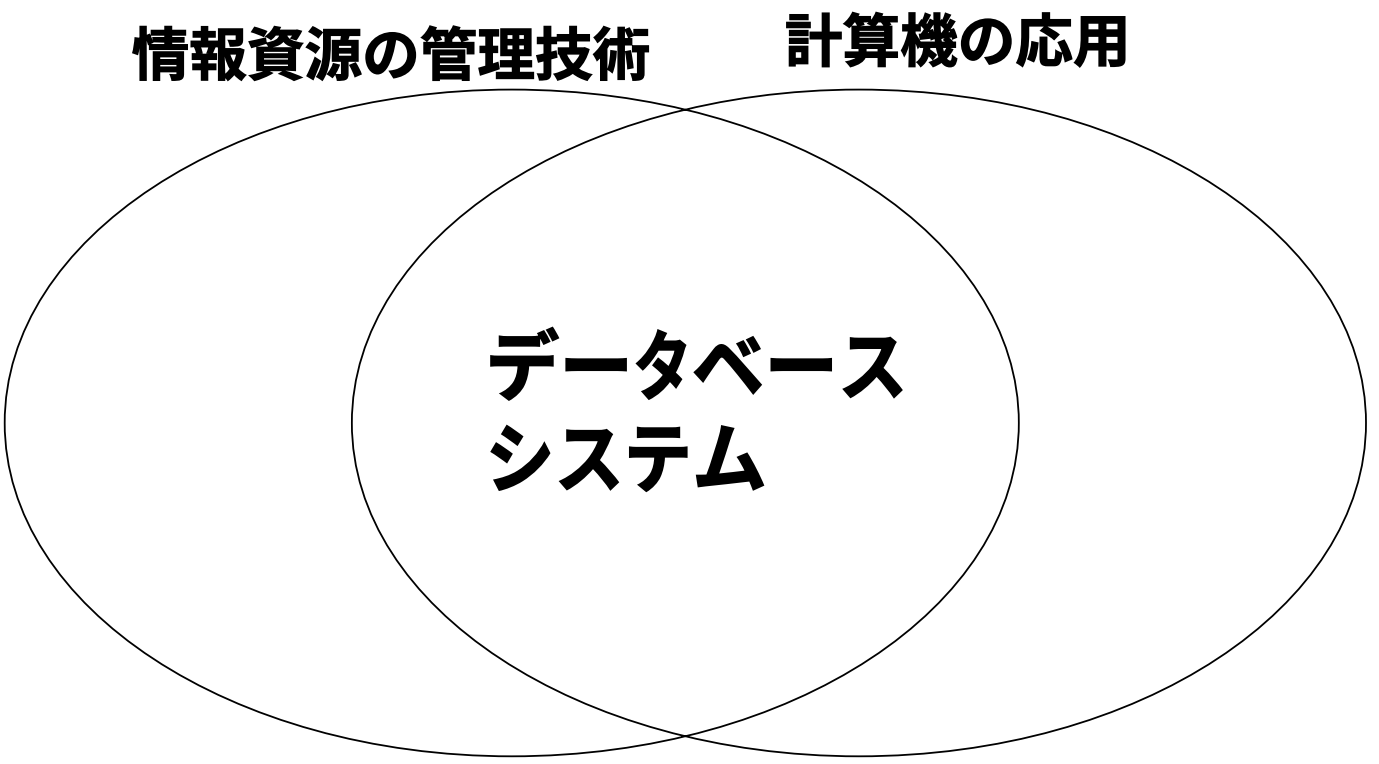
情報資源管理・有効利用のための データベース技術最前線

筑波大学知的コミュニティ基盤研究センター
知の情報基盤研究部門
森嶋厚行

データベースシステム



技術開発の歴史(系譜ではない)

- **ネットワークデータベース, 階層型データベース**

- **関係データベース**
  - E.F. Codd. A Relational Model of Data for Large Shared Data Banks, CACM, June 1970

- **オブジェクト指向データベース**(OODB)
  オブジェクト関係データベース(ORDB)
  演繹データベース

- XML**データベース？(あやしい)**

XMLデータベース?

