

機械学習 と その最近の話題

～ 機械学習概論＋ネットワーク＋クラウドソーシング ～

(と ヒューマンコンピューテーション)

鹿島久嗣
数理情報学専攻
情報理工学系研究科



概要： 機械学習の概要を紹介したあと、 この分野で最近注目されている話題を紹介します

1. 機械学習概論

- データからの予測と発見

2. ネットワークと機械学習

- 個々のデータから、データ間の関係へ

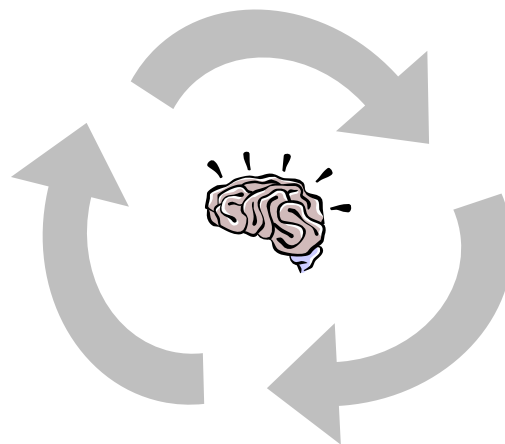
3. 機械学習とクラウドソーシング

- 機械と人間の協調問題解決

10年間の企業研究所勤務ののち、2009年に大学へ異動 機械学習の応用研究に携わる

- 1999年に京都大学工学研究科システム科学専攻を修士課程修了、以降、10年間IBM東京基礎研究所にて研究員として勤務
 - バイオインフォマティクス、コンピュータシステムの障害解析、ビジネス・データ解析（購買管理、人材マネジメント、マーケティング）、製造システム/自動車のセンサーデータ解析、特許データ分析
 - データ解析コンサルティング
 - グラフ構造データを対象とした機械学習手法
- 2009年から東京大学情報理工学系研究科数理情報学専攻数理6研 准教授
- 「機械学習（データ解析）をより多くの重要な場面で活躍できるようにする」
 - これまで扱うことができなかった形式のデータや問題設定などを見つける

機械学習概論



例 1 あるなしクイズ：これは「あり」？「なし」？

- ヒント：「あり」なものと、「なし」なもの

あり	なし
うさぎ	ねずみ
はがねのつるぎ	てつのおの
きんとき	あんこ
たわし	わたし

- では…
 - 「ししゃも」は？
 - 「ほっけ」は？
 - 「しゃけ」は？

部分文字列に注目してみると… 判別するルールが 出てきます

- ヒント：「あり」なものと、「なし」なもの

あり	なし
うさぎ	ねずみ
はがねのつるぎ	てつのおの
きんとき	あんこ
たわし	わたし

- では…

- 「ししゃも」は？ ⇒ あり
- 「ほっけ」は？ ⇒ なし
- 「しゃけ」は？ ⇒ なし

「あり」のグループには鳥の名前が含まれている

例 2 なかまはずれさがし：仲間はずれはどれ？

- 以下のうち、仲間はずれは どれでしょうか？

くも
やどかり
たこ
いか
たらばがに
毛がに
えび

グループ分けしてみると…なかまはずれが 見えてきます

- 「足の数」と「かたさ」で分類してみると…

		足の数	
		8本	10本
かたさ	やわらかい	くも たこ	いか
	かたい	たらばがに やどかり	毛がに えび

グループ 1 (くも, たこ)

グループ 2 (くも, たらばがに, やどかり)

グループ 3 (いか, 毛がに, えび)

- あるいはもっと安直に、棲んでいる場所に注目すると「くも」であろう

棲んでいる場所	
陸上	水中
くも	その他

前述の例は、それぞれ機械学習の2大タスクである
「教師つき学習（予測）」と「教師なし学習（発見）」に対応しています

- あるなしクイズの場合：
 - 「ある」「なし」を区別するルールを与えられた事例から見つける
 - 未知の対象に対してルールを適用し分類する
- なかまはずれ探しの場合：
 - ある視点から対象をグループ分けする
 - それぞれのメンバーを評価
- これらはそれぞれ機械学習の2大タスク
 - 「教師つき学習」= 予測
 - 「教師なし学習」= 発見に対応している

教師付き学習と教師無し学習は機械学習の基本問題です

- 機械学習では、学習者を、入出力のあるシステムと捉え、学習者に対する入力と、それに対する出力の関係を数理的にモデル化する
 - 入力：視覚などからの信号（実数値ベクトルで表現）
 - 出力：入力を表す概念、入力に対してとる行動
- どうやら2つの重要な基本問題があるらしいということになった
 - 教師付き学習：入力に対する出力を試行錯誤するうちに、どういう入力
のときにどういう出力をすればよいかがわかってくる
 - 教師無し学習：入力を見ているうちに、どんなものが現れやすいかなどの
パターンが分かってくる

形式的にいうと 教師つき学習は、入出力関係の推定問題です

- 目的 : 入力 x が与えられたとき、対応する出力 y を予測したい
 - 入力 x : 「ししゃも」や「ねずみ」
 - 出力 y : 「あり」か「なし」か

※ 厳密にはこれは教師つき学習の「分類」と呼ばれるタスク
- つまり、 $y = f(x)$ となる関数 f がほしい
- しかし、ヒントなしではこれではできない…
そこでヒント（過去の事例＝訓練データ）が必要
 - 「うさぎ」は「あり」、「ねずみ」は「なし」、など
- 訓練データをもとに入出力関係 f を推定するのが教師つき学習
 - 正しい出力を与えてくれる「教師」がいるというイメージ
 - 訓練データは f を「訓練する」ためのデータ

一方、教師なし学習は、入力データのグループ分け問題です

- 教師なし学習では入出力関係についてのヒントがない
(出力が与えられず、入力のみが与えられる)
 - 入力だけから出力らしきものをつくる必要がある (= 自習)
 - 「あり」「なし」などのラベルが明示的に与えられないので、グループ分けくらいしかできない
 - 目的 : 入力 x が与えられたとき、これらをグループ分けしたい
 - 入力 x : 「くも」や「やどかり」
 - 出力 y : グループ 1、グループ 2、... など
(明示的なラベルを付ける必要は無い)
 - 通常グループの数は指定される
- ※ 厳密には教師なし学習の「クラスタリング」と呼ばれるタスク

歴史的経緯： 機械学習とは、データ分析技術の一流派のようなものです

- 機械学習とは、本来「人間のもつ”学習能力”を機械（計算機）にも持たせる」ことを目指す研究分野
 - もともとは人工知能の一分野として始まる
 - 論理推論がベース
 - 現在では、「統計的」機械学習が主流（≡機械学習）
 - 遺伝子情報処理、自然言語処理、他、ビジネス分野での成功
 - つい最近ではクイズ王に勝利したIBMのワトソン
- 現在では、データ解析技術一般を指すほかの言葉とあまり変わらない
 - 統計／データマイニング／パターン認識など。
（多少のニュアンスの違いはあるが、基本的に好みの問題）

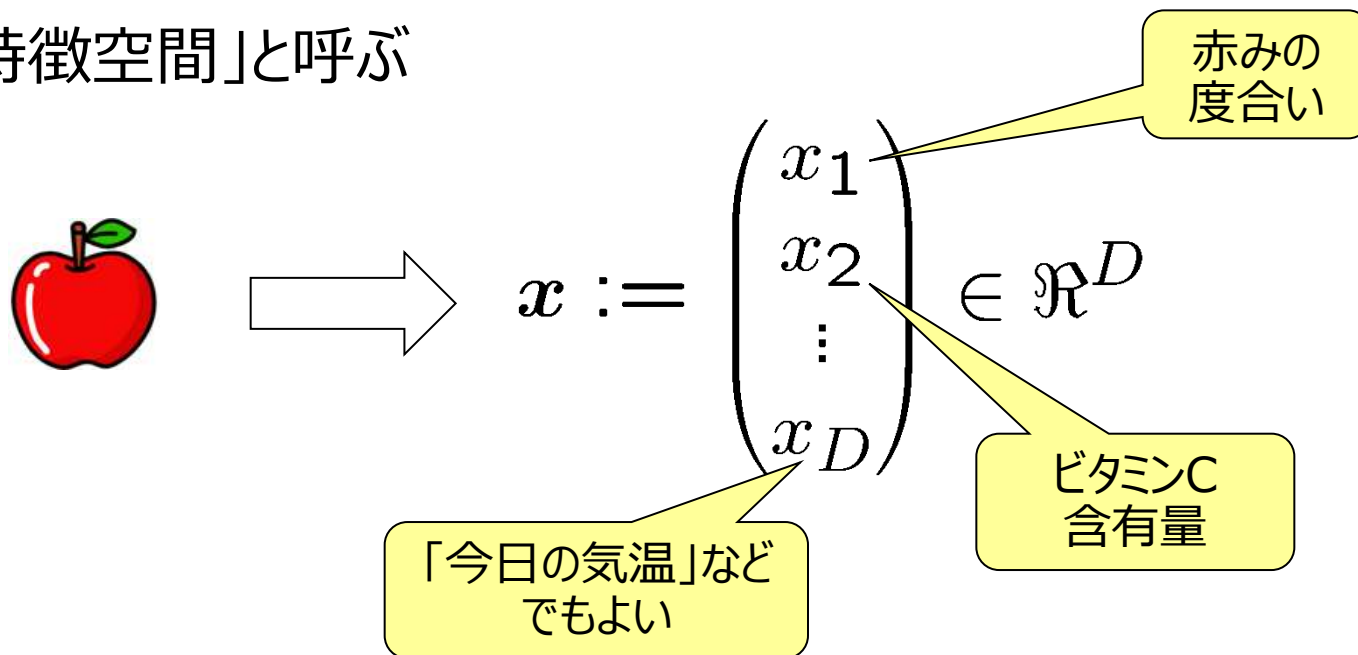
機械学習のモデル

機械学習を実現するためには、入力の数理的表現が必要です

- 学習機能を計算機上に実現するために、まず、学習問題を数理的にとらえる必要がある
- まずは、入力をどう数理的（＝計算機可読な形式）に表現するか？
 - 「やどかり」「ねこ」「りんご」は計算機上でどのように扱うか？
- 出力については比較的自明
 - 「あり」を+1、「なし」を-1と割り当てる

入力の表現： 通常、実数値ベクトル（特徴ベクトル）として表現します

- 入力を、その特徴量を列挙した D 次元の実数値ベクトル x として表現する
 - x を「特徴ベクトル」と呼ぶ
 - その領域を「特徴空間」と呼ぶ



- 特徴ベクトル x はどのようにデザインしたらよいか？
 - 完全にドメイン依存
 - 一般的解はなく、目的に合わせユーザーがデザインする


訓練データ：

教師付き学習では、入力ベクトルと出力の組が複数与えられます

- 訓練データは、 N 個の入力と出力のペア

$$\{ \underbrace{(x^{(1)}, y^{(1)})}_{\text{1つ目の入出力ペア}}, \underbrace{(x^{(2)}, y^{(2)})}_{\text{2つ目の入出力ペア}}, \dots, \underbrace{(x^{(N)}, y^{(N)})}_{\text{N個目の入出力ペア}} \}$$

- $x^{(i)}$: i 番目の事例の入力ベクトル
- $y^{(i)}$: i 番目の事例に対する正しい出力

( ならば +1, 違うなら -1)

- 教師付き学習：与えられた入力信号に対する、あるべき出力を教師信号として、入出力の関係を学習する

教師無し学習では、入力ベクトルのみが複数与えられます

- データは N 個の入力信号

$$(x^{(1)}, x^{(2)}, x^{(3)}, \dots, x^{(N)})$$

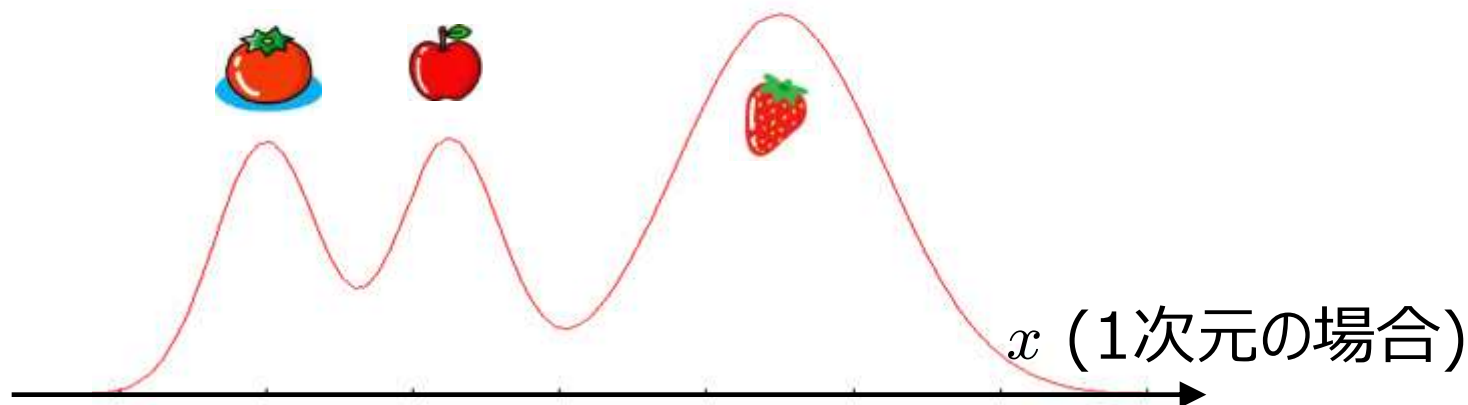
1つめの
データ

2つめの
データ

...

$$x := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix} \in \mathbb{R}^D$$

- 教師無し学習は、（大げさにいえば）明示的に指定されることなしに、「概念」を形成するプロセスを表している



線形モデル： もっともシンプルな教師つき学習の予測モデル

- 入力 $\mathbf{x} = (x_1, x_2, \dots, x_D)^\top$ に対し、
出力 $\{+1, -1\}$ を予測する分類モデル f を考える

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}) = \text{sign}(w_1 x_1 + w_2 x_2 + \dots + w_D x_D)$$

