データマイニング

序論

データマイニングの役割

- 記憶装置に蓄えられたデーター つ一つを普通、データレコードもしくはデータタプルと言う。

- 多くのデータレコードの集合体をデータベースと呼ぶ。

データマイニングの役割

- データベースのモデルと構築のブレイクスルーはリレーショナルデータベースであり、現在でもDBの中心的技術として使われている。

- データベースが効率的に管理されるようになると、もっと積極的にDBを利用したくなる。

データマイニングの役割

- DBで最も権威であるACM SIGMOD国際会議の１９９８年の論文集の前書きでデータマイニングという単語が出てくる。これがテーマである。

- データマイニングは巨大なデータベースをざっと見渡すことの自動化に対応した技術である。

データマイニングの役割

- DBで最も権威であるACM SIGMOD国際会議の１９９８年の論文集の前書きでデータマイニングという単語が出てくる。これがテーマである。

- データマイニングは巨大なデータベースをざっと見渡すことの自動化に対応した技術である。

データマイニングの役割

- ータの良し悪しを判別するのは困難であるが、これを数理的なモデルを設定して知識形態と評価基準をキメて基準値が高く形式に載っ飛んた知識を抽出するという形で具体化することが可能。このようなコンセプトをKnowledge Discoveryと呼ばれる。

- このノーレッジディスカバリーはかなり昔から研究が始まり、脳神経回路網を用いた実験研究も行われている。

データマイニングの役割

- そのため、現在ではKnowledge discovery and Data mining (KDD)と呼ばれることが多く、AI、DB、計算理論、統計学、経営学にまたがる横断的な学問分野となっている。

データマイニングの役割

- 以上より、データマイニングの役割は大規模データのIntelligentな圧縮、或いは視覚であるといえる。
- 特に、データマイニングが機械学習と区別される点はこれらが実用技術である点である。＝＞実際の巨大DBを高速処理し、圧縮された情報が実用化地を持つようにSystem設計自体を考えなければいけない。

データサイエンスと
データマイニング

- 統計の世界に於けるデータマイニングは「適切な仮説を前もってもたずに、虱潰し、デタラメにパターンを探すこと」と言うように、否定的な意味で用いられていたそうだ。

- computerの記憶装置の巨大化に伴い、この否定されてきたデータマイニングが脚光を浴びてきた。

データサイエンスと
データマイニング

- データマイニングの定義 「大規模なデータから思いがけない（unsuspercted）パターンを発見すること」

- 仮説検証型のアプローチ＝過程を立てるために、思いがけないパターンを発見することは出来ない。

データサイエンスと
データマイニング

- 統計学：データ全体を説明する様な大域モデルを構成する方法を採る。

- データマイニング：データの細かな一部でしか成り立たないようなパターンにしばしば注目する。

データサイエンスと
データマイニング

- データマイニングー＞探索的データ解析にごく近い目的を持っているが、あくまでも実用技術である。

データサイエンスとデータマイニング

- データマイニングの特徴まとめ

- 1.　非常に大規模なデータを対象とする

- 2.　データの収集法に対してコントロールが聞かないことが多い

- 3.　新しい種類のデータやパターンに注目している

- 4.　人間とcomputerが以下に役割を分担できるかに注目している]]]]]

簡単なデータマイニングの例：バスケット解析

- 誰が何を買ったというデータはPOSSystemにドンドンと溜まっていく。

- これらのデータを文字や数字の羅列、或いは表として見るだけでは非常に使いづらい。

- 例えば、「ソーセージを買った人の多くがロールパンも買っている」などの法則がデータに内在している。・・・①

- バスケット解析とは、①のような法則を自動的に抽出してデータを分かりやすい情報に変換する事であり、データマイニングの最も基礎的な例である。

簡単なデータマイニングの例：バスケット解析

- 上記の様な簡単な法則はグラフで表すことが出来る。

- 性格には有向グラフと呼ばれる物を用いる。

- このようなグラフを画面に書く技術をグラフドローイングと言う。

データマイニングシステムの構成

- バスケット解析でグラフの言葉で掛けるAならばBというルールを拡張し、A,Bの代わりに論理積を用いる事を許した物を相関ルールと言う。

- 相関ルールを幾つも複雑に組み合わせる事によりどんな概念でも記述が可能であるはずだが、組み合わせ爆発により計算コストが大きい。将来的には脳神経回路網などを用いることで利用できる様になるかもしれない。

データマイニングシステムの構成

・データマイニングシステムの設計は法則の明快性：「法則はできるだけシンプルで、ユーザーにとって分かりやすいものでなければならない」という基本欲求に乗っ取らなければならない。