<研究科リスト>

<研究科><名前>理工学</>

<種別>2年制</>

<1年定員>120</><2年定員>110</>

</> …

<研究科> <名前>図書館情報メディア</>

<種別>2+3区分制</>

<前期課程><1年定員30</><2年定員> 30</></>

<後期課程> <1年定員>10</> … </>

</>

- 半構造性(完全に規則的では無いこと)を許す
- (注意)XMLはデータモデルではない.

データと情報の違い

- データ … 記号の集まり
- 情報 … 受け手の知識を増加させるもの

計算機が扱うデータ

- データの構造

- データの寿命
  - 揮発データ
  - 永続データ
    - 更新する
    - 更新しない

- データの内容に関する制御
  - 集中制御可能か？

データの構造

- **構造データ**
  - **完全な規則性がある.**
- **半構造データ**
  - Not as rigid, regular, or complete as the structure required by the traditional database management systems.
- **非構造データ**
  - **計算機上での表現において明示的な規則構造が認められない**

半構造データ(例：XML)

<研究科リスト>
  <研究科><名前>理工学</>
    <種別>2年制</>
    <1年定員>120</><2年定員>110</>
  </>　…

  <研究科> <名前>図書館情報メディア</>
    <種別>2+3区分制</>
     <前期課程><1年定員>30</><2年定員> 30</></>
    <後期課程> <1年定員>10</> … </>
  </>