- $\text{sign}(\cdot)$ は引数が0以上なら+1、0未満なら-1を返す関数
- $\mathbf{w} = (w_1, w_2, \dots, w_D)^\top$ はモデルパラメータ
 - w_d は x_d の出力への貢献度を表す
 - $w_d > 0$ なら出力+1に貢献、 $w_d < 0$ なら出力-1に貢献

学習とは、訓練データからパラメータベクトル w を決定することです

- パラメータ w がきまるとモデル f がきまる

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x}) = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_Dx_D)$$

- 訓練データから w を決定するのが「学習」

$$\{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\} \xRightarrow{\text{学習}} \mathbf{w}$$

- 基本的には、訓練データの入出力を再現できるように w を調整する

- 出力が $y = +1$ のデータについては $\mathbf{w}^\top \mathbf{x} > 0$ となるように
- 出力が $y = -1$ のデータについては $\mathbf{w}^\top \mathbf{x} < 0$ となるように
- まとめてかくと $y \mathbf{w}^\top \mathbf{x} > 0$

教師つき学習の応用例

- 信用リスク評価
- テキスト分類
- 画像認識

教師付き学習の応用例：信用リスク評価

「この人にお金貸して、返ってくるんだろうか？」

- ある顧客に、融資を行ってよいか
 - 顧客 x を、さまざまな特徴を並べたベクトルで表現
 - 融資を行ってよいか y
 - 融資を行ってよい（返済してくれる） : +1
 - 融資してはいけない（貸し倒れる） : -1
 - マーケティングの文脈では、買ってくれる(+1),買ってくれない(-1)

$$x := \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix}$$

過去に延滞したことがあるか? (1/0)

リボ払い使用率

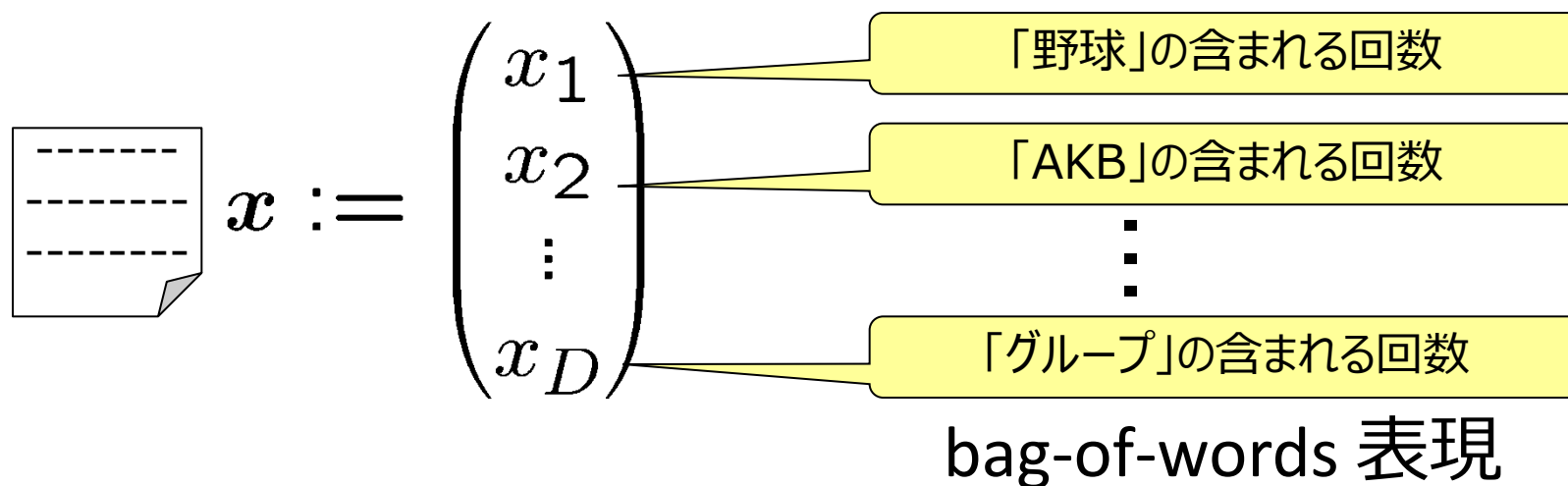
⋮

使用限度額

教師付き学習の応用例：テキスト分類

「あのタレントの不祥事、世間の評判はどうだろう？」

- 自然言語の文書が、あるカテゴリーに入るかどうか
 - 文書 x を、含まれる単語ベクトルで表現
 - (たとえば) ある事柄に好意的かどうか y
 - 好意的：+1
 - 否定的：-1
 - トピック y ：「スポーツ」「政治」「経済」… (多クラス分類)

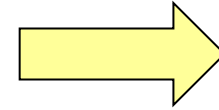


教師付き学習の応用例：画像認識、脳波解析、...

「これ、何て書いてあるの？」「いま何考えてる？」

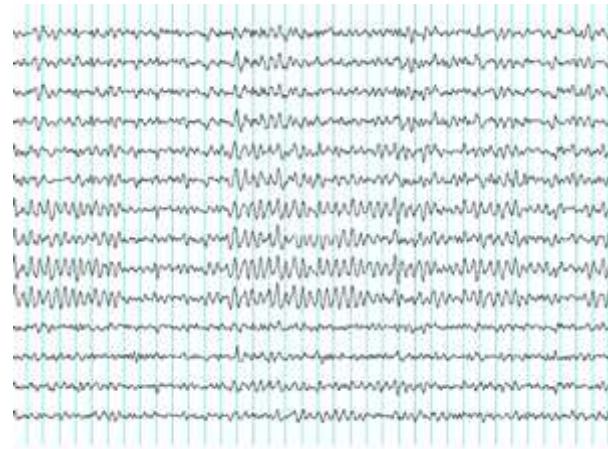
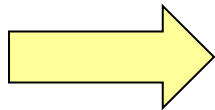
■ 手書き文字認識

7210414959
0690159734
9665407401
3134727121
1742351244

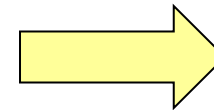


ある文字か(+1)否か(-1)
どの文字か？ {"0","1","2",...}

■ BCI (Brain Computer Interface)



どちらを思い浮かべている？



右(+1)？左(-1)？

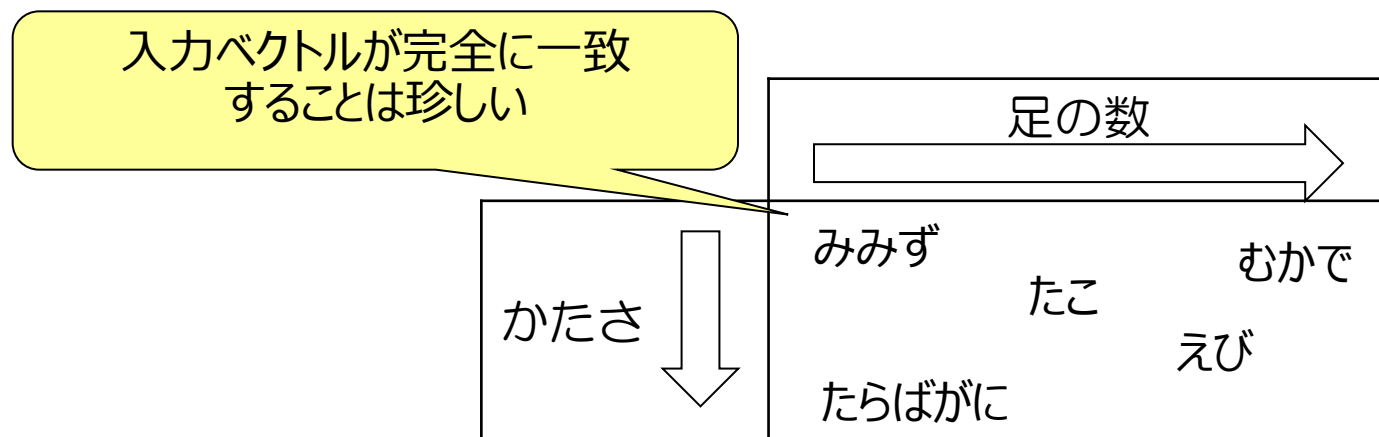
■ ほか、顔画像認識や、動画認識

教師なし学習では入力データを K 個のグループに分けますが
データは完全に一致することは珍しいので工夫が必要です

- N 個の入力ベクトル $\{x^{(1)}, x^{(2)}, \dots, x^{(N)}\}$ を K 個のグループに分ける
- 先の例では完全に一致するデータがあったのでグループ分けは自明

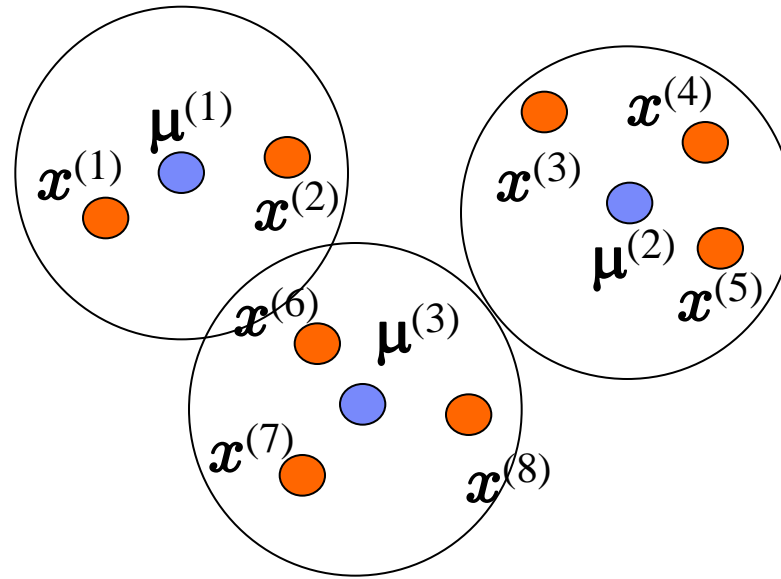
		足の数	
		8本	10本
かたさ	やわらかい	くも たこ	いか
	かたい	たらばがに やどかり	毛がに えび

- 通常はそうではないので、グループ分けは自明でない



教師なし学習の典型的アプローチのひとつは、グループごとの代表点を考え、代表点への距離でグループ所属をはかることです

- K ($=3$) 個のグループそれぞれの代表点 $\{\mu^{(1)}, \mu^{(2)}, \mu^{(3)}\}$ を考える
- 代表点に近い入力データは、そのグループに属するとする

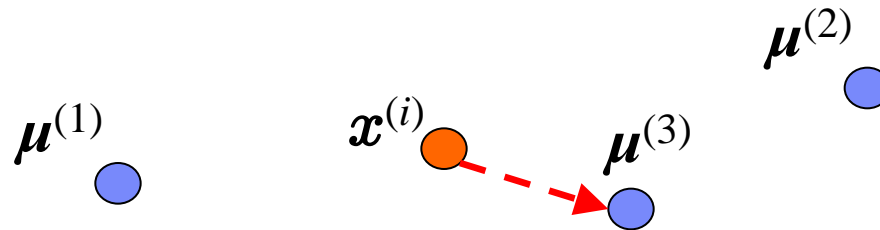


- 代表点への「近さ」（距離）はどう定義するか？
 - 距離関数 $d(\mu^{(k)}, x^{(i)})$ を目的によって適切に定義する
 - たとえばユークリッド距離 $d(\mu^{(k)}, x^{(i)}) = \|\mu^{(k)} - x^{(i)}\|_2^2$

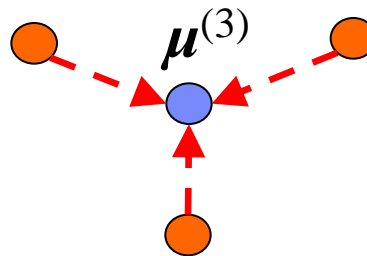
K-meansアルゴリズム：グループ割り当てと代表点推定を交互に行うアルゴリズムです

- 以下のステップを収束するまで繰り返す

1. 各データ $x^{(i)}$ を、最寄の代表点 $\mu^{(k)}$ に割り当てる



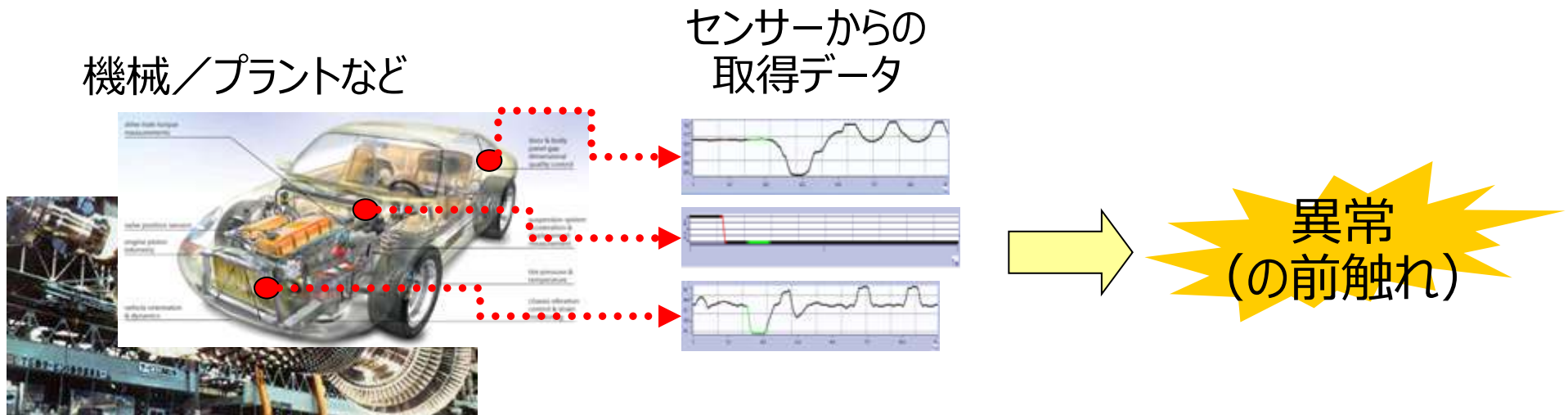
2. 各代表点に所属したデータの平均として代表点を新たに求める
(ユークリッド距離の場合)