データマイニングシステムの構成

- 法則の明快性は、何故その法則が成り立つのかをユーザーに理解してもらう必要が有るため、順守せねばならい。

- 理由不明だが、何故か答えが出てくるシステム：BLACK BOXはデータマイニングに置いて好ましくない。

データマイニングシステムの構成

- 例えば一万個の相関ルールが複雑に絡み合ったシステムは、実際の所BLACK BOXである。これは好ましくない。

- また、明快性は、「法則とは、類似データに対して普遍的に成立する必要が有る」ことからも重要である。

データマイニングシステムの構成

- sampleから全体に適用できるルールを予測類推しようと言うのがデータマイニングのアイディアである。

- これはサンプルからの学習であり、法則が複雑であれば有るほど、sampleの個性（統計的な学習の揺れ）による誤差が大きくなってしまう。

データマイニングシステムの構成

・この誤差を考えた法則の制度は予測精度と呼ばれる。

・これは知識抽出では最も大切な評価基準となる。

データマイニングシステムの構成

- BLACK BOX化を避けるために用いられる代表的なシステム構成は、単純なルールを用いて組み立てられた規則的な階層構造である。

- 相関ルールを部品として持ち、代表的な階層構造である木構造と回帰木。 これらは構造の大きさと予測精度に間sる理論も整備されており、効率が良い上に、安心して利用できるため、現在実用化されているシステムで成功したものだと言える。

データベース

関係データベース

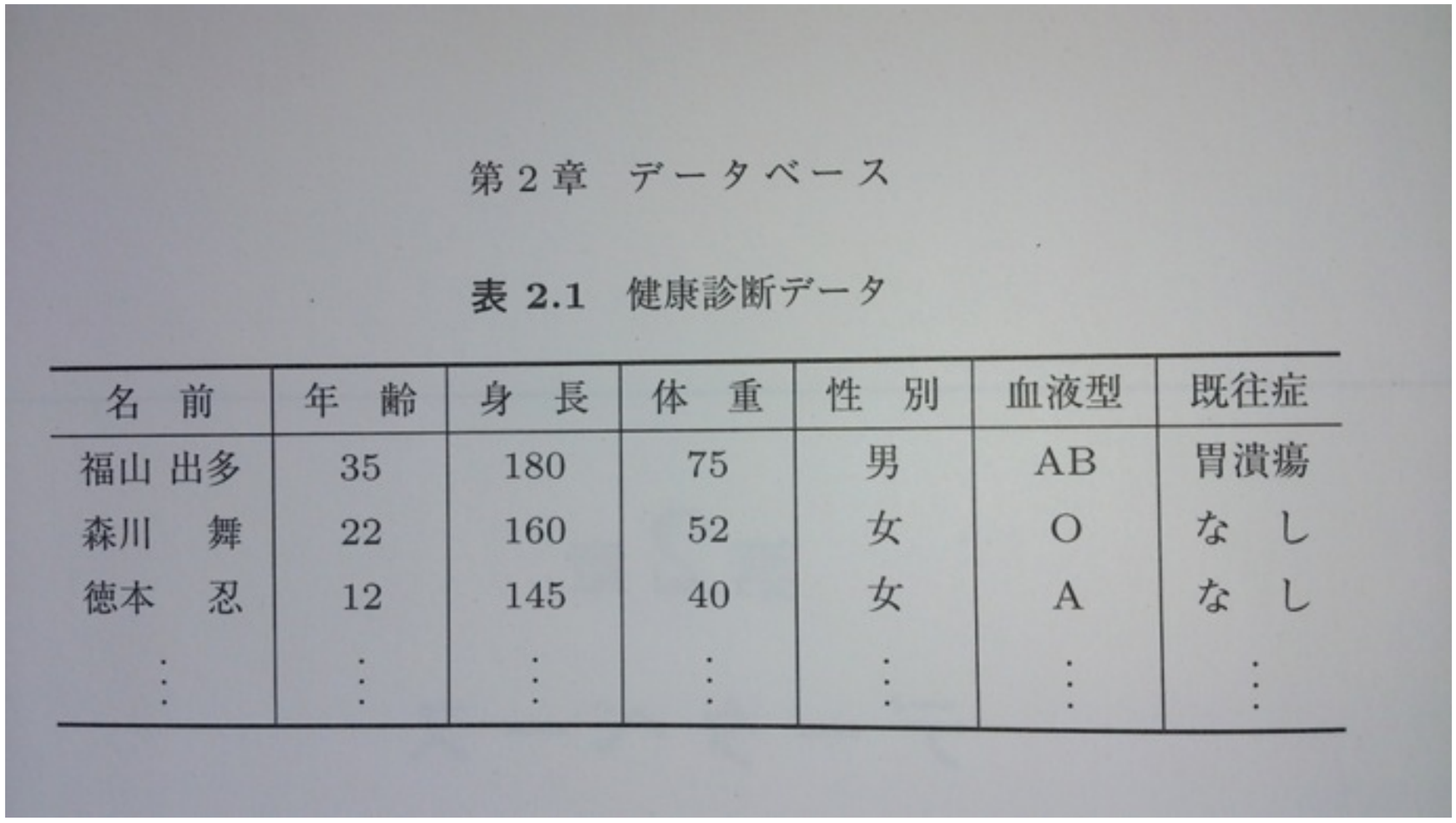
- データベース：複数の応用目的での共有を意図して組織的かつ永続的に格納されたデータ群
- データベース管理システム：このようなデータ群を作るための仕組み
- データベースシステム：データベース管理システムとデータ群を合わせたもの

