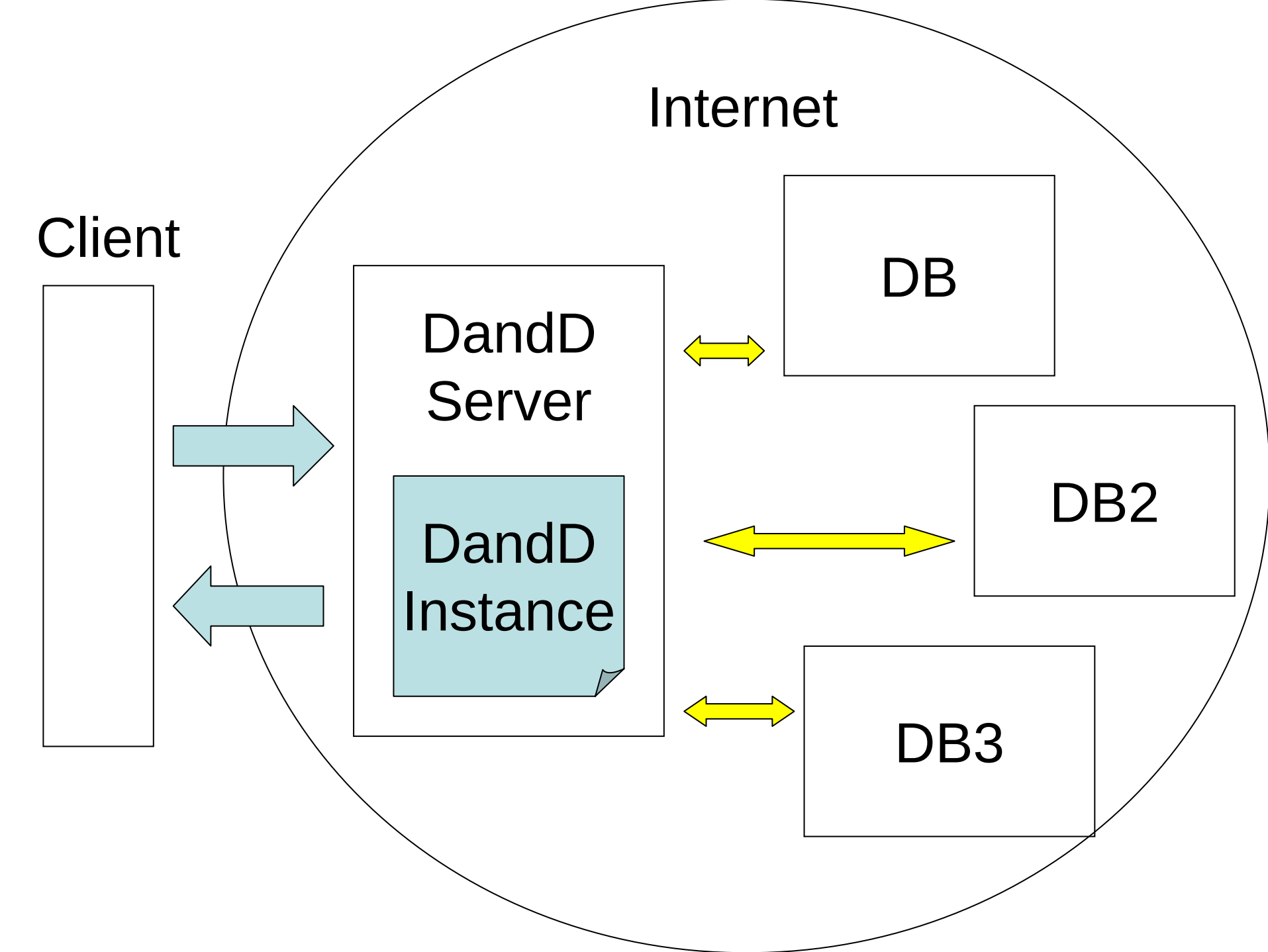
```
<DataBody>
  <DataVector Id="v1" Access="a1" Protocol="b1"
                      Query="c1" PostProcessing="d1" />
  .......
</DataBody>
```

External DataVector

```
<Appendix>
  <Access  Id="a1" IP="131.113.65.1" UserId="anonymous" />
  <Protocol Id="b1" Physical="tcp">
    <JDBC DatabaseServer="131.113.65.1" DatabaseName="KobeQuake"/>
  </Protocol>
  <Query ="c1" Type="SQL">select date from kobequake</Query>
  <ScanFormat Id="d1"> %*s,%s,%*s </ScanFormat>
  .......
</Appendix>
```