教師なし学習の応用例：異常検知

「ちょっと出かけてくるけど、ヤバそうだったら教えて」

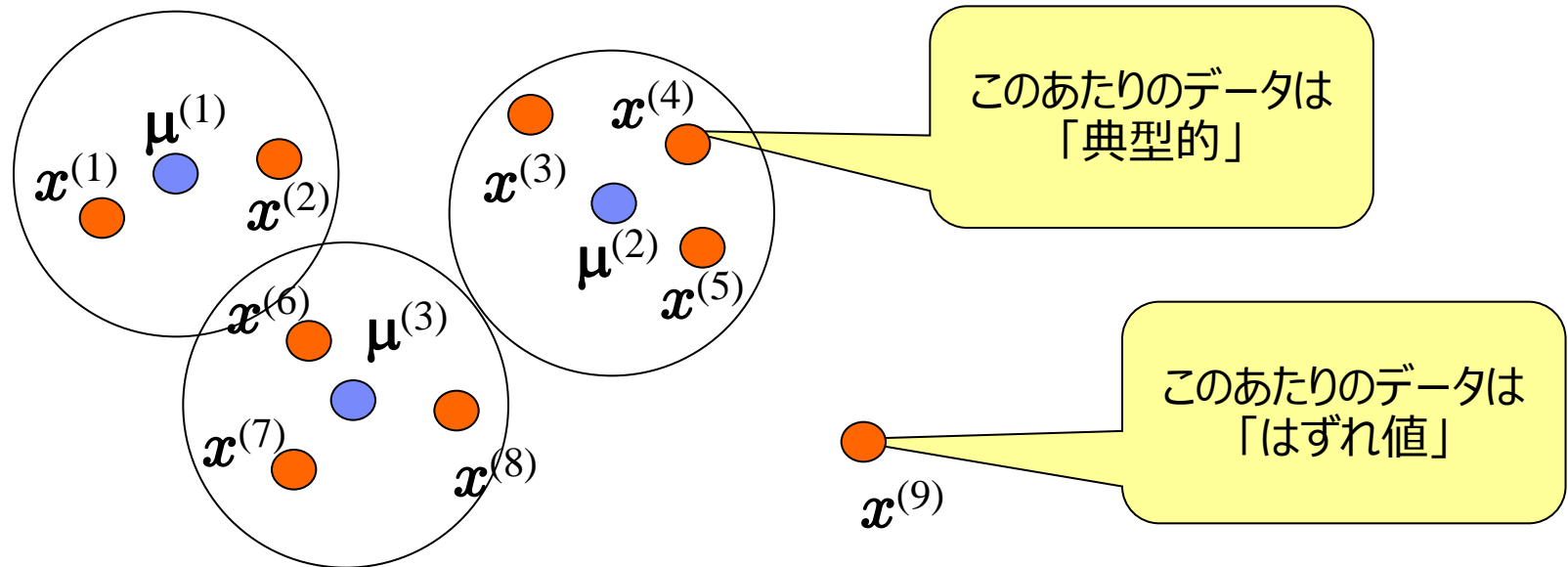
- 機械システム／コンピュータシステムの異常を、なるべく早く検知したい
 - 早い段階で検出できれば、それだけコスト減
- システムに仕込まれたセンサーからの取得データを分析する
 - システムの異常／変化、不正な操作により、システムが通常とは異なった振る舞いをするようになる
 - 計測機器の異常によって、通常とは異なった計測値が得られるようになる



教師なし学習の応用例：異常検知

グループに属さないデータ＝異常と考えます

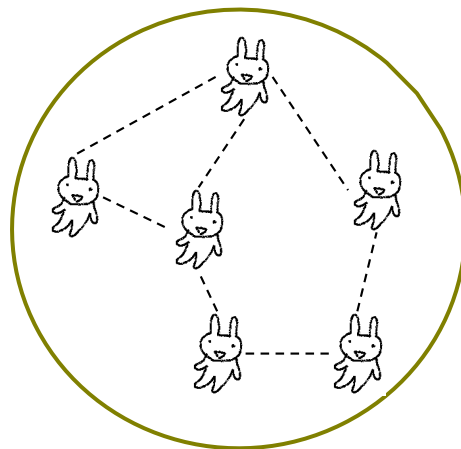
- システムの状態をベクトル x で表現し、教師無し学習によるグループ分けを行う
 - コンピュータ間の通信量、各コマンドやメッセージ頻度
 - 各センサーの計測値の平均、分散、センサー同士の相関
- 代表点から遠い x は「めったに起こらない状態」＝システム異常、不正操作、計測機器故障などの可能性がある



まとめ

- 機械学習はデータ解析の手法である
- 教師つき学習：予測
 - 入出力の関係を導き、出力未知の入力に対し予測を行う
- 教師なし学習：発見
 - 入力に潜むパターン（グループ）を発見する
 - 異常検知は重要な応用
- データは実数値ベクトルとして表現される
 - その表現がきわめて重要だが、それは機械学習の枠の外

ネットワークの機械学習



近年、機械学習の対象が、 個々のデータから、データ間の関係へと移行しつつあります

- 従来：「個々のデータを対象とした解析」

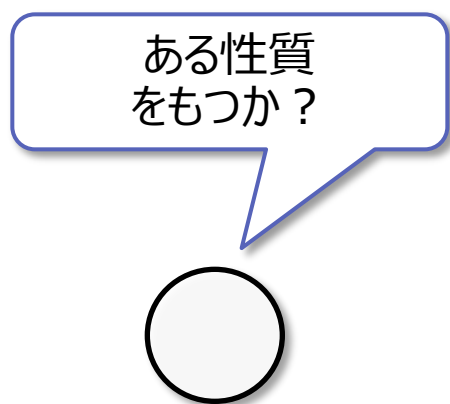


近年：「データの間の関係の解析」

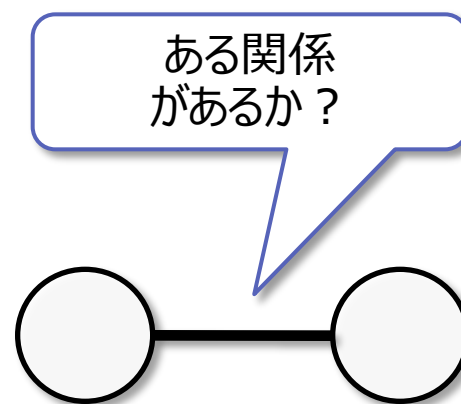
- 様々な領域において「関係の分析」は盛んになりつつある
 - ソーシャルネットワーク分析：人間関係
 - オンラインショッピング：顧客と商品の間の関係
- データ間の関係に注目することで、
個々のデータに注目しているだけでは見えない性質が見えてくることもある
 - コンピュータネットワーク上のプロセス依存関係から異常を予測
 - 複数の脳波時系列の相関関係から思考を読みとる

関係データとは ものごとの関係を表現したデータ です

- 通常 of データ解析では、ひとつのデータについて成り立つ性質を推論する
- 関係データとは： データの組についてのデータ
- 関係の成立や、関係のもつ性質についての推論を行う



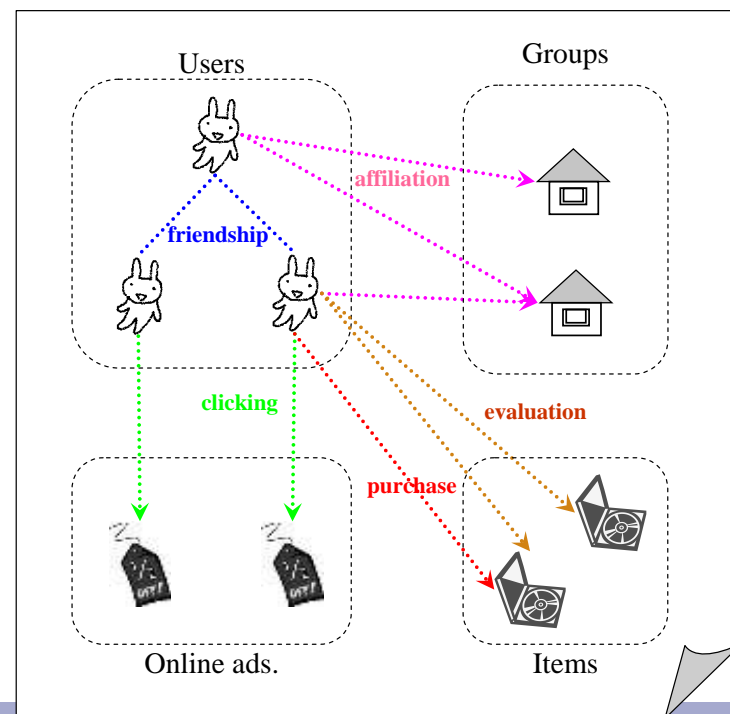
単一データ
についての予測



2つのデータの関係
についての予測

関係データの例：マーケティング、Web、バイオ、…

- オンラインマーケティング
 - 顧客と商品との間の関係（購買、評価）
- ソーシャルネットワーク
 - SNS内の人間関係 (facebook, twitter, mixi, ...)
 - 企業間取引
- 生体ネットワーク
 - タンパク質相互作用ネットワーク
 - 化合物-タンパク質相互作用



関係データを用いたタスク：予測と発見

- 予測
 - 推薦システム（協調フィルタリング）
 - 顧客と商品との間の関係（購買、評価）を予測
 - 例：Netflix challenge
 - SNSの友人推薦
 - 新規薬剤候補の探索
- 発見
 - 顧客セグメンテーションの発見
 - 協調するタンパク質グループの発見
 - 例外の発見



関係データの表現：2項関係はグラフや行列などで表現できます

- 通常、データは表形式で与えられる

顧客番号	顧客氏名	年齢	性別	住所	...
0001	〇〇	40代	男性	東京都	...
0002	××	30代	女性	大阪府	...

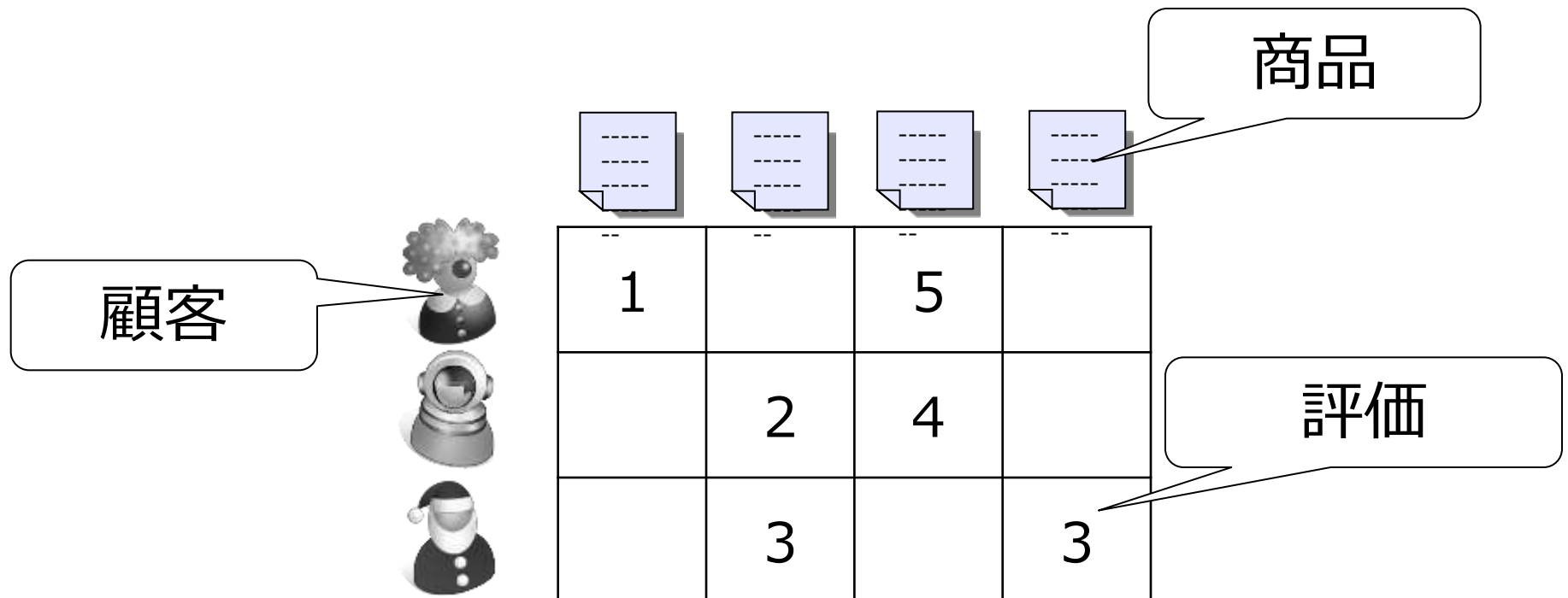
- 関係データはこれらの間の関係を表す



- 数学的な表現
 - 行列／多次元配列
 - グラフ／ハイパーグラフ

2項関係の集合は行列として表現できます

- 2項関係は行列として表現できる
 - 行と列がデータの集合に対応
 - 各要素がデータ間の関係を表す
- グラフ（重みつき）の隣接行列としてもみることができる