関係データベース

- 関係データベース（RDB）：現在最も一般的に使われているDB

- relation：ぶち抜きや入れ子のない、二次元の単純な表（テーブル）

関係データベース



第２章　データベース

表 2.1　健康診断データ

名　前	年　齢	身　長	体　重	性　別	血液型	既往症
福山 出多	35	180	75	男	AB	胃潰瘍
森川　舞	22	160	52	女	O	な　し
徳本　忍	12	145	40	女	A	な　し
⋮	⋮	⋮	⋮	⋮	⋮	⋮

関係データベース

- レコード、タプル：行、一つのデータのまとまり
- タプル：列、タプルの表している対象のもつ属性
- relationはタプルの集合である＝重複は無し

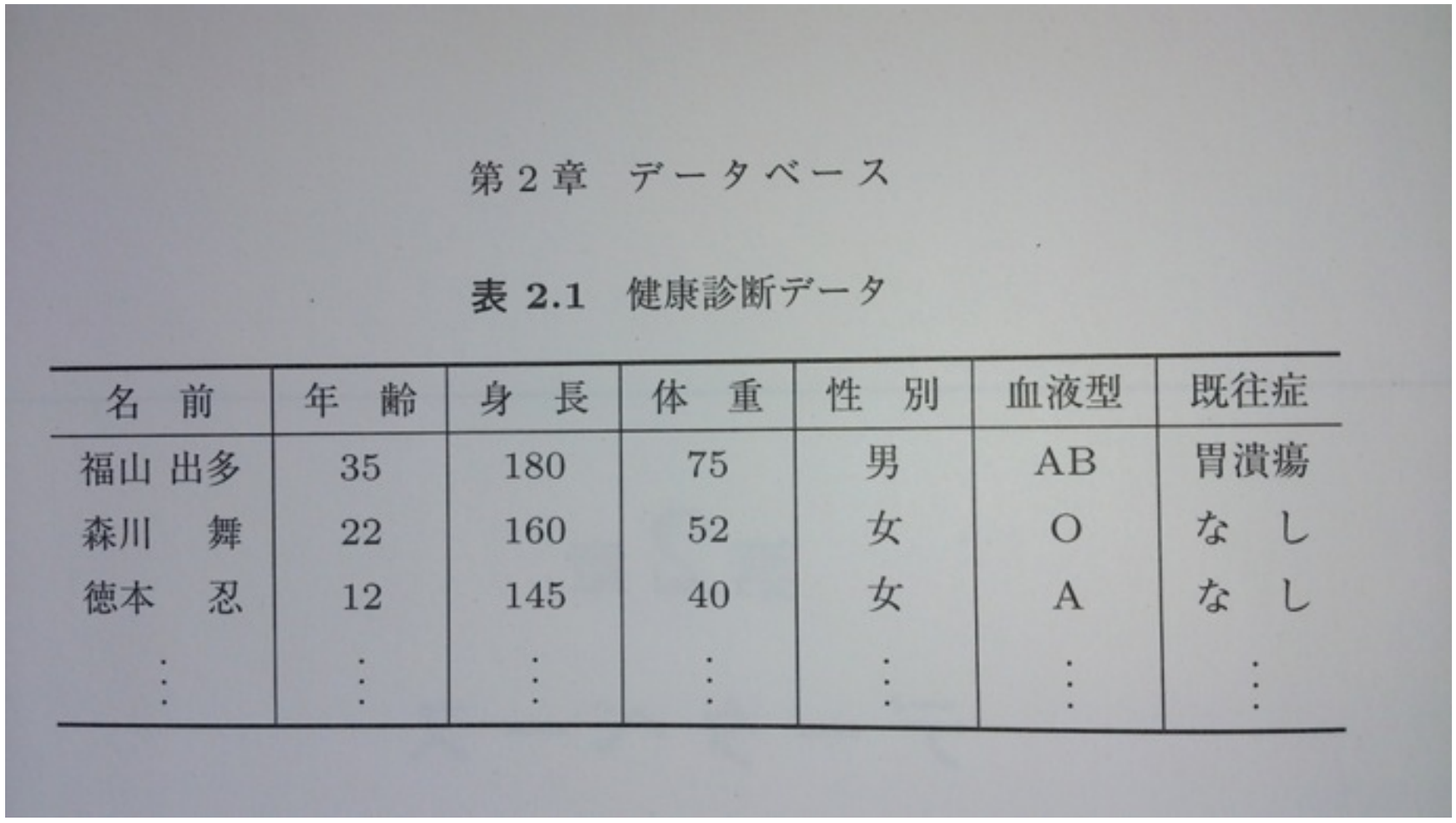
関係データベース

- DBの表記法
- R：あるrelation（表）
- t∈R：Rはタプルt（行）の集合である
- R(A,B,C)：属性A,B,Cを持つrelation

関係データベース

- t[A]：あるタプルt∈Rの属性A
- dom(A),Aの定義域,(domain):ある属性Aの取りうる値の全体集合

関係データベース



第２章　データベース

表 2.1　健康診断データ

名　前	年　齢	身　長	体　重	性　別	血液型	既往症
福山 出多	35	180	75	男	AB	胃潰瘍
森川　舞	22	160	52	女	O	な　し
徳本　忍	12	145	40	女	A	な　し
⋮	⋮	⋮	⋮	⋮	⋮	⋮