DandD インスタンス(XML document)
DataVector
DandD ルールに従った
ExternalDataVectorの記述
RDB1
RDB2
Data File

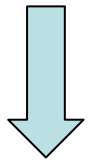

DandD インスタンス
InterDatabase
DataVector
新たな一つのデータ
RDB1
RDB2
Data File


Internet
Client
DandD
Server
DandD
Instance
DB
DB2
DB3

インターネット時代のデータを取り巻く環境

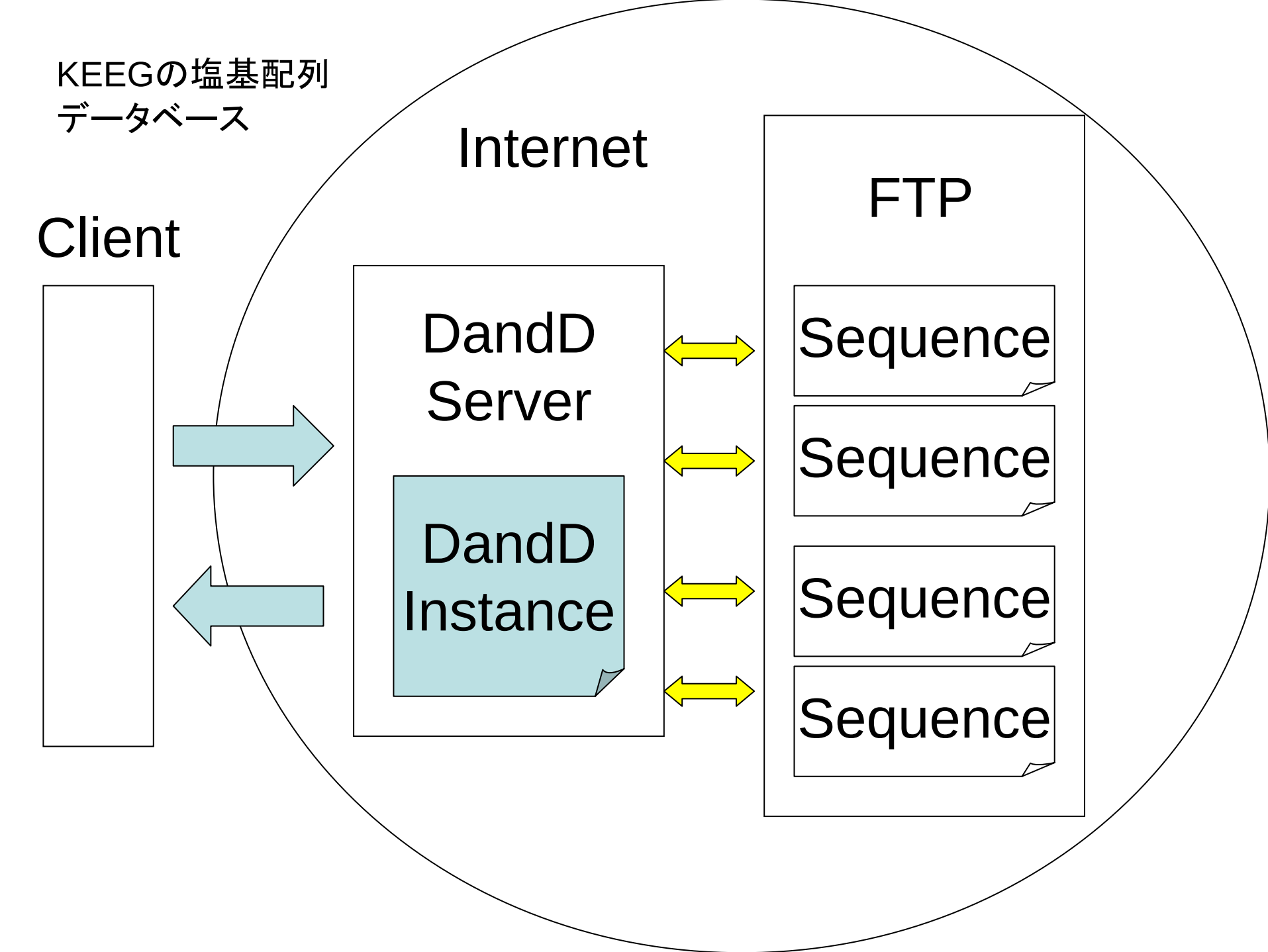
- バイオインフォマティックスデータ
  - インターネット上に多数のデータベースが公開（KEGG，GenBankなど ）
  - データの種類
    - 分子
    - ゲノム
    - 分子間の相互作用
    - ネットワーク（相互作用の集合）