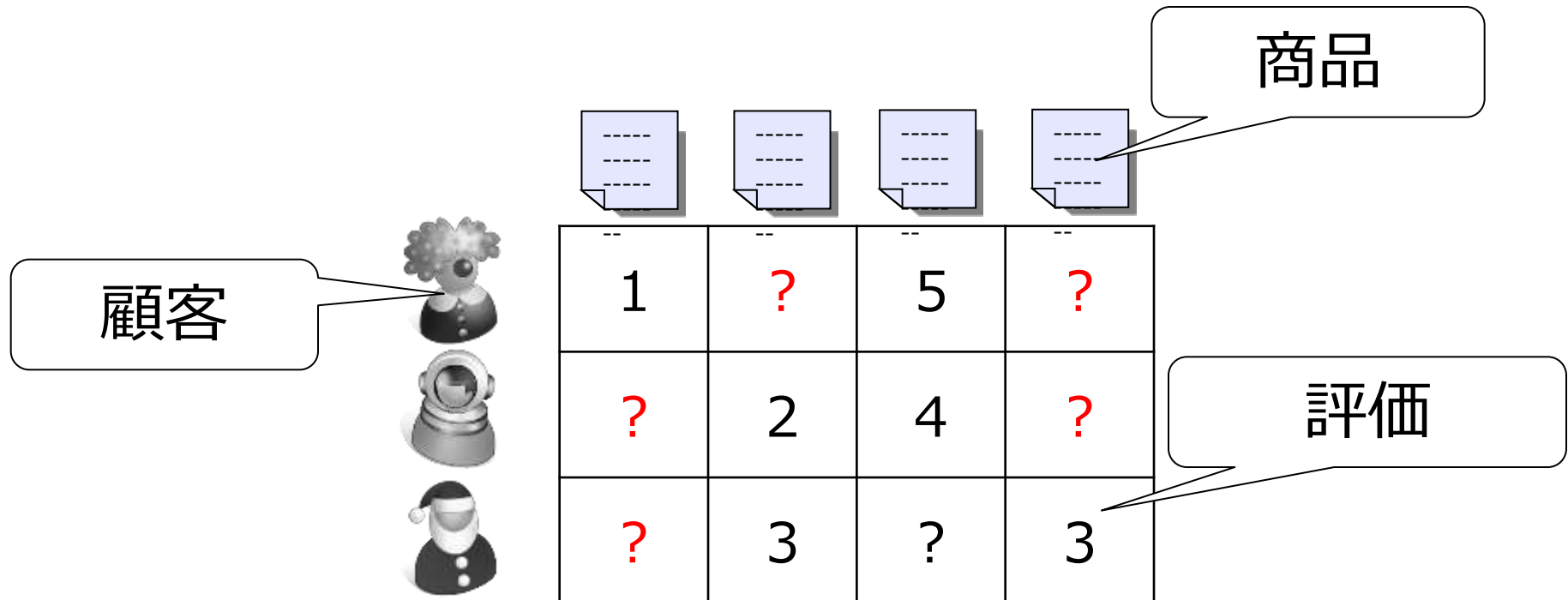


行列データの解析手法

- 行列の補完問題
- 協調フィルタリングの初等的手法：GroupLens
- 行列の低ランク分解

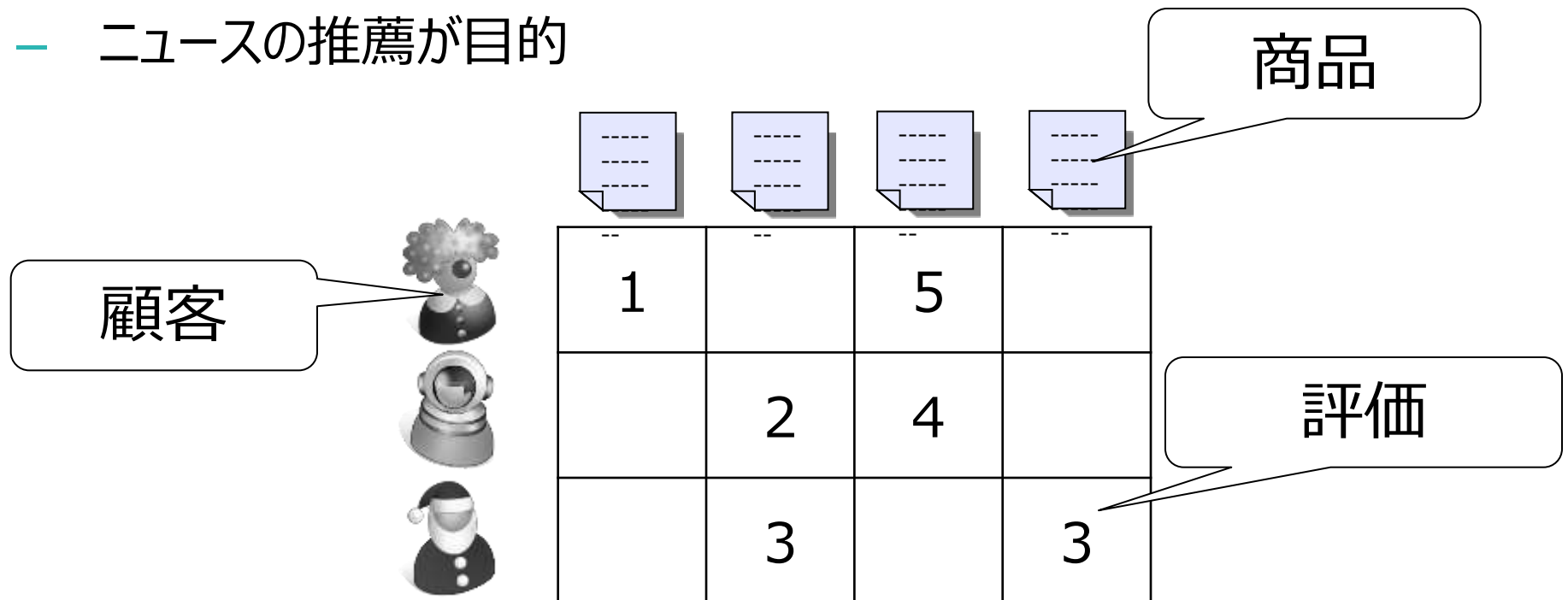
行列の補完問題は、行列の観測部分をもとに、未知の部分进行予測する問題です

- 見えている部分をもとに、見えていない部分进行予測する
- 推薦システムにおける評価予測



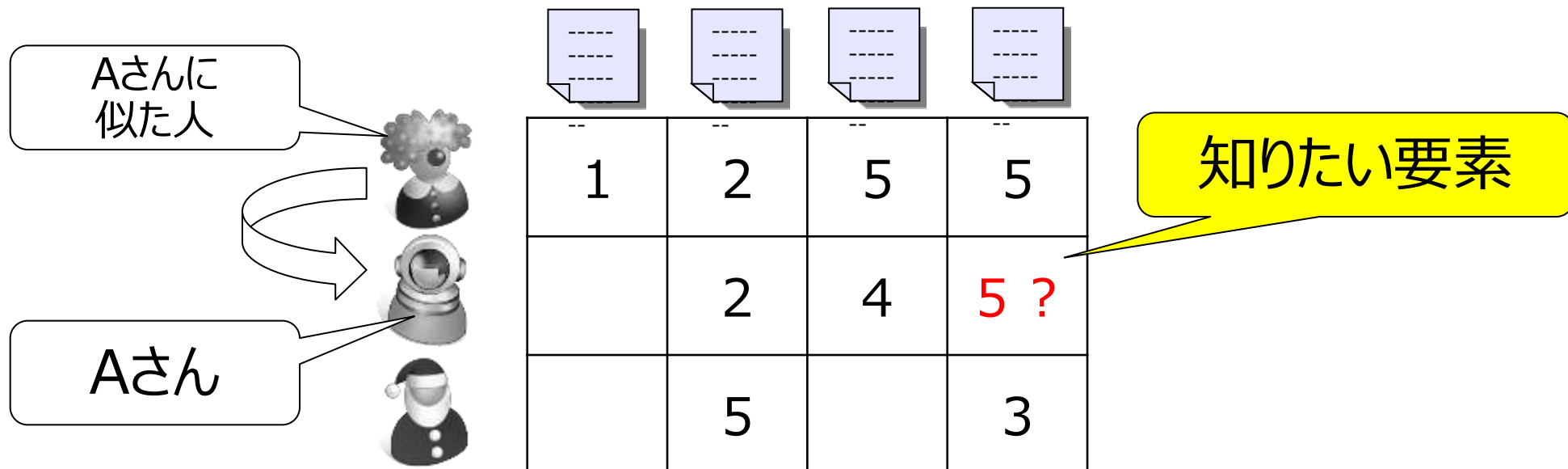
GroupLens：協調フィルタリングの初等的手法

- 推薦システム（協調フィルタリング）は、顧客と商品との間の関係（購買、評価）を予測する
- 値が分かっている部分を手掛かりに、未知の部分の予測したい
- GroupLens：初期の予測アルゴリズム
 - ニュースの推薦が目的



GroupLensでは、ある顧客の評価を、 似た顧客の評価を持ってきて予測します

- 予測したい顧客と似た顧客を集め、類似顧客の評価を用いて予測を行う
 - Aさんの未知要素を予測したいとする
 - Aさんと良く似た評価を行っている別の顧客を集めてきて、彼らの評価を用いて予測する

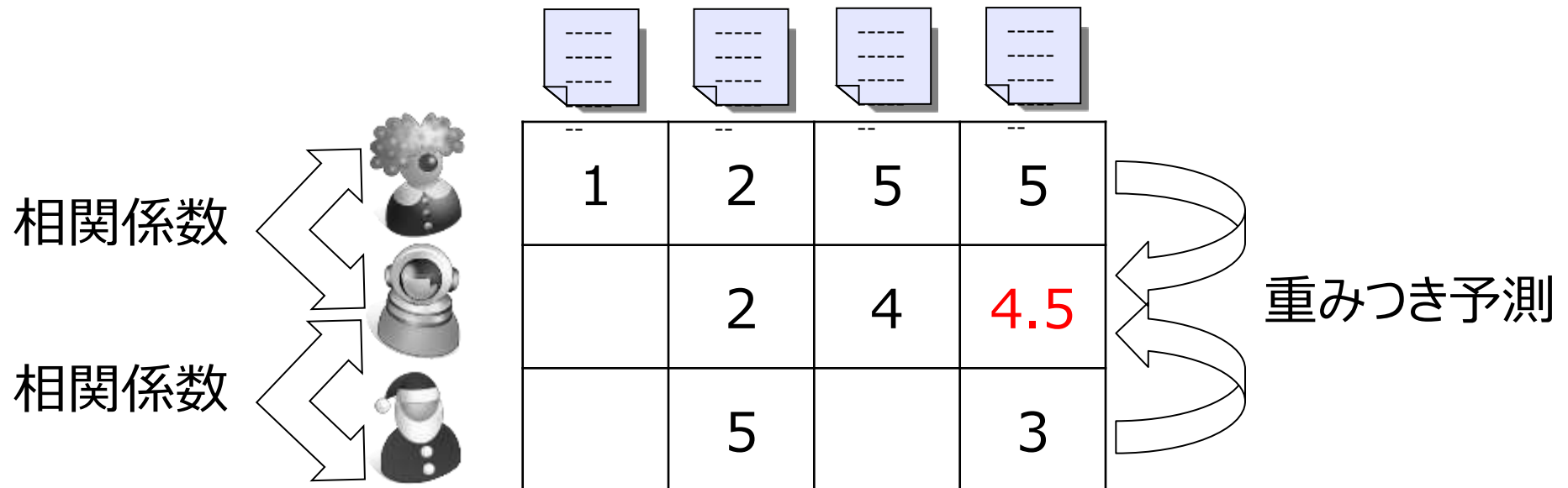


「似ている」の定義は 評価値の相関係数で測り、 相関係数で重みづけして予測します

- 2人の顧客の類似度を（共に評価値が観測されている部分の）相関係数で測る
- **相関係数**で重みづけし予測を行う

$$y_{i,j} = y_i + \sum_{k \neq i} \rho_{i,k} (y_{k,j} - y_k) / \sum_{k \neq i} \rho_{i,k}$$

- 同様に、商品間の類似度を用いることも可能

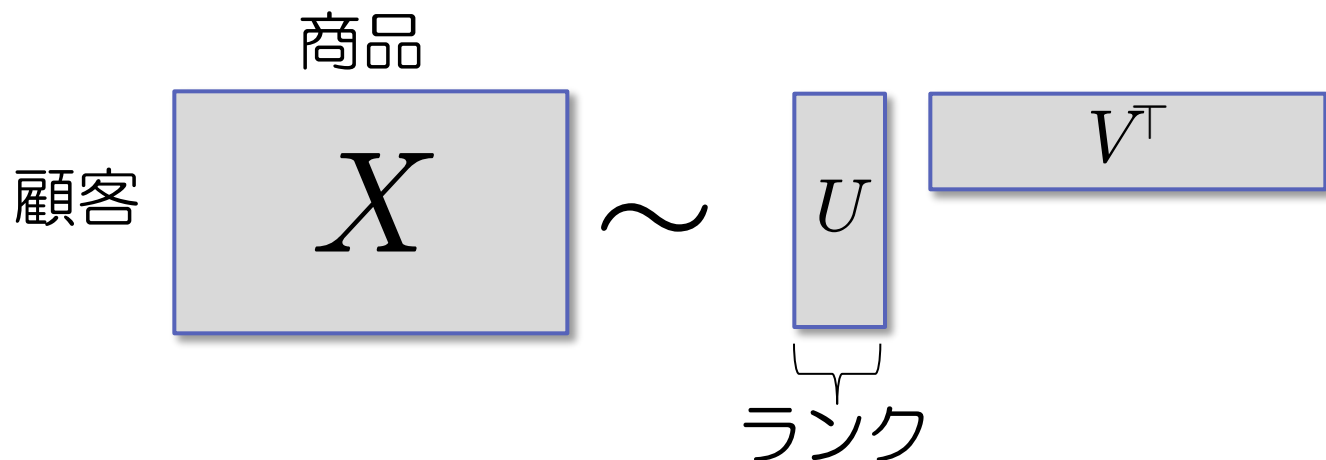


協調フィルタリングの初等的手法は 行列の低ランク性を暗に仮定しています

- 行列の各行が、別の行の（相関係数で重み付けた）線形和によって表せるとしている
 - 線形従属
- 対象となる行列のランクがフルランクではない（ \Rightarrow 低い）ことを暗に仮定した方法ということになる
- 低ランク性の仮定は行列の穴埋めに有効であろう
 - データよりもパラメータが多い状況では、なんらかの事前知識を用いて解に制約を設ける必要がある
 - 低ランク性の仮定は、実質パラメータ数を減らす

行列の低ランク性を仮定することで分解を行います

- 低ランク性の仮定：行列が2つの（薄い）行列の積で書ける



$$\text{minimize}_{\mathbf{Y}} \|\mathbf{X} - \mathbf{Y}\|_F^2 \quad \text{s.t.} \quad \text{rank}(\mathbf{Y}) \leq k$$

- 実効パラメータ数が減っている
- U (V) の各行：顧客（商品）の特徴を捉えた低次元の潜在空間にデータを配置
 - この空間で近いものが似た顧客（商品）：グループ構造

行列分解には特異値分解がよく用いられます

- 行列分解 $X = UV^T$ の仮定だけでは、解の不定性があるので、制約を入れる
- 特異値分解

Diagram illustrating the SVD decomposition of matrix X :

$$X \sim U D V^T$$

Where:

- U is the left singular matrix.
- D is the diagonal matrix of singular values (特異値).
- V^T is the transpose of the right singular matrix.