関係データベース

・ 健康診断データ(名前、年齢、身長、…)と表記され、t=(福山　出多，35,180,75，…)に関して属性「身長」の値はt[身長]=180である。属性「血液型」の定義域はdom(血液型)=｛A,B,O,AB｝である。

トランザクション

- Transaction：ひとまとまりの処理単位
- データベースのACID性(atomicity, consistency,integrity,durability)：トランザクション処理に求められる特性
- “Atomicity”(原子性)、“Consistency” (一貫性)、“Isolation”(独立性)、“Durability”(耐久性)

トランザクション

- Atomicity(原子性、不可分性)：トランザクションに含まれる個々の手順が「すべて実行される」か「一つも実行されない」のどちらかの状態になるという性質

- Consistency(一貫性)：トランザクションの前後でデータの整合性が保たれ、矛盾の無い状態が継続される性質

トランザクション

- Isolation(独立性、隔離性)：トランザクション実行中の処理過程が外部から隠蔽され、他の処理などに影響を与えない性質

- Durability(耐久性、持続性)：トランザクションが完了したら、その結果は記録され、システム障害などが生じても失われることがないという性質である。

トランザクション

- 同時並行処理されるトランザクション同士が干渉してはならない。システム障害が起こってもトランザクションが途中終了する様なことが無い様に保証する。これらがデータベースのACID性を保証する機能を提供する

問合せ

- 問い合わせ(query)：DBが受け取るデータに対する問い合わせ。命令。

- RDBに対するqueryは関係代数と呼ばれる体系に基づき、定義されている。

問合せ

- 関係代数では基本代数演算子を順次適用する事によりqueryを表現する。

- 基本代数演算子：和、差、直積、斜影、選択

問合せ

- 和(union)：RvS、2つのrelationの和集合

- 差(difference)：R-S、2つのrelationの差集合

- 直積(Cartesian product)：R×S、2つのrelation
  それぞれから一つづつタプルを取り出し、ソレを繋げたタプルを含むrelationを作る。