小

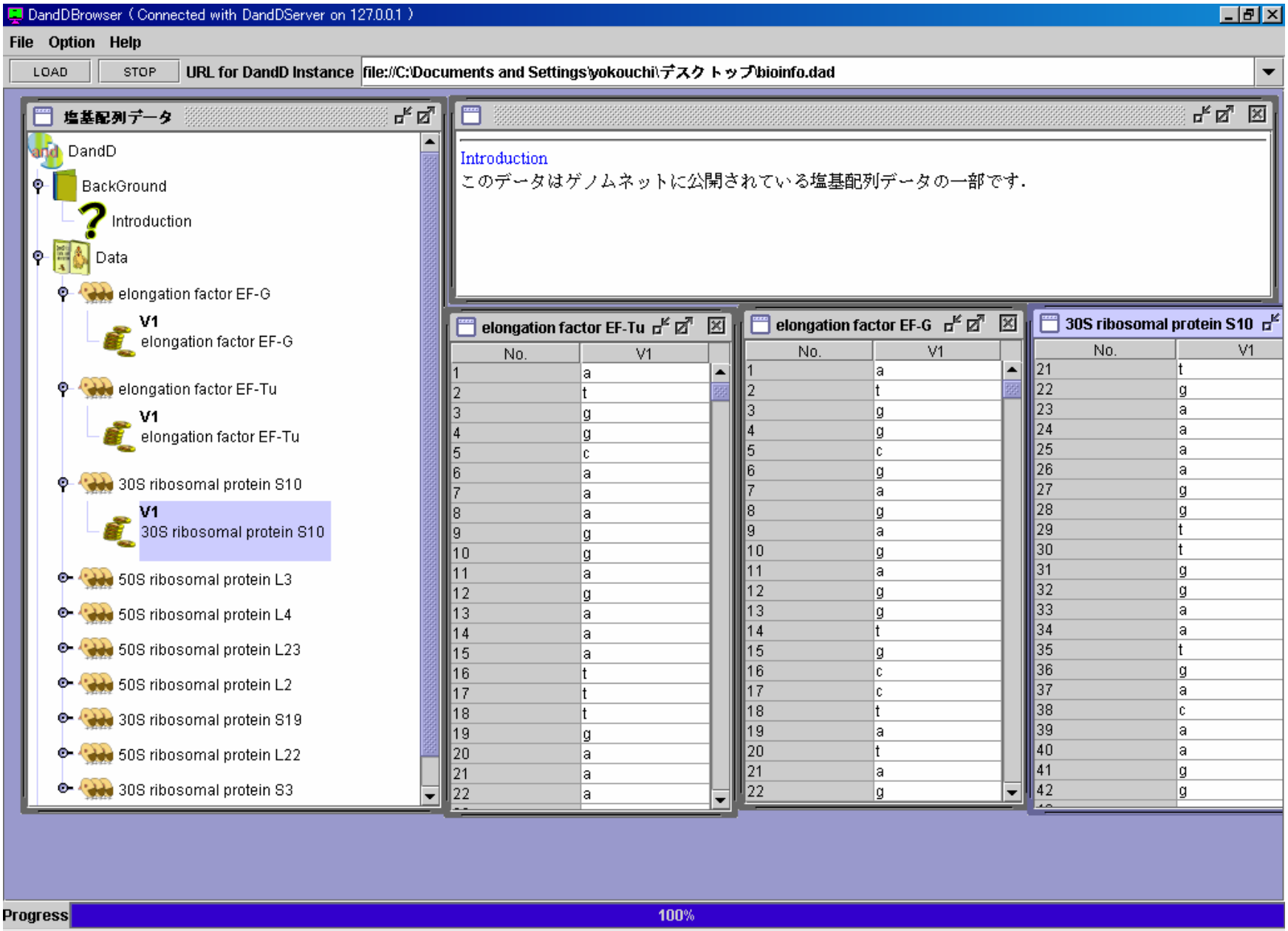


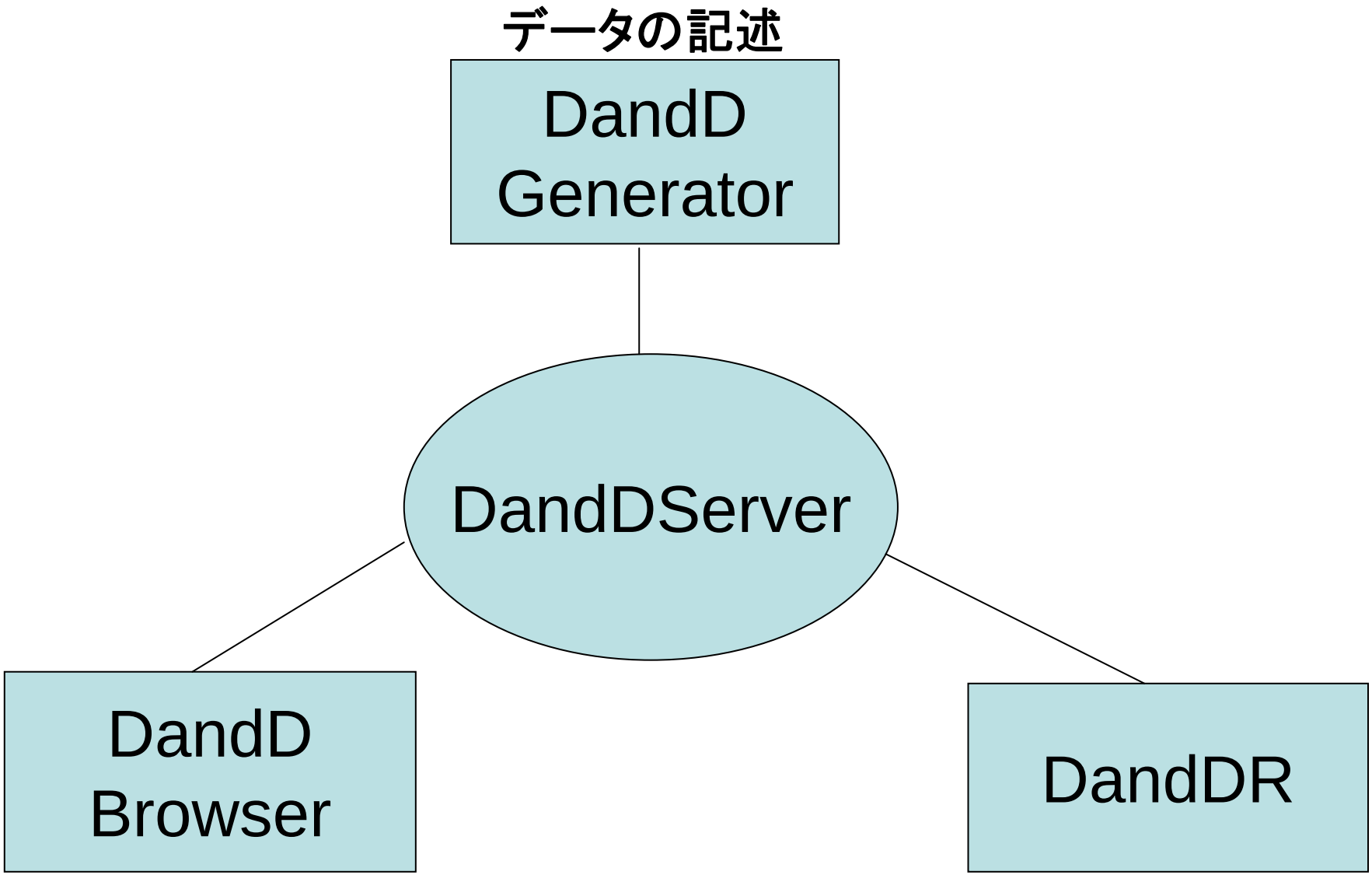
大

それぞれのデータが互いに密接に絡みあっている.


KEEGの塩基配列
データベース
Internet
Client
DandD
Server
DandD
Instance
FTP
Sequence
Sequence
Sequence
Sequence

例 GenomeNet（KEEG）




データの記述
DandD
Generator
DandDServer
DandD
Browser
DandDR
データのブラウジング
データ解析，モデル化

DandD IV 統合環境

- DandD IV
  - DandD ルール（Syntax）
    - DandD_2.2.0.dtd
  - DandDBrowser
    - DandDBrowser040903
  - DandDR
    - DandDR
  - DandDGenerator

http://www.stat.math.keio.ac.jp/DandD/