- 制約 : $U^\top U = I \quad V^\top V = I$
- $X^\top X$ の固有値問題になる
 - 固有値を大きい方から k 個とる

欠損値がある場合には特異値分解は使えません

- ランク制約をもった最適化問題は凸最適化問題ではない

- ランク k 以下の行列は凸集合ではない

- 目的関数 = 復元誤差（凸関数） + ランク制約

$$\text{minimize}_Y \|X - Y\|_F^2 \text{ s.t. } \text{rank}(Y) \leq k$$

もしくは分解を UV^\top と明示的におくと誤差項が非凸になってしまう

$$\text{minimize}_Y \|X - UV^\top\|_F^2$$

- 全データが観測されている場合には、固有値問題としてたまたま解ける
- 欠損値がある場合には困る

欠損値がある場合には、EM的アルゴリズムが用いられる

- ひとつの方法としては気にせず、勾配法などで適当に解く
 - データが大きいときにはこちら
- EMアルゴリズム：未観測部分には暫定的な推定値をあてはめ、完全観測として問題を解く
 1. 未観測部分を適当に初期化（平均など）
 2. 低ランク行列分解を適用
 3. 復元した値で未観測部分の値を置き換えるステップ 2～3を収束まで繰り返す

凸最適化としての定式化：トレースノルム正則化

- 行列のランク制約は凸集合ではないので、凸集合であり、ランク制約のよい近似となるような制約がほしい
- 行列の特異値の和を用いる

特異値

$$\|\mathbf{Y}\|_* = \sigma_1(\mathbf{Y}) + \sigma_2(\mathbf{Y}) + \dots$$

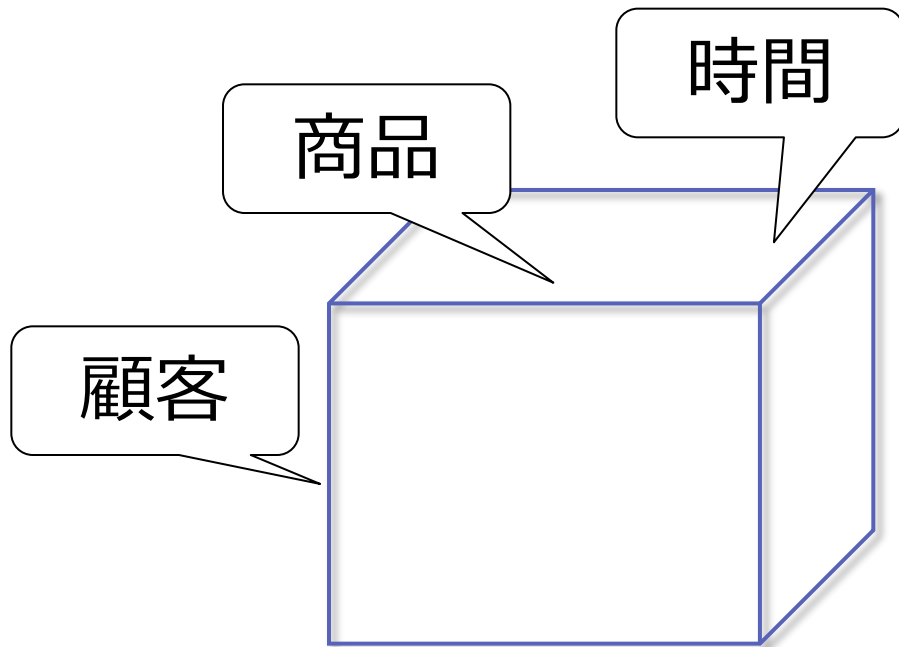
- 特異値の集合 $(\sigma_1(\mathbf{Y}), \sigma_2(\mathbf{Y}), \dots)$ に対する L_1 ノルム制約と等価であるため、疎になる \Rightarrow ランクが落ちる
- 一方、ランクは、非零の特異値の個数
- 目的関数 = 観測部分の復元誤差 + トレースノルム制約

$$\text{minimize}_{\mathbf{Y}} \|\mathbf{O}^*(\mathbf{X} - \mathbf{Y})\|_F^2 \quad \text{s.t.} \quad \|\mathbf{Y}\|_* \leq c$$

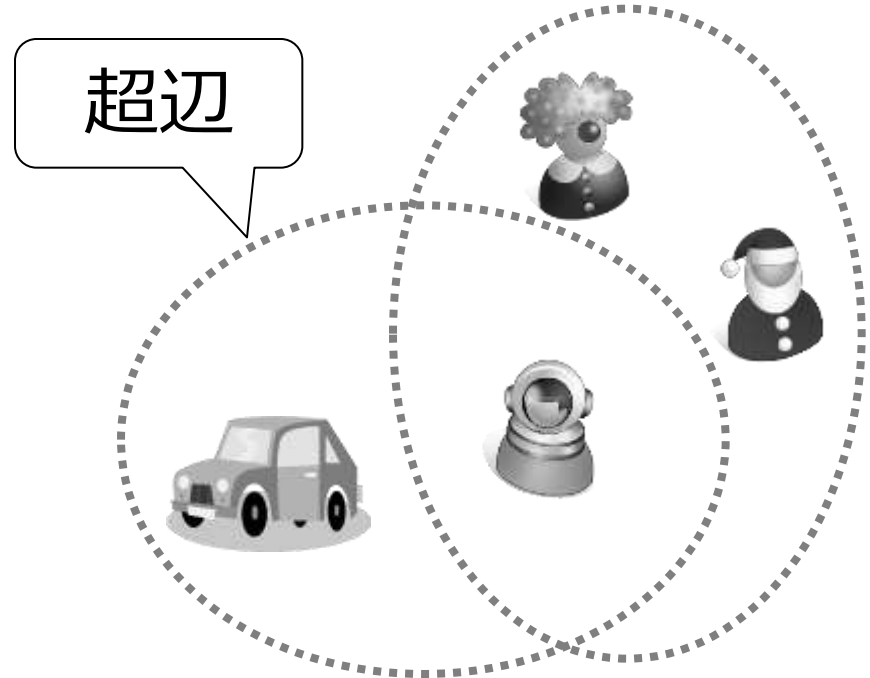
- 最適化は勾配法と特異値分解の組み合わせ

多項関係の集合は 多次元配列やハイパーグラフとして表現できます

- 多項関係の集合は多次元配列として表現できる
- ハイパーグラフとしても表現可能
 - こちらのほうがより一般的：関係に参加するデータの数が可変



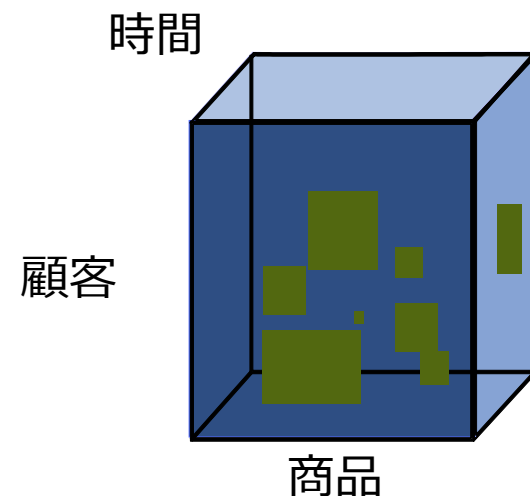
多次元配列



ハイパーグラフ

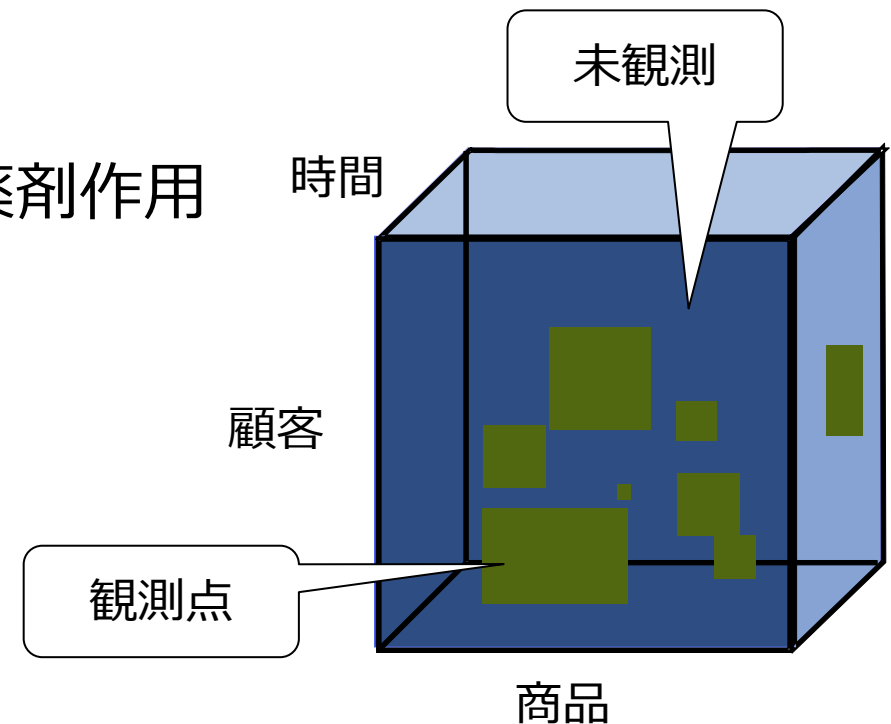
テンソル（多次元配列）は行列よりも一般的な関係の表現です

- テンソルはさまざまなデータ間の複雑な関係を表すことができる
 - (顧客, 商品, 時間)の関係は「ジョンが2011/09/01にIPadを買った」ことを表現できる
 - (顧客, 行動, 商品)の関係は「アリスがハリーポッター最新刊についてレビューを書いた」ことを表現できる
- テンソルは動的で異種混合的な関係を表すことができる：
 - 関係の時間変化
 - 例：顧客の興味の時間的うつりかわり
 - 関係の関係
 - 「購買」と「商品レビュー」には正の相関がある



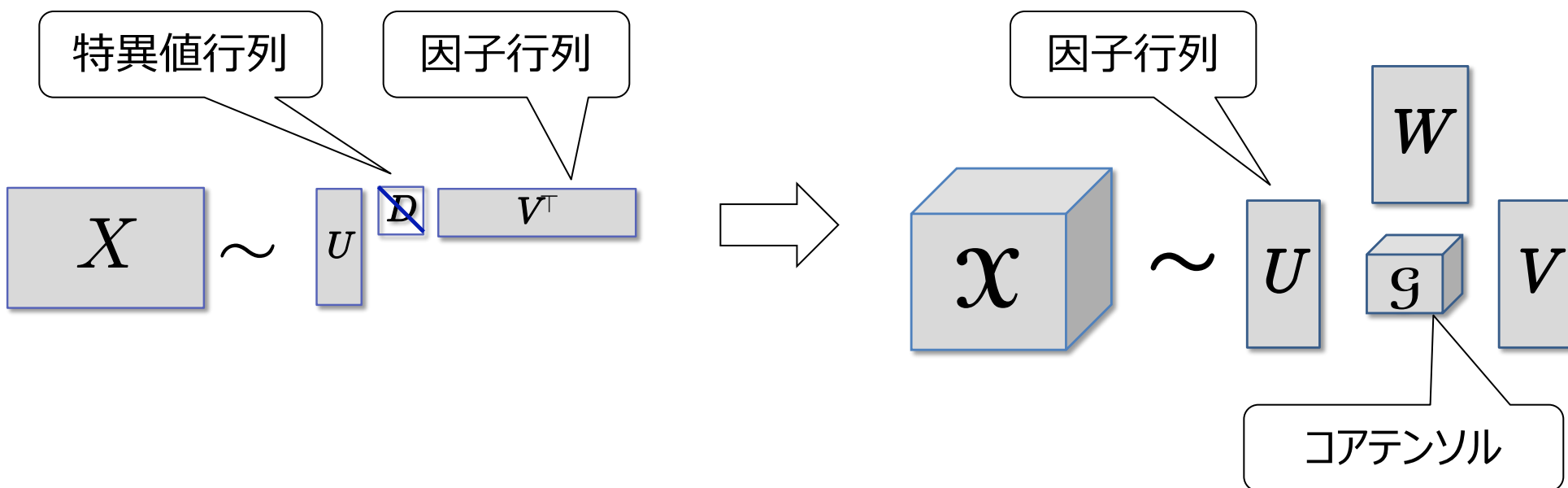
テンソル補完問題： より高次の関係の予測問題を扱います

- テンソル補完問題:
テンソルが部分的に観測されたとき、のこりの部分を予測する問題
 - テンソル分析の典型的問題
 - マーケティング、社会科学、生物学など幅広い応用がある
 - オンラインショッピングでの商品推薦
 - SNSでの友人推薦
 - タンパク質相互作用、タンパク質-薬剤作用
- 予測精度の向上は：
 - 売上増加
 - ユーザ満足度
 - 新たな科学的知見



