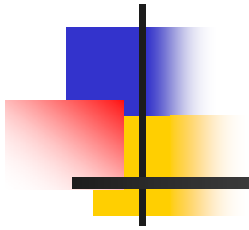
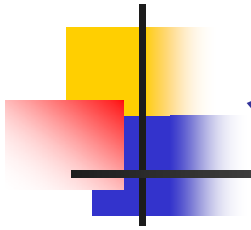


インターデータベース

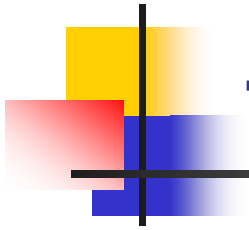
分散した異種データの統合

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 ルールと インターデータベース



- DandD（Data and Description)
- XML (eXtensible Markup Language)
- データはデータベクトルの集まり
- 実体，属性，関係の分離
- XML 文書（DandDインスタンス）上でデータベクトルを組織化し，データを表現する
- インターデータベースはDandDルールを背景に実装している

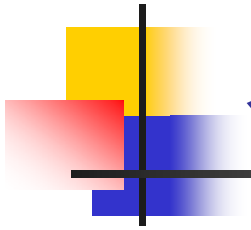


データベクトルの外部実体

- データベクトル
- 内部実体
  - DandDインスタンスの内部に直接記述する方式
- 外部実体
  - DandDインスタンスの外にあるデータを指定するための方式
  - データベクトルの実体の変わりに，データ取得に必要な属性を埋め込む

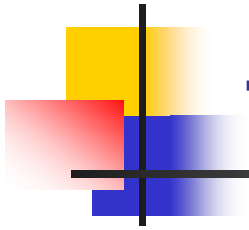
データベクトルの記述

- 内部実体
  - Gas.xml
  - 乗用車の給油記録データ
- 外部実体
  - JapanWeather.xml
  - 気象庁　地上気象観測原簿データ



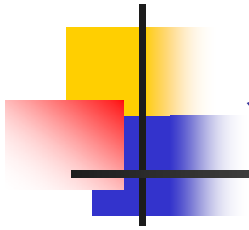
インターデータベース

- データベクトルの外部実体をもちいてネットワーク上に分散した異種データを統合する
- つまり，異種データを表すデータベクトルをDandD インスタンス上で構造化するによって一つのデータとして扱うことができる.
- ネットワーク上をまたぐ(つまりインターな)データベース



データの統合における諸問題

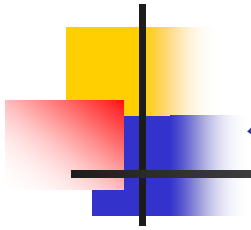
- 文字コード
  - データ間のエンコードの違い→DandD では全てUTF-16で扱う
  - DatabaseEncoding
- コーディング
  - データベクトルは数値のみ(欠損である NA を除く)
  - 欠損値の扱い　Missing　MissingType
  - コードの違い　DatabaseCode Code
- 記述形式
  - アプリケーション
  - 内容



インターデータベースの対象

- すべてのデータを扱うことは困難
  - ファイル1つをとっても新しく出てくるアプリケーションは新しい形式を生み出す
- DandD ルールが対象とするデータ
  - SQLの利用できるリレーショナルデータベース
  - 関係形式と見なすことのできるテキストファイル
- 理由
  - 大規模データはリレーショナルデータベースが主流
  - 表計算ソフトでもCSV形式のような自由欄形式のテキストに落とすことができる
  - どちらも関係形式なので SQL による統一的扱いが可能

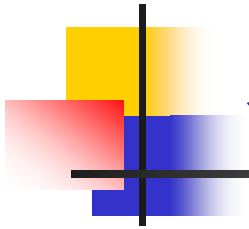
外部実体におけるデータベクトルの属性(1)



- **アクセスのための情報**
  - DatabaseServer　URL　UserId
  - DatabaseSever は「IPAdress：Port」
  - URL のプロトコルはhttp://,ftp://
- **データ特定のための情報**
  - DatabaseName　DatabaseTable　HTTPParameter
  - HTTPParameter は CGI 等のWebインターフェイスを介してデータを取得する際のパラメータを指定する属性
  - Web ブラウザを介して取得しなくてよい

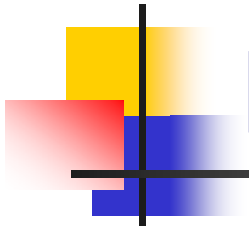
外部実体におけるデータベクトルの属性(2)

- **データ取得時に必要な情報**
  - Column DatabaseTable SelectCondition
  - これらは　SQL のSELECT文を構成
  - SELECT「Column」FROM「DatabaseTable」「SELECTCondition」
- **データ取得後に必要な情報**
  - ScanFormat DatabaseEncoding Offset DatabaseCode
  - Offset はヘッダーのような不要な行を読み飛ばすための属性
  - SELECT 文による取得
  - 日付型のデータ 1999-10-22 の月を指定するならば ScanFormat = “%*d-%d-%*d”