問合せ

- 斜影(projection)：relationが持つ属性の内、指定した属性だけを残して他を削除する

- 選択(selection)：relationが持つタプルの内、指定した条件Fを満たすものだけを残し、他を削除

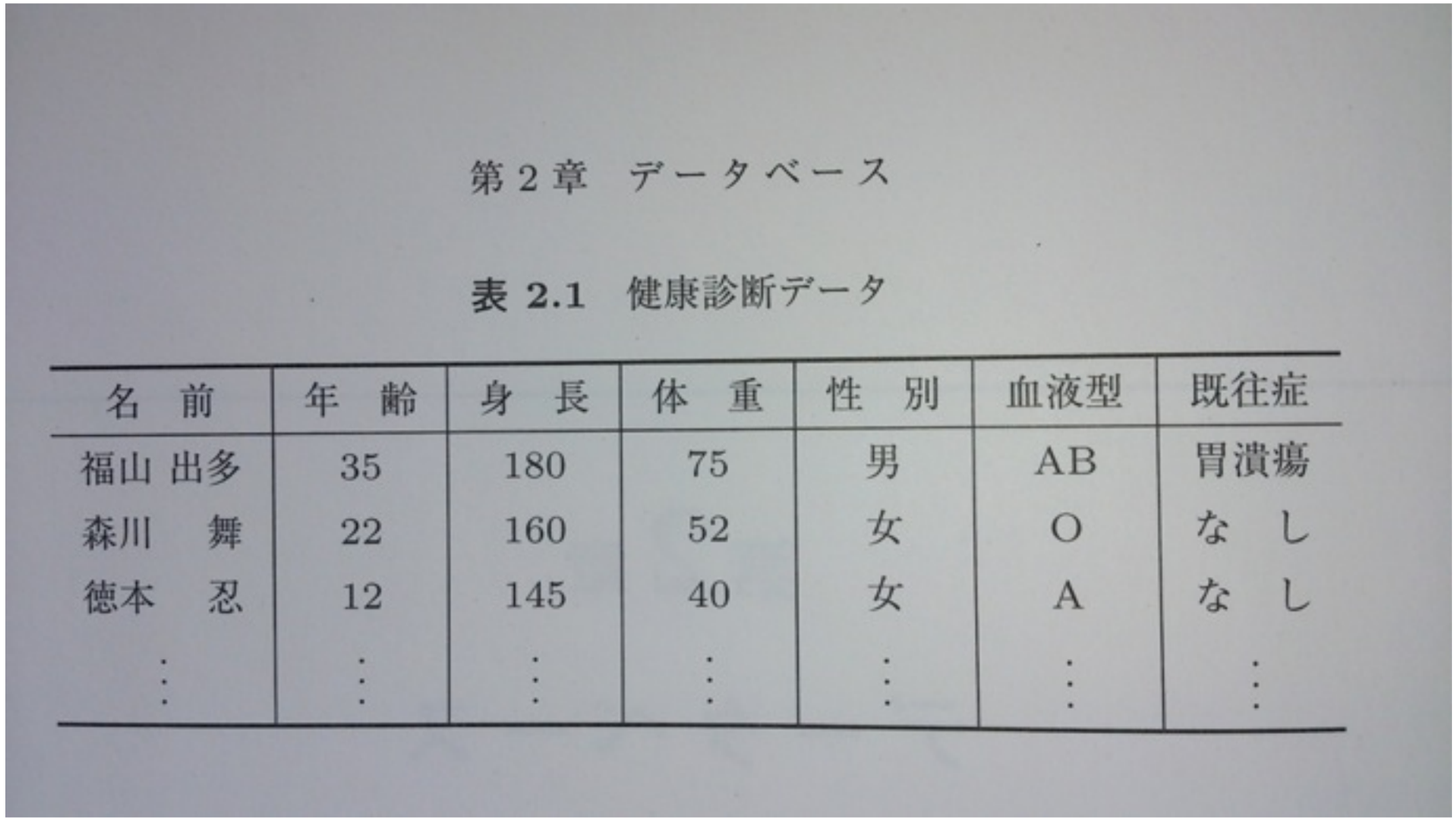
問合せ

- 問合せ言語（query　language)：queryを記述する為の言語

- SQL：最もよく使われる問合せ言語

関係データベース

- 身長が150以上を満たすレコードを取り出す



第２章　データベース

表 2.1　健康診断データ

名　前	年　齢	身　長	体　重	性　別	血液型	既往症
福山 出多	35	180	75	男	AB	胃潰瘍
森川　舞	22	160	52	女	O	な　し