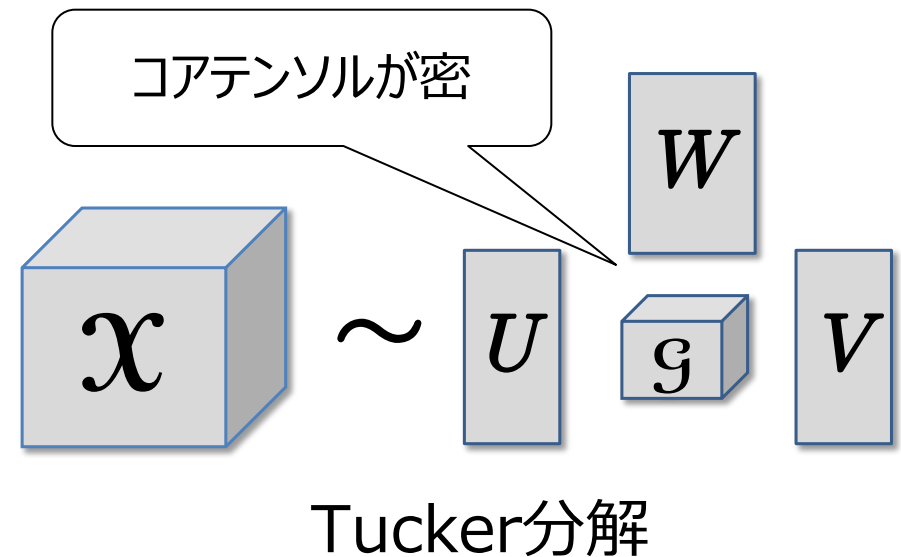
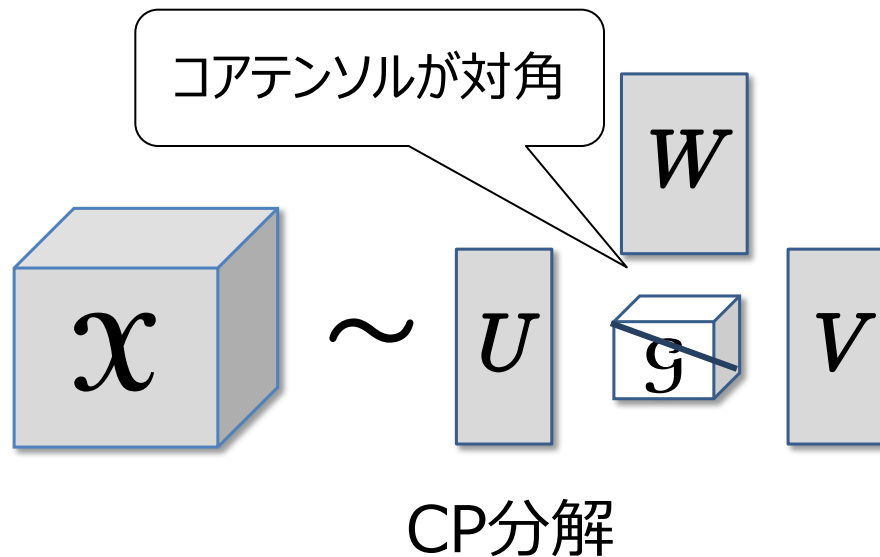
行列分解は多次元配列（テンソル）の低ランク分解に一般化されます

- 行列の低ランク分解の多次元配列への一般化
 - ちいさな（コア）テンソルと因子行列に分解する
- 近年、機械学習やデータマイニングで盛んに用いられている



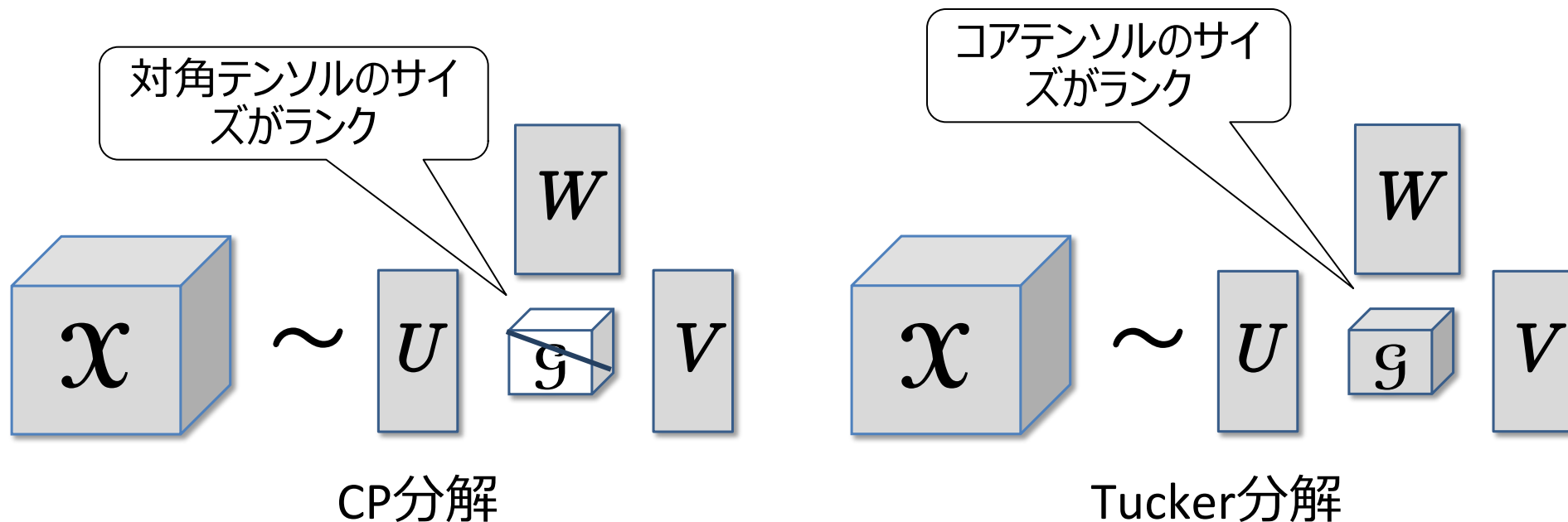
テンソル分解のタイプ：CP分解とTucker分解

- よく用いられるのがCP分解とTucker分解
- CP分解：特異値分解の自然な拡張（コアテンソルが対角；正方）
- Tucker分解：よりコンパクトな表現（みっちりコア；各モードの次数が異なる）



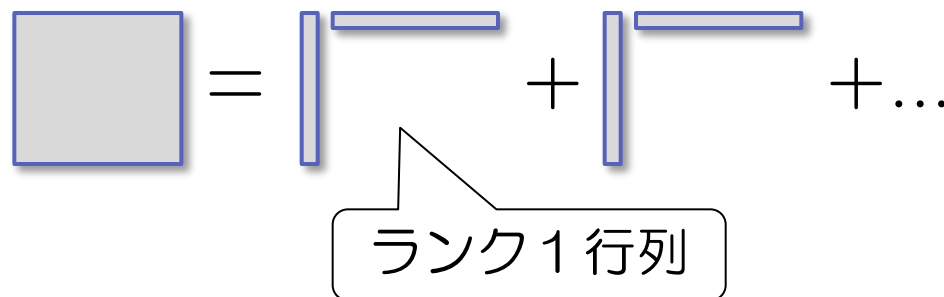
テンソルのランクは分解のタイプによって決まります

- 行列のランクはSVDの非零の特異値の数で決まった
- テンソル分解の場合には分解のタイプによって決まる
 - CP分解、Tucker分解それぞれでランクの定義がある

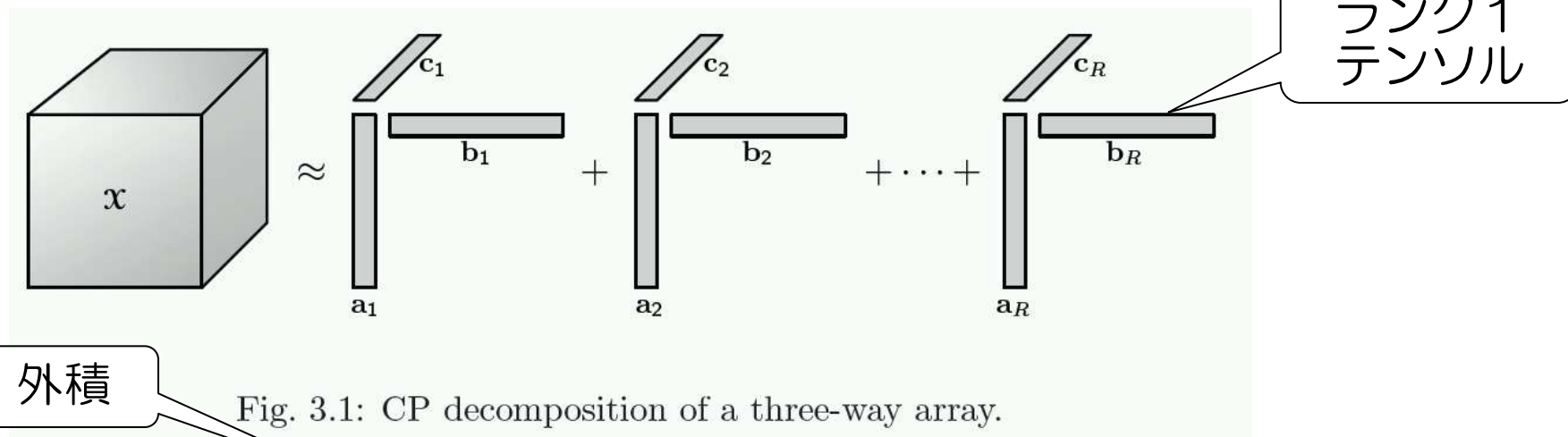


CP分解はランク 1 テンソルの和として定義されます

- 行列はランク1行列の和



- CP分解はランク1テンソルの和



外積

Fig. 3.1: CP decomposition of a three-way array.

$$\mathcal{X} \sim \sum_r \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r$$

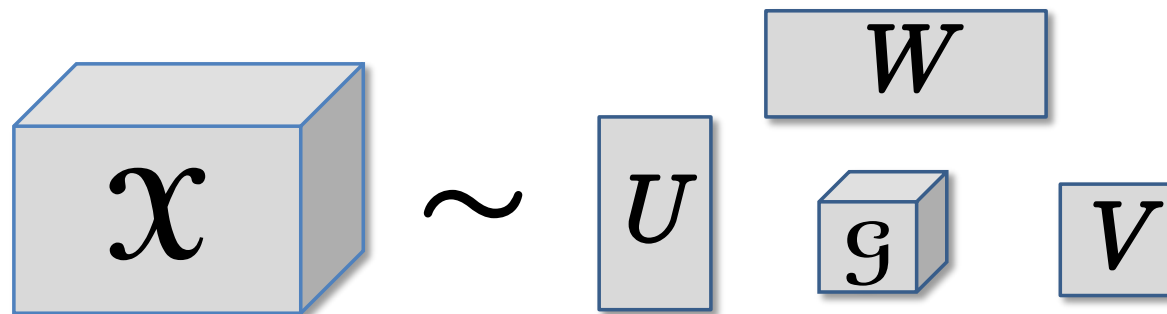
$$x_{ijk} = \sum_r \lambda_r a_{ri} b_{rj} c_{rk}$$

* The figures are taken from T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Tucker分解は小さいテンソルと行列によって定義されます

- Tucker分解はコアテンソルと、因子行列によって定義される
 - モード積を使って定義される

$$\mathcal{X} \sim \mathcal{G} \times_1 U \times_2 V \times_3 W \quad (x_{ijk} = \sum_{pqr} g_{pqr} u_{ip} v_{iq} w_{ir})$$



- 多くの場合因子行列の列ベクトルが正規直交であると仮定
- CP分解はコアテンソルが対角であるようなTuckerの特殊ケース



ソフトウェア：Matlabでの実装が公開されています

- Matlabのツールボックスとして公開されている
 - Tensor Toolbox
 - N -way Toolbox

応用事例

- ソーシャルネットワーク分析 (人×人×時間)
- Webリンク解析 (Webページ×Webページ×アンカーテキスト)
- タグ推薦 (人×Webページ×タグ)
- 画像認識 (画像×人×向き×明るさ×…)
- 脳波解析 (場所×場所×時間)

タグ推薦タスクへの応用例 (Rendle *et al.* (2010))

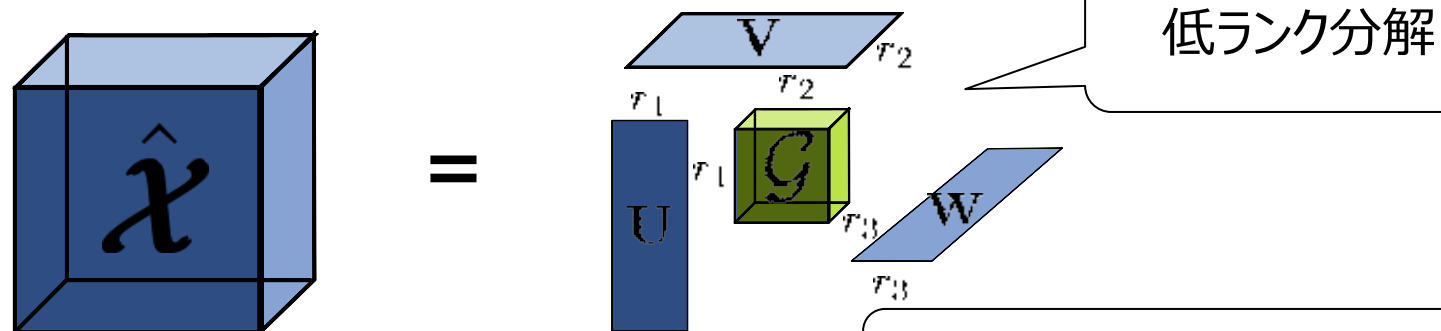
Personalized Tag Recommendation



Task: Recommend a user a (personalized) list of tags for a specific item.