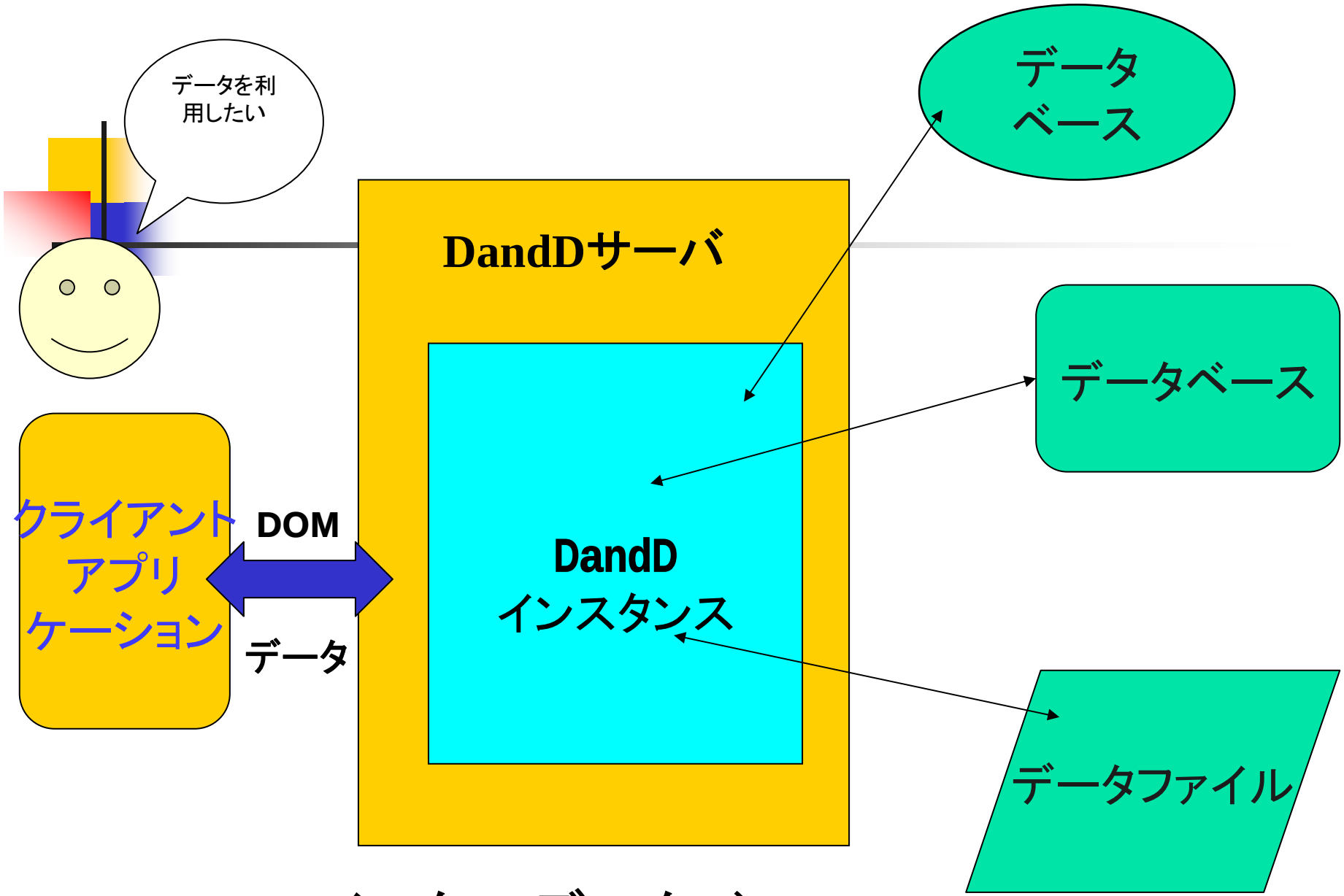
インターデータベースの実現

- DandD インスタンスは単なるXML文書
- 実現のためにはこの記述を解釈し，ネットワーク上のデータの取得をサポートするソフトウェアが必要
- DandD サーバ



DandD サーバ

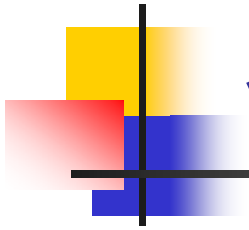
- DOM (Document Object Model)
- XMLのようなマークアップ言語の操作を定義したモデル
- データベースアクセスやファイル操作などの独自機能を追加
- ソケット通信


データを利
用したい
データ
ベース
DandDサーバ
データベース
クライアント
アプリ
ケーション
DOM
データ
DandD
インスタンス
データファイル

インターデータベース

インターデータベースのデモンストレーション

- DandD ブラウザ
- ユーザの要求に応じてDandD サーバとソケットによる通信を行い，データを表示するシステム
- １０の観測地点の気象データ（1990年12月，19991年1月）
- 最も大きなリレーショナル構造は１万記録（行）
- キーとなるデータベクトルを参照する Keys 属性を用意し，必要なデータをユーザに選択させる



今後の課題

- 大規模データ取得の効率化
  - Java から他言語への移植
  - 複数のDandDサーバによる分散処理
- DandD 化支援システム
  - 既存のデータベースのDandD化をサポート
  - DandD ルールを完全に理解していない人でも利用できることを目指す
- 解析ソフトウェアとの連携
  - 現在 C 言語とJava言語によるDandDサーバ用インターフェイスを作成(DandD ブラウザで利用している)