徳本　忍	12	145	40	女	A	な　し
⋮	⋮	⋮	⋮	⋮	⋮	⋮

問合せ

- 身長が150以上を満たすレコードを取り出す

```
SELECT 名前，年齢，身長，体重，性別，血液型，既往症
FROM 健康診断データ
WHERE 身長 >= 150
```

表 2.2　問合せ結果

名　前	年齢	身長	体重	性別	血液型	既往症
福山 出多	35	180	75	男	AB	胃潰瘍
森川　舞	22	160	52	女	O	な　し

問合せ

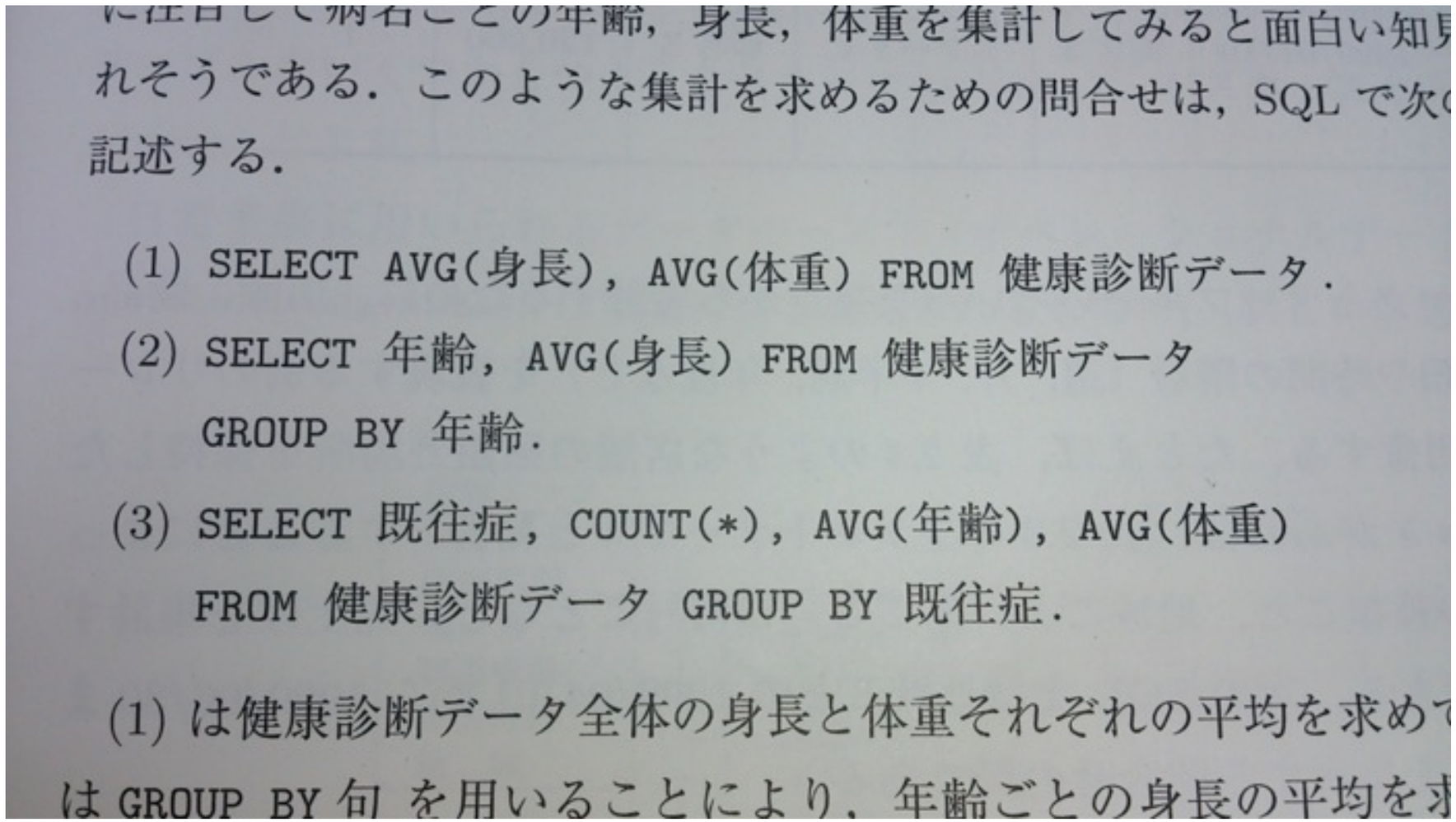
- SQLのqueryではFROMの後ろに問合せ対象のrelation名を指定、SELECTの後に問合せ結果の属性を記述、WHEREの後ろに条件を記述する。