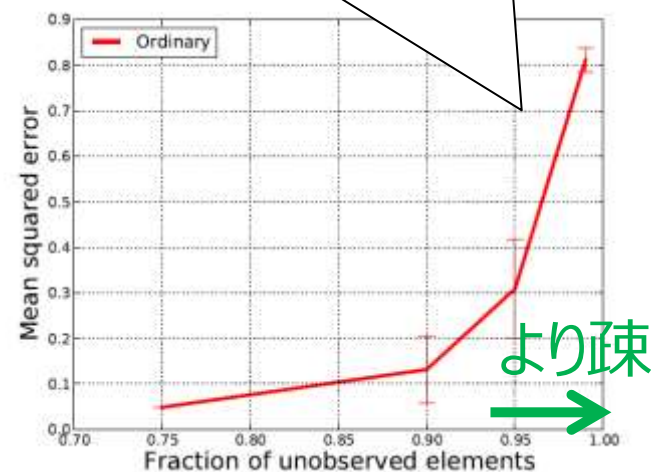
高次関係の予測ではデータの疎性が課題です

- テンソルの分析では低ランク性の仮定を行うのが通常
 - Tucker分解など



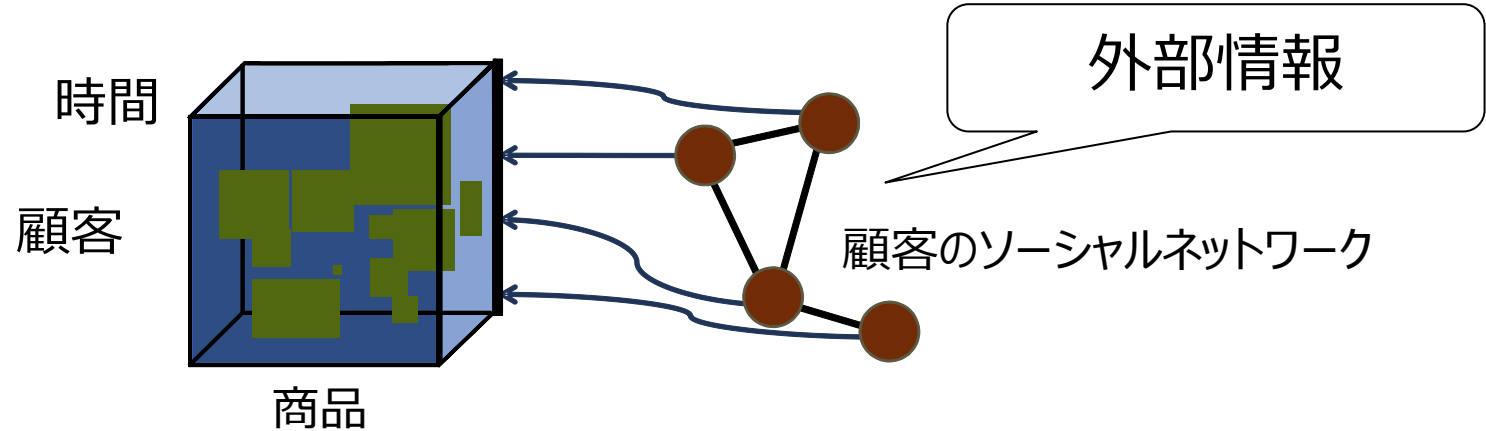
- 課題：疎なデータの予測
 - 観測部分が少ないときに、予測精度が著しく悪化してしまう
 - 可能な関係の数は組み合わせ的に増加する
- 低ランクの仮定だけでは足りない！

予測精度の悪化



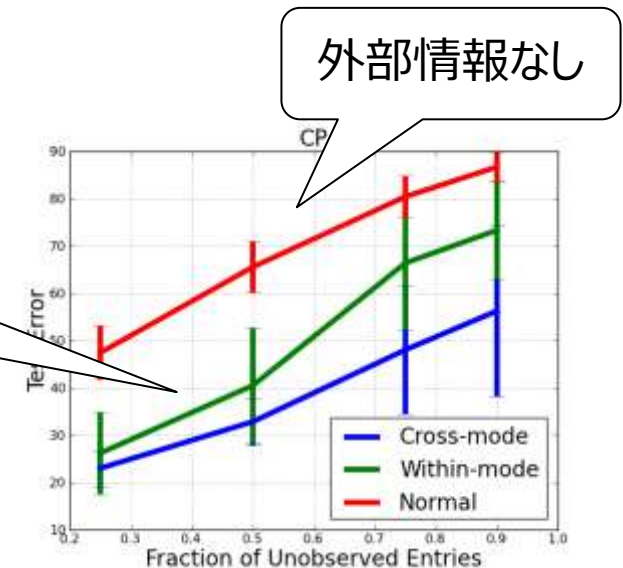
疎性への取り組み：低ランク性の仮定だけでは足りないので、併せて外部情報を利用します

- 実際には、予測したい関係データのほかに、データ間の関係が外部情報として利用可能な場合が多い（例：友人同士の振る舞いは似ている）



- データ間の関係を用いると予測精度が改善する

外部情報の利用が
精度を大きく向上させる



Narita, Hayashi, Tomioka & Kashima:
Tensor Factorization Using Auxiliary Information
In ECML PKDD 2011 (won the Best Student Paper Award)

ネットワーク正則化によって、外部情報を取り込み、予測の助けとすることで、予測精度が向上します

- ネットワーク正則化： 外部情報として与えられる関係情報を推論のガイドに用いる（最適化問題の目的関数に導入）
 - 隣り合ったデータが振る舞いをするように働く

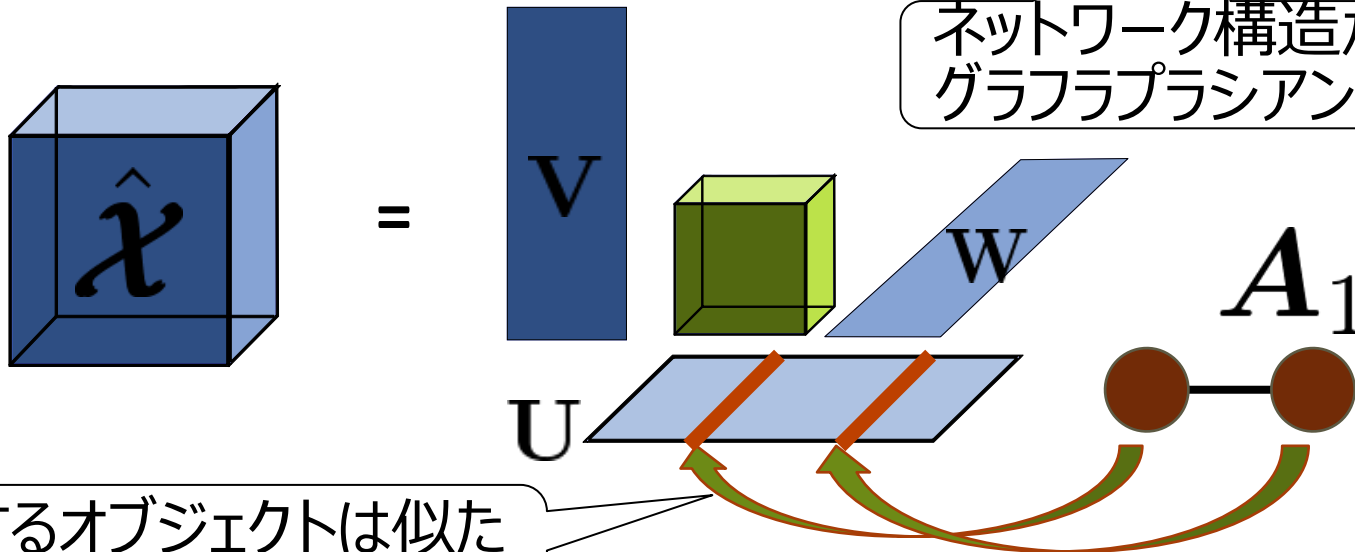
近似誤差の項

$$\min \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2$$

ネットワーク正則化項

$$\text{tr} (U^\top L_1 U)$$

ネットワーク構造から導かれる
グラフラプラシアン行列



「隣接するオブジェクトは似た振る舞いをするべき」

- データ解析の興味の対象は、単一のデータから、データ間の関係へ
- データ間の関係は、行列やテンソルで表現される
- 行列、テンソルともに低ランク分解を中心とした分析手法が用いられる

クラウドソーシングと機械学習



クラウドソーシングとヒューマンコンピューテーション： 機械学習研究とのかかわり

- 機械学習におけるクラウドソーシング利用
- ヒューマンコンピューテーションにおけるクラウドソーシング利用

クラウドソーシング

クラウドソーシング：不特定多数に仕事を依頼するしくみ

■ クラウドソーシングとは：

「（インターネットを通じて）不特定多数の人に仕事を依頼すること、もしくはその仕組み」一般を指す言葉

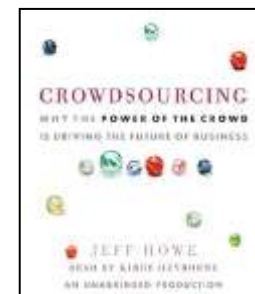
- 米Wired 誌の寄稿編集者ジェフ・ハウ氏によって命名

― クラウドソーシングでは（時には匿名の）不特定多数の相手に仕事を依頼

- 業務の一部を外部に委託する「アウトソーシング」を振ったもの
- アウトソーシングの委託先は素性の知れた特定の相手

■ メリット：

- ― 社員を抱えるよりも安価
- ― 必要なスキルを必要に応じて発見、調達可能
- ― 「群衆の叡智」の利用



Howe, J. / Crowdsourcing (2004)

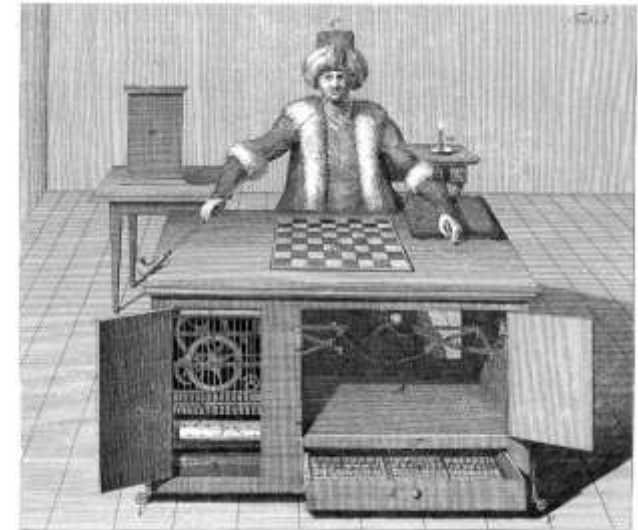
クラウドソーシングの事例：

- 初期の事例：
 - 米P&G 社が技術的な課題の解決を公募
 - Wikipedia：不特定多数の人間がその編集に関わる
- クラウドソーシングをサポートする様々なサービスの出現：
 - 米InnoCentive 社：研究開発の委託を仲介するサービスを提供
 - Amazon Mechanical Turk (AMT)：Amazon の提供するクラウドソーシング市場
 - 計算機から呼び出し可能なAPIを提供
 - CloudCrowd：Facebook上で作業を行うことができる
 - ...
 - 利用は米国内に限定されている

- 世界中にいるワーカー（Turker）に比較的単純な作業を、Web経由で安価で依頼できるプラットフォーム
 - 例：このWebサイトの感想をください（→ テキストデータ）
 - 例：この画像に鳥は写っていますか（→ Yes／No）
- 自然言語処理、コンピュータビジョンなどのアノテーションづくりに盛んに利用されている
- 現在、（発注側は）US内のみに限定

※ クラウドソーシング (crowd)

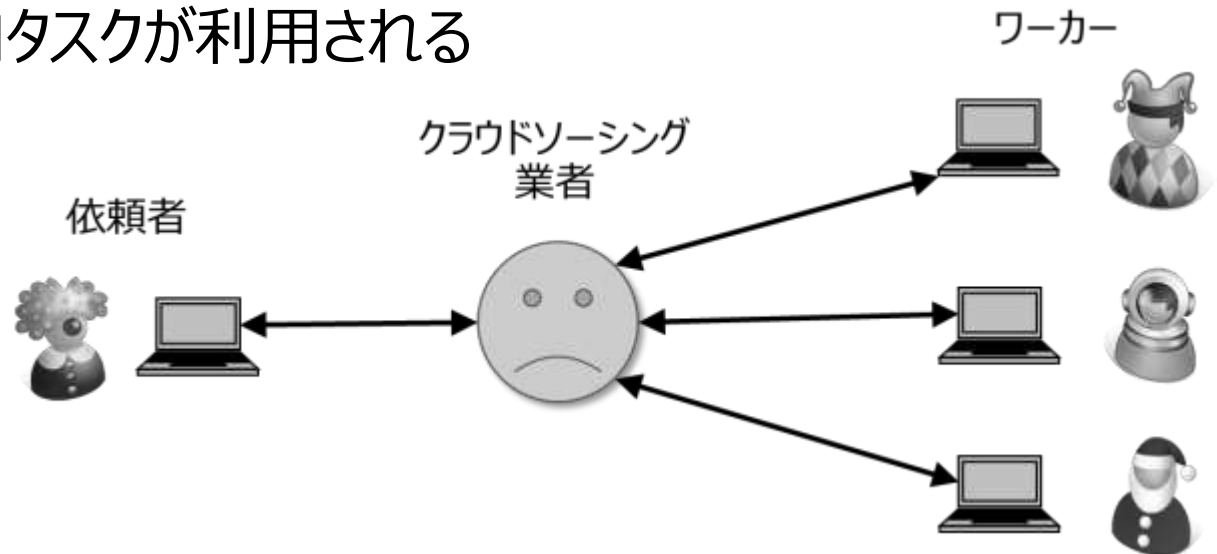
≠ クラウドコンピューティング (cloud)



<http://ja.wikipedia.org/wiki/チェス>

クラウドソーシングにおけるタスク粒度： 情報工学で利用されるのは主にマイクロタスク

- クラウドソーシングのタスクの粒度：
 - 複雑なタスク：Webサイトの作成、ソフトウェアの開発
 - 単純なタスク：ロゴのデザイン、レポートの作成
 - マクロタスク：レストランのレビュー、Webサイトの機能チェック
 - マイクロタスク：画像のラベル付、住所の確認、単純な参照解決
- 情報工学では主にマイクロタスクが利用される



クラウドソーシングと機械学習

機械学習を用いた知的システム実現： 訓練データの収集にクラウドソーシングが使われています

- 画像処理における認識

- 自動で画像内容を認識



自動画像認識

鎌倉の大仏

- 認識器の自動構成には多くの「正解データ」が必要

- 教師付き学習によって認識器をデータから学習
- 正解ラベルは人間が与える



鎌倉の大仏



奈良の大仏

- 正解データの収集にクラウドソーシングを利用

- 自然言語処理： Webテキストのカテゴライズ、情報抽出
- 画像処理： 検索のためのタグ付け、物体認識

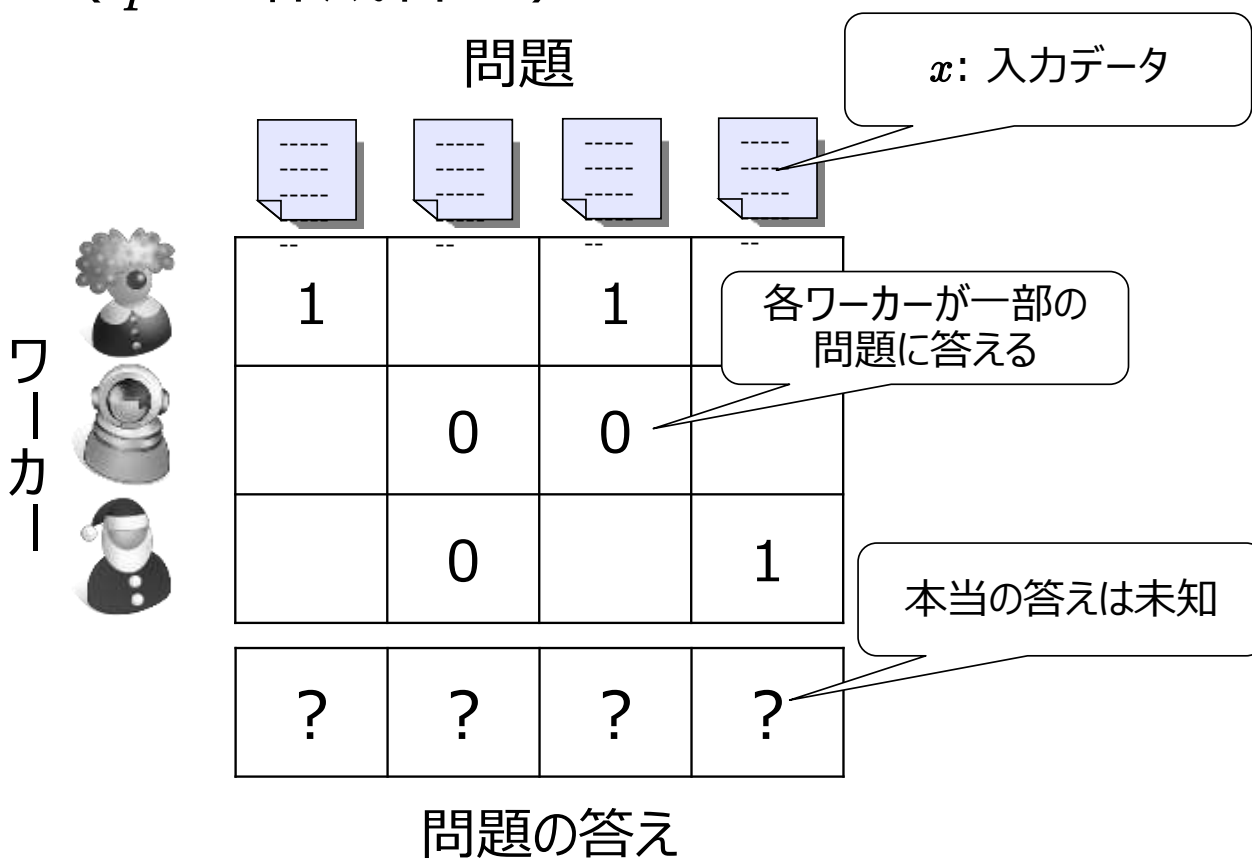
課題： クラウドソーシングの品質管理問題

- 自然言語処理における注釈データの作成は、十分に訓練された人間によって行われる
- クラウドソーシングにおいてはワーカーが課題を達成するための能力を十分にもっているということは保証されない
 - 高い能力の者もいれば、そうでないものもいるという玉石混合
 - 報酬を得ることだけを目的として不誠実に働く「スパムワーカー」
- クラウドソーシングサービスの品質管理機能
 - フィルタリング：ワーカーの遂行タスク数や依頼者承認率など
 - 多数決 ⇒ 十分な数が必要
 - 資格テスト ⇒ 大変
 - 正解セットの利用 ⇒ 必ずしもあるとは限らない

機械学習問題としての品質管理問題：

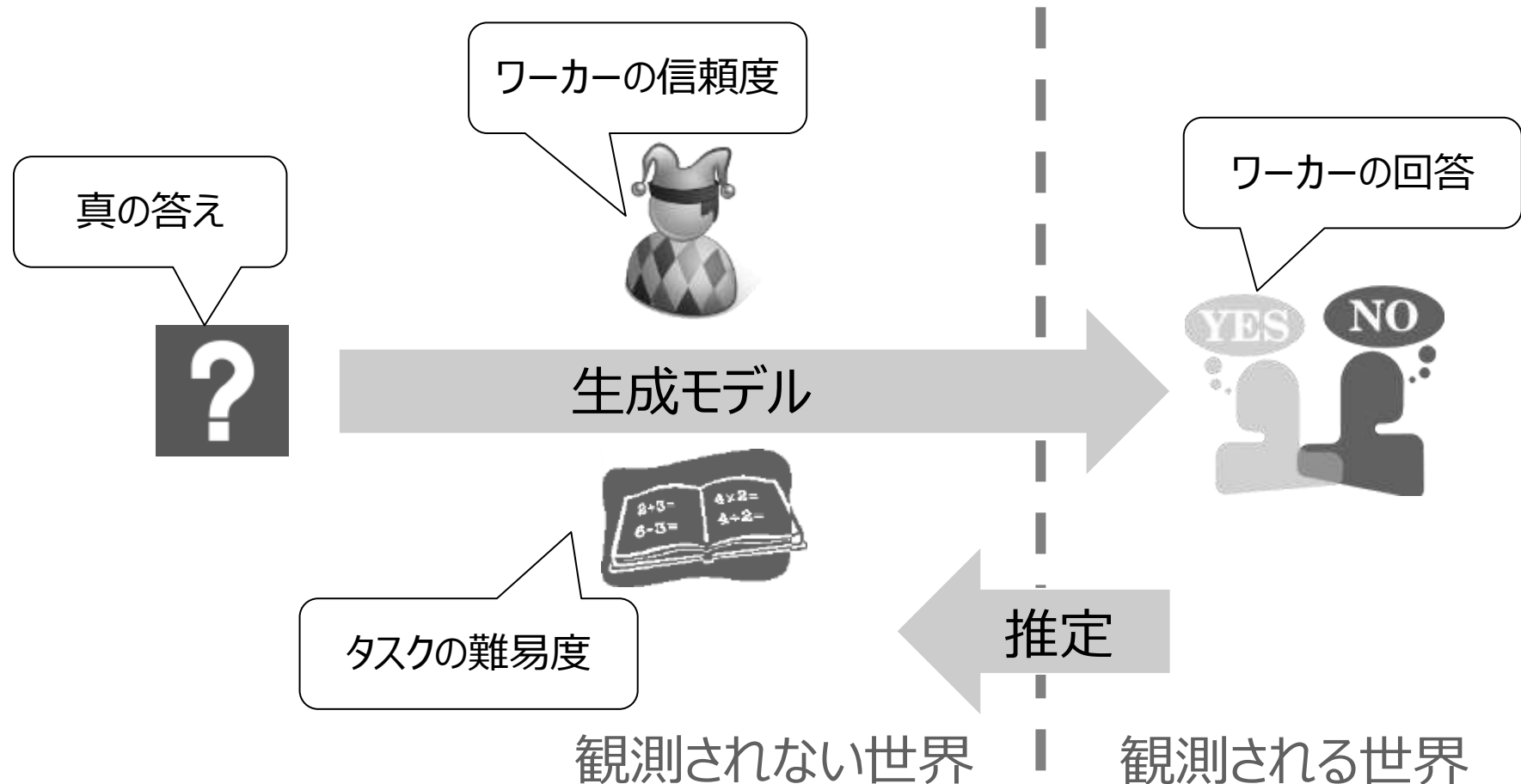
データソースについての情報がたよりに品質のばらつくデータから学習

- 通常は訓練データは $\{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1, \dots, N}$ （入出力ペア） の形式
- データソースについての情報が付加されたデータからの学習
 - 誰がそのデータをつくったか （ $p^{(i)}$: 作成者ID ）
 - いつ作ったか （ $t^{(i)}$ ）
 - 作業条件 （ $c^{(i)}$ ）
 - それぞれの特徴ベクトル
- 新たに考慮すべき項目：
付加データに依存した
データ信頼度



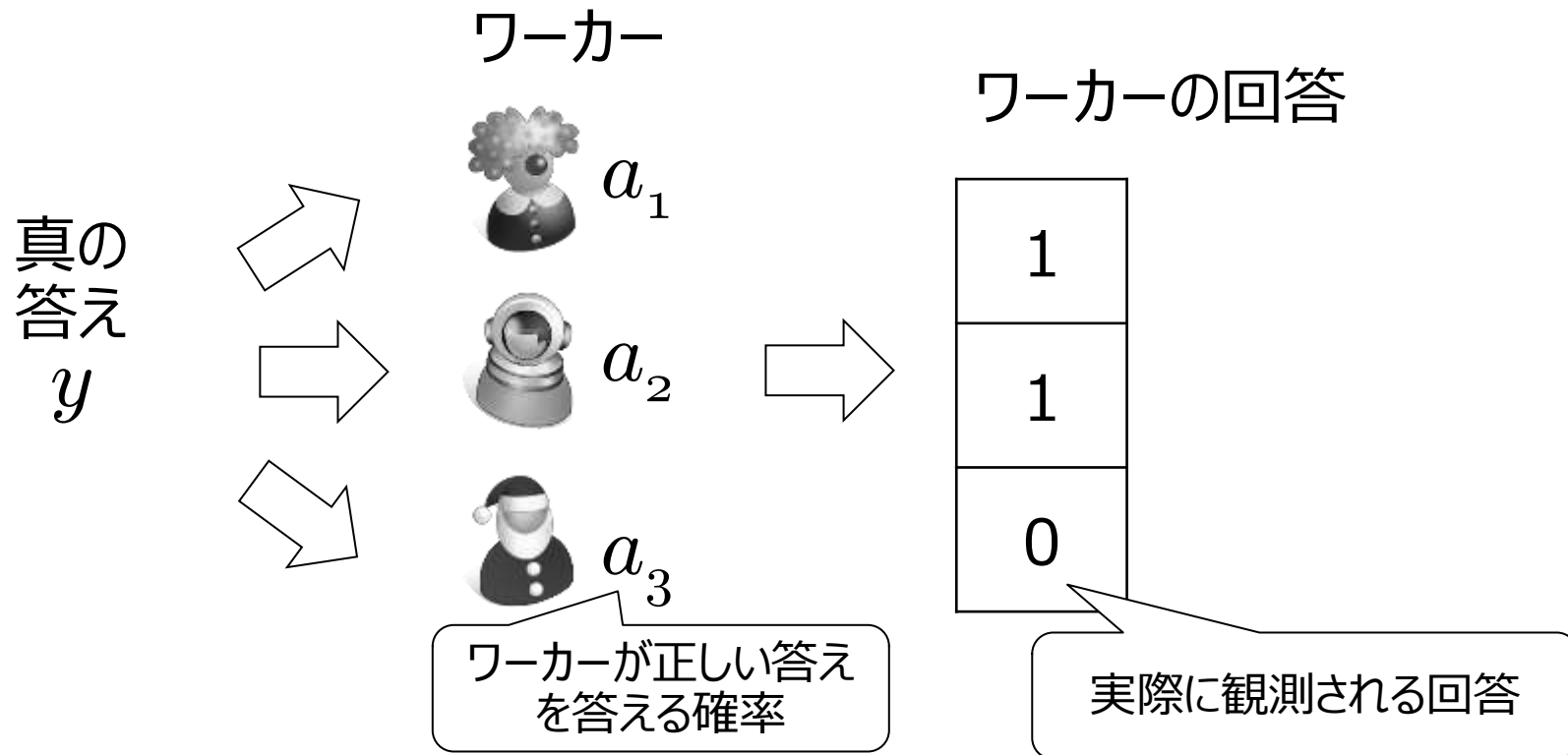
標準的アプローチ： 真の答えを潜在変数とした生成モデル

- 正解から、問題固有の要素とワーカー固有の要素に依存して、回答（観測値）が生成されるようなモデル



Dawid&Skene (1979) による先駆的研究： 各ワーカーの能力と真実を相互に推定する繰り返しアルゴリズム

- 繰り返しアルゴリズムでワーカーの信頼度と真実の推定を繰り返す：
 1. 各ワーカーの信頼度で重みづけを行い、真実の答えを推定する
 2. (推定した) 真実の答えに近いワーカーの信頼度を上げる
- 実際には EM アルゴリズムとよばれる方法で、これを統計的におこなう



ヒューマン コンピューテーション

ヒューマンコンピューテーション： 人間と機械の協調問題解決

- 機械は領域を限定すれば人間を超える
 - ⇔ しかし、依然として「人間にしかできない」領域は多々存在する
- ヒューマンコンピューテーションとは：
 - 計算資源としての人間の労働力を明確に意識し
 - コンピュータと人間の一方のみでは解決できないような問題解決を行うという考え方
- 人間計算資源の調達方法にはさまざまある
 - クラウドソーシングサービス
 - ゲーム化
 - ……



Law & Von Ahn (2012), Human Computation

ヒューマンコンピューテーションのさきがけ： 労働の「ゲーム化」

- 初期の試み：2005年ごろのESPゲーム
 - 地理的に離れた2人のプレイヤーによるWeb上の協力ゲーム
 - 同一の画像に対して二人のプレイヤーがその画像にふさわしいと思うキーワードを独立に与え、これが一致したときに得点が得られる
- ESPゲームは「目的をもったゲーム」（GWAP）
 - 人間による画像へのタグ付け作業をゲームの形で実現したもの
 - 不特定多数のプレイヤーに対して、ゲームの形式を持ちながら何らかの作業を暗黙的に行わせる
 - 暗黙的なタスクは機械にとって不得意、人間には得意
 - 他、音楽のタグ付け、タンパク質の畳み込み等のゲーム化
- ReCAPTCHA：認証にOCRタスクを埋め込む

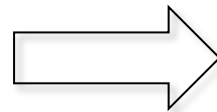



ヒューマンコンピューテーションによる人間と機械のハイブリッドシステム： クラウドソーシングサービスを基盤としたシステムを想定

- ヒューマンコンピューテーション：「計算の一部を人間が行う」というアイデア
 - 検索エンジンの検索結果を人間が並べ替える
 - クイックソートの（2項）比較を人間が行う
 - データベース検索のマッチング判定を人間が行う



```
SELECT name  
FROM people  $p$   
WHERE isFemale( $p$ )
```