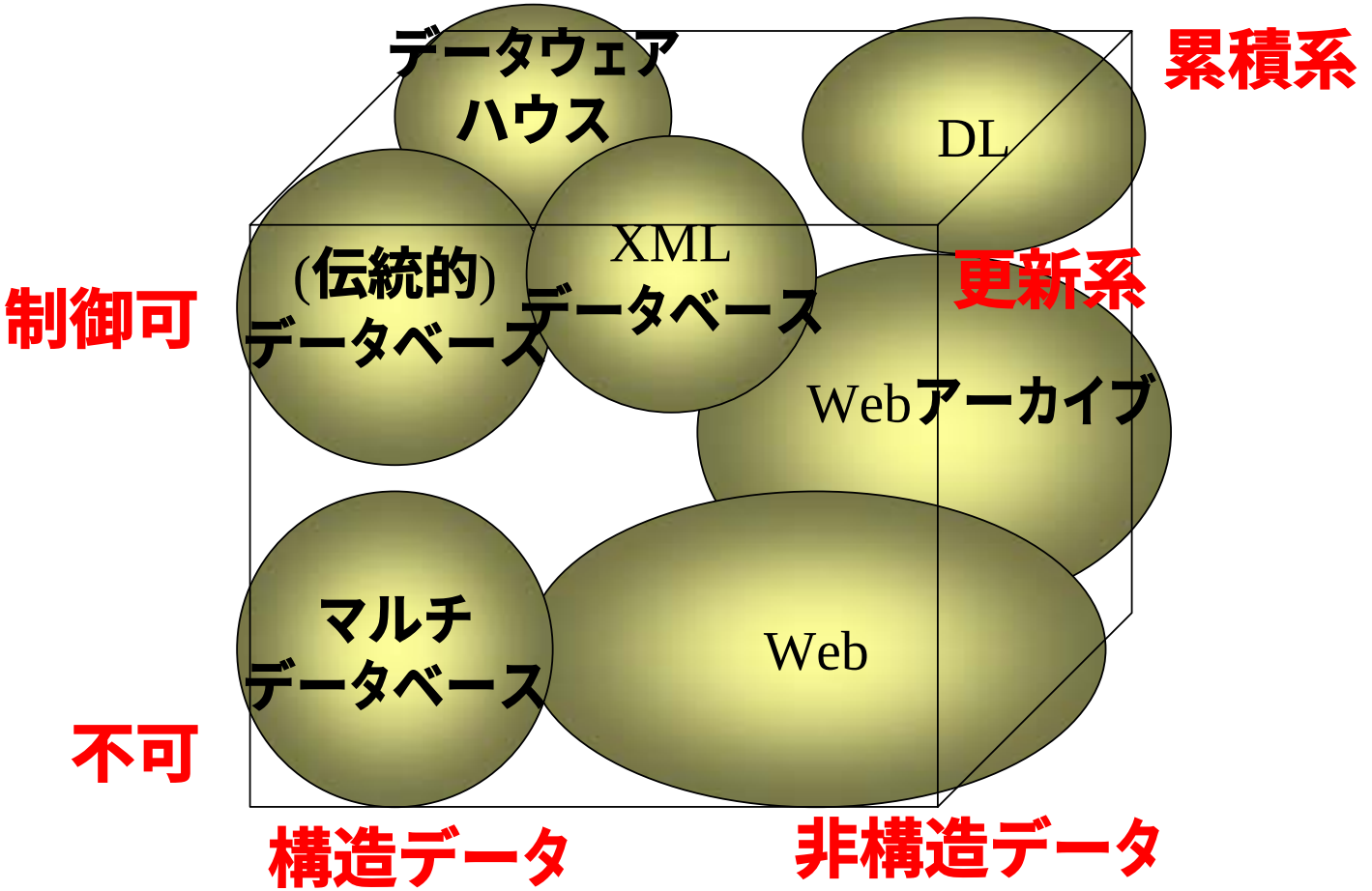
データの寿命

- 揮発データ
  - Webページ(の大多数)
  - スケジュール帳のデータ
- 永続データ
  - 更新系データ
    - 銀行の預貯金データ
  - 蓄積系データ
    - アーカイブ, データウェアハウス

データ内容に関する制御

- **管理者がデータの構造や制約を集中制御**
  - **例) 銀行預貯金データ**
- **不可能**
  - **例**) Web

様々な“情報資源”


データウェア
ハウス
DL
累積系
(伝統的)
データベース
XML
データベース
更新系
制御可
Webアーカイブ
マルチ
データベース
Web
不可
構造データ
非構造データ

(伝統的)データベース研究のキーワード

- Persistency
- Data Independence
- Structured Query
- Updates and Integrity
- Scalability
- Transactions

データベース研究者のコミュニティ

- Technologyの開発者
- Scalabilityは最重要
- ad-hoc approachesを嫌う
- Market-driven, Demand-driven

データベース研究の現在

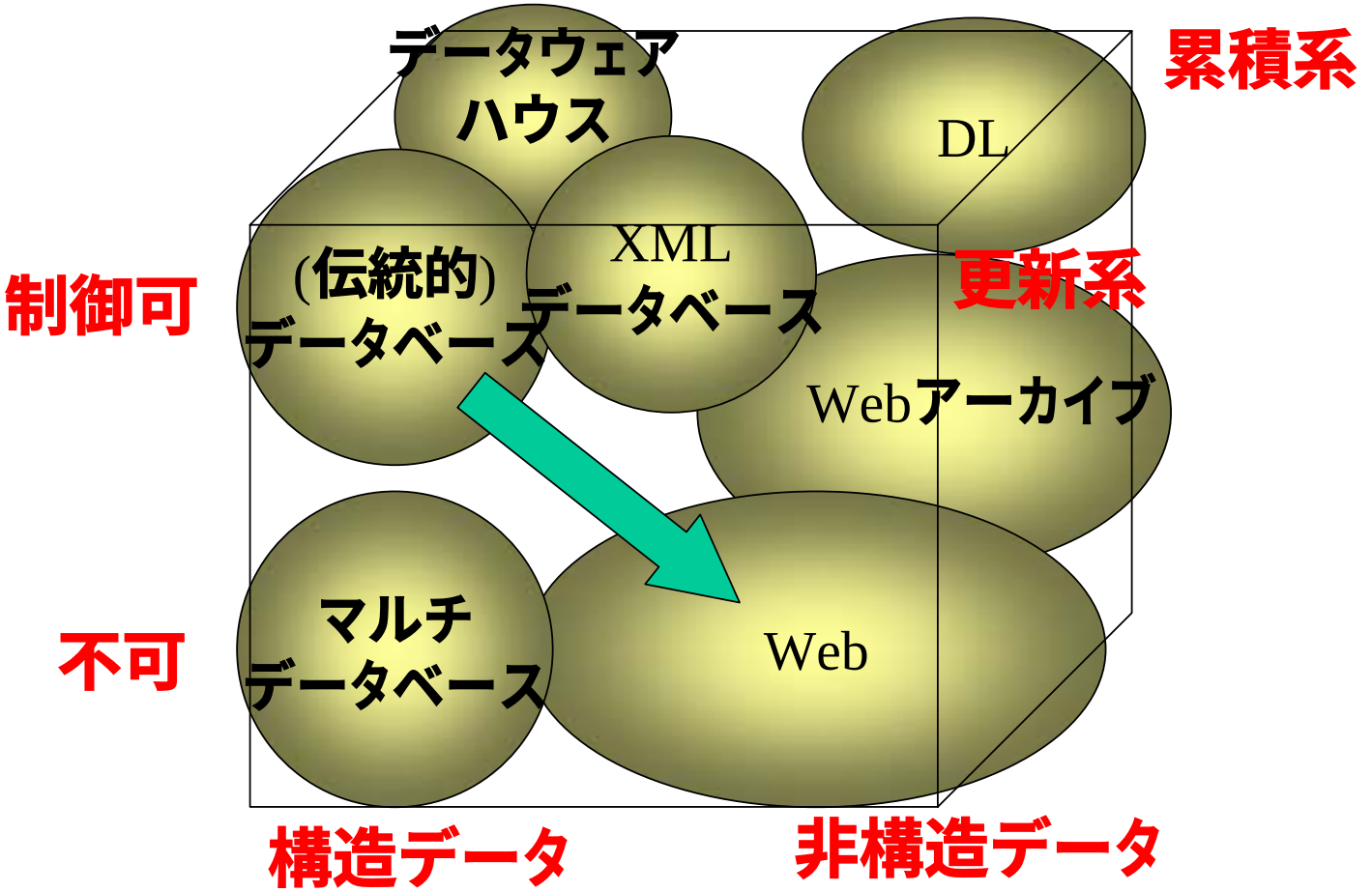
「Web Changed Everything」

Asilomar Report, 1998

The Grand Challenge

The Information Utility: Make it easy for everyone to store, organize, access, and analyze the majority of human information online

Web Changed Everything



Webの特徴

- データ: 半構造データ・非構造データ
- 管理：しない(できない)
- 寿命: 揮発データ ～　非更新永続データまで
- アーキテクチャ：分散

データ工学に関する研究の最近の傾向 (のいくつか)

- 構造：構造データ->半構造データ
- 管理：する->できない
- 他技術，とくにWeb技術，情報検索技術などとの統合
- 複雑さへの対応-> 自動化，半自動化

世界のデータベース技術研究動向
(ACM SIGMOD'03より)

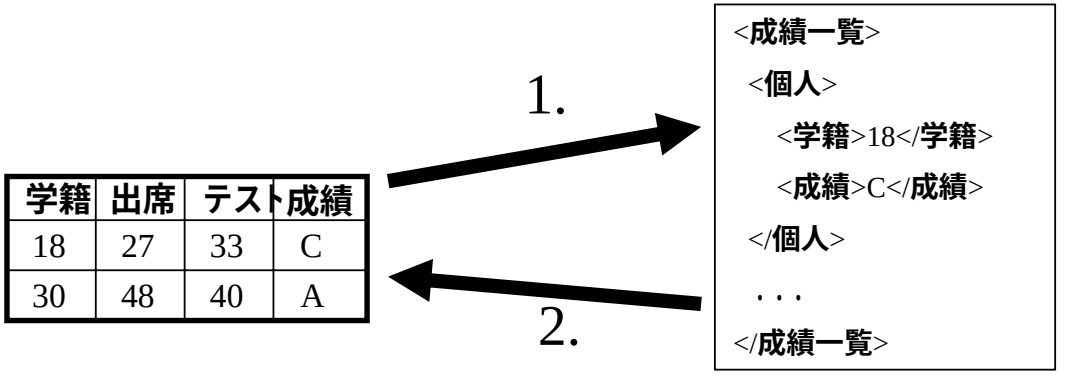
- XML　2+3+(2)+2+2 = 9+(2)
- **ストリーム処理** 2+2+(2)+2=6+(2)
- OLAP 2
- **データセキュリティ** 3
- **問合せ処理一般** 2+3 =5
- **時制問合せ** 2
- **メタデータ** 2
- **統計** 2
- **情報統合と共有** 2+3 =5
- **類似検索** 2+2 =4
- **理論** 2
- Spatial**問合せ** 4
- **センサーデータベース**2
- Approximate**検索** 2