集約演算

- データの大まかな特徴を理解したい→平均、最大、最少、分散等の代表値に集約すると役立つ

集約演算

- SQLで集計を求めるためのquery



に注目して病名ごとの年齢，身長，体重を集計してみると面白い知見
れそうである．このような集計を求めるための問合せは，SQL で次の
記述する．

(1) SELECT AVG(身長), AVG(体重) FROM 健康診断データ.

(2) SELECT 年齢, AVG(身長) FROM 健康診断データ
GROUP BY 年齢.

(3) SELECT 既往症, COUNT(*), AVG(年齢), AVG(体重)
FROM 健康診断データ GROUP BY 既往症.

(1) は健康診断データ全体の身長と体重それぞれの平均を求めて
は GROUP BY 句 を用いることにより，年齢ごとの身長の平均を求

データウェアハウスとOLAP

- ファクトテーブル：表2.3の様な事実を集めたrelation。この会社がどの商品でどの顧客からいくら収益を上げた、などを調べることができる。

るかといったことを調べることができる．このような事実を集めたリレーションをファクトテーブル (fact table) とよぶ．

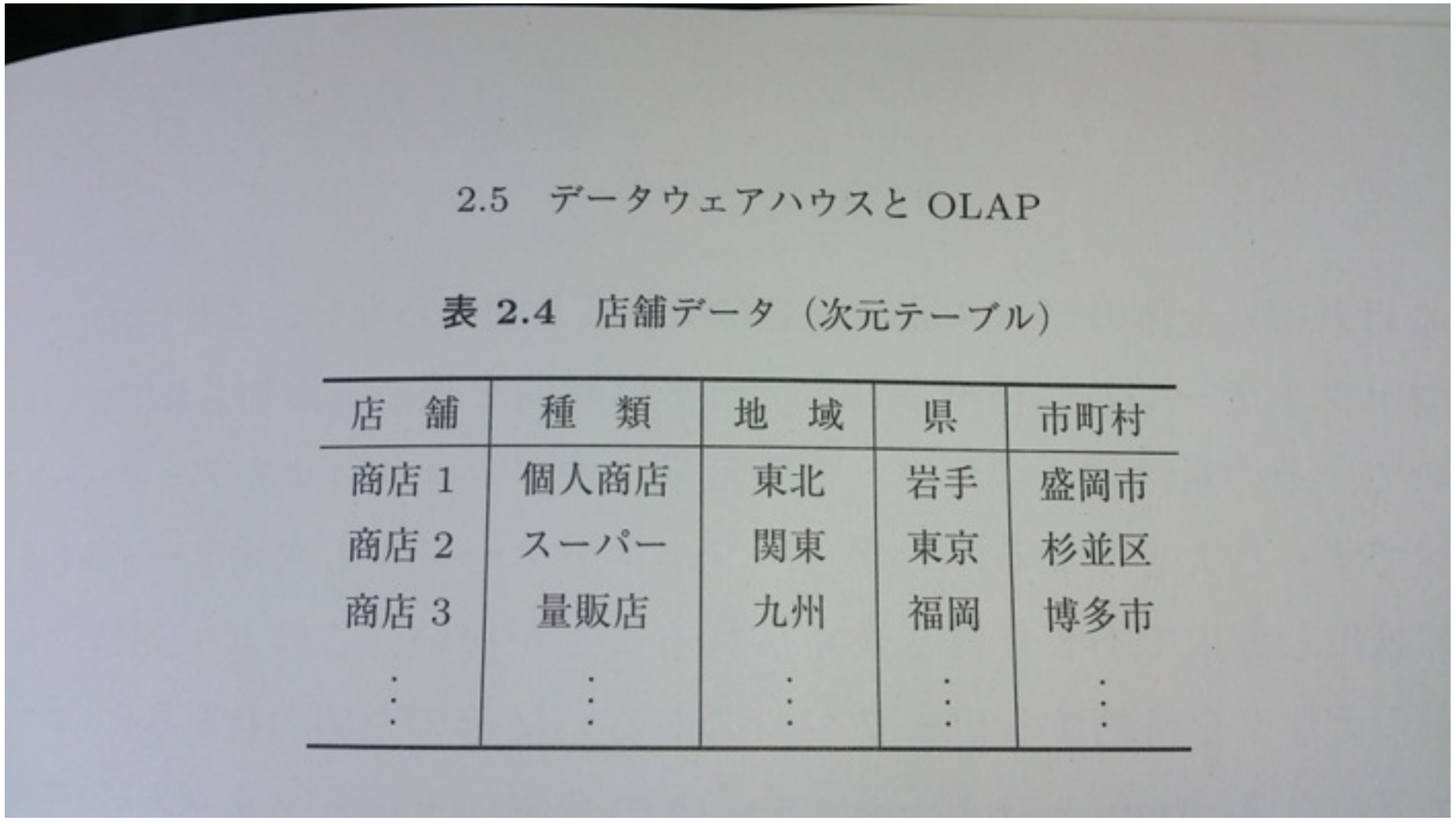
表 2.3 販売データ（ファクトテーブル）

日　時	店　舗	商　品	顧　客	単　価	数　量
1999/07/16	商店 1	自転車 A	顧客 X	20,000	2
1999/07/16	商店 2	一輪車 B	顧客 Y	8,000	4
1999/07/19	商店 3	スクータ C	顧客 Z	120,000	1
⋮	⋮	⋮	⋮	⋮	⋮

このファクトテーブルをいろいろな視点から集計するために，店舗，商品

データウェアハウスとOLAP

- 先ほどのファクトテーブルを様々な視点から集計するために、店舗や商品、顧客の分類や時間の階層を表現する別のrelationを用意する。



2.5　データウェアハウスと OLAP

表 2.4　店舗データ（次元テーブル）

店　舗	種　類	地　域	県	市町村
商店 1	個人商店	東北	岩手	盛岡市
商店 2	スーパー	関東	東京	杉並区
商店 3	量販店	九州	福岡	博多市
⋮	⋮	⋮	⋮	⋮

データウェアハウスとOLAP

- 表2.4の様なrelationがあれば、2.3と結合することで店舗の種類ごと、地域ごと、県事、市町村ごとなどの売上を集計することが出来る。

データウェアハウスとOLAP

- 次のSQL文は地域ごとの1999/04/01~1999/06/30までの売上高を集計するqueryである。

```
SELECT 地域, SUM(単価 * 数量)
FROM 販売データ A, 店舗データ B
WHERE A.店舗 = B.店舗 AND
      日時 BETWEEN 1999/04/01 AND 1999/06/30
GROUP BY 地域.
```

データウェアハウスとOLAP

- データキューブ：幾つもの分類の軸を自由に組み合わせてその組合せの任意の位置が対応するので、dataCubeと呼ぶ
- 次元：集計の軸となる属性
- 次元テーブル：次元を保持するrelation