XML

- XMLの半構造性をいかに扱うか？
  - 問合せ言語
- 大量のXMLをいかに扱うか
  - 問合せ処理
  - 圧縮
- リレーショナルデータベースに格納されたデータをいかに効率よくXML化するか
- Intentional XML

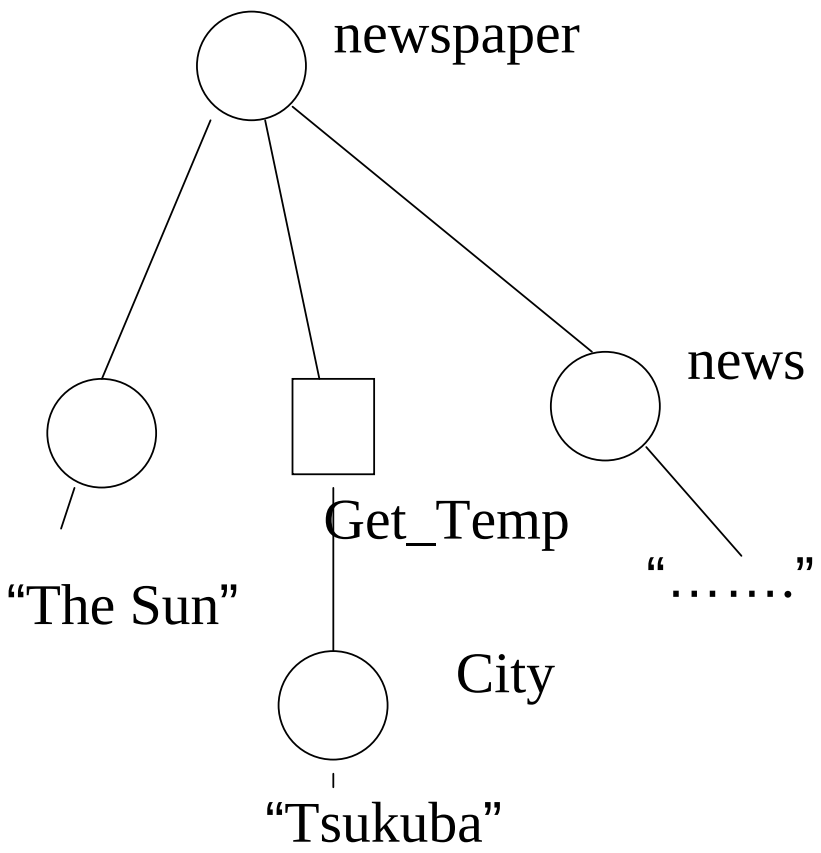
データベースコミュニティによるXML研究の最近の成果

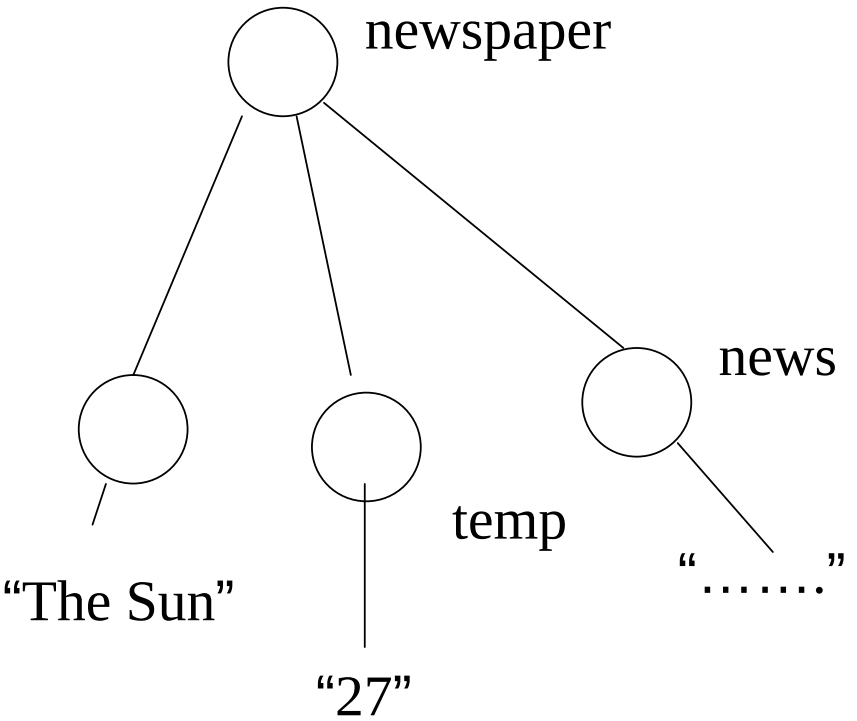
1. RDBをXMLに変換するための効率の良い方式の研究
2. XMLを効率よく扱うためにRDBエンジンを利用するための研究
3. Intentional XML



学籍	出席	テスト	成績
18	27	33	C
30	48	40	A

Intentional XML

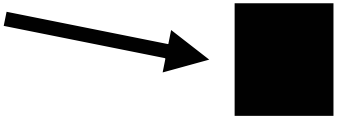




XMLの圧縮に関する成果

- XMLの構造を利用した効率の良い圧縮
- 問合せ可能な圧縮

```
<成績一覧>
 <個人>
   <学籍>18</学籍>
   <成績>C</成績>
 </個人>
 ・・・
</成績一覧>
```



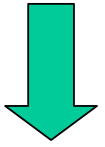
より大量のデータを格納可能へ

情報統合に関する問題

- スキーマ統合
- インスタンスレベルでの統合
- 構造情報の抽出
- 統合データに対する問合せ処理

スキーマ統合に関する最近の成果

- 複数のRDBの情報をXMLとして統合．その際，指定された制約を満たすようにする
- 複数の情報源のデータを比較し，スキーマ間のマッチングを発見



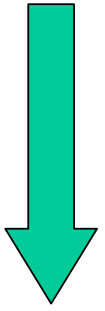
名前	学位
工学	博士 (工)
理工学	修士(理工)

研究科名	授与
工学	博士 (工)
理工学	修士(理工)

情報統合のための基礎技術

インスタンスレベルの統合に関する最近の成果

- Peer-to-Peer**環境におけるデータ統合**

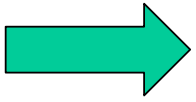


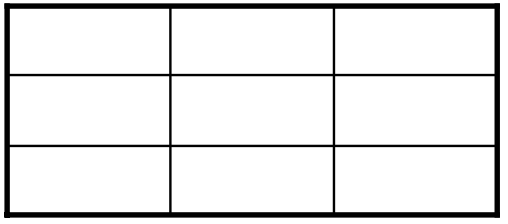
ID1	ID2
20020123	n123
20030064	m064

データインスタンス間の対応付けの推論へ

構造情報の抽出に関する最近の成果

- Webデータからの構造データ抽出





データベース技術の得意フィールドへ

最後に

- **データベース研究者は巨大データの扱いが得意です.**
- “Technology Driver”**となる，アプリケーションの視点からの新しい問題を求めています.**