データウェアハウスとOLAP

- スタースキーマ（star　schema）：ファクトテーブル周りに次元テーブルをいくつか持つようなスキーマ

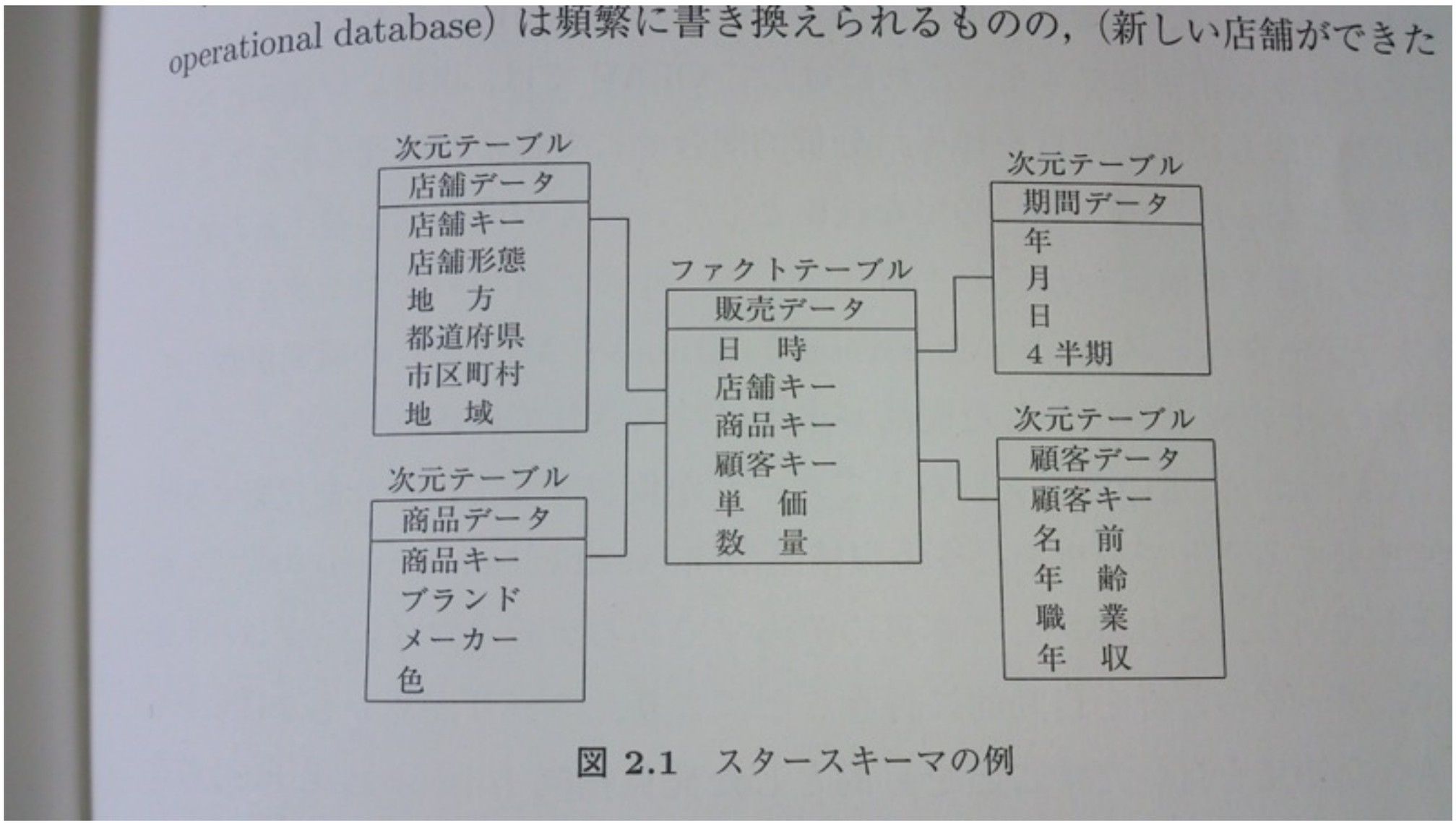


図 2.1　スタースキーマの例

データウェアハウスとOLAP

- データウェアハウス（DWH ）：時系列に蓄積された大量の業務データの中から、各項目間の関連性を分析するシステム。単純な例をあげると、コンビニの売上データから「月曜日に雑誌を買う30代の男性は一緒にコーヒーを買うことが多い」「肉まんは雨の日に最もよく売れる」など、従来の単純な集計では明らかにならなかった各要素間の関連を洗い出してくれるのがデータウェアハウスシステムである。

データウェアハウスとOLAP

- オーラップ
- OLAP 【 On-line Analytical Processing 】 オンライン分析処理
- 企業が顧客データや販売データを蓄積したデータベースを多次元的に解析し、視覚化するシステム。データウェアハウスなどを使って集められた大量の元データを多次元データベースに格納し、これを様々な角度から検索・集計して問題点や解決策を発見する。例えば、顧客の購入履歴を解析し、売上を地域別や製品別、月別など様々な次元から瞬時に分析することができる。情報技術部門ではなく、解析結果を必要としている部門の人間(エンドユーザ)が直接システムを操作して解析を行う点が従来の解析システムと異なる。設計の違いからROLAPとMOLAPの2種類に大別される。

データウェアハウスとOLAP

- オーエルティーピー
- OLTP 【 On-Line Transaction Processing 】 オンライントランザクション処理

- ネットワークに接続された複数の端末がホストコンピュータに処理要求を行い、ホストコンピュータが処理要求にもとづいてデータを処理し、処理結果を即座に端末に送り返す処理方式。データベースアクセスを伴うことが多く、途中で処理が中断されてしまうとデータの整合性が取れなくなるため、高い信頼性が要求される。

データウェアハウスとOLAP

- 仮説検証指向：OLAPが対象としている、大量のデータに対してユーザが仮説を建てて、ソレを検証するqueryを発することで分析を行う課題
- 発見指向：データマイニングがサポートしているものであり、仮説検証指向から一歩進んだ、データの分析を自動的に行うことによりユーザが思いもかけなかった様な知見を見出すことを目的とした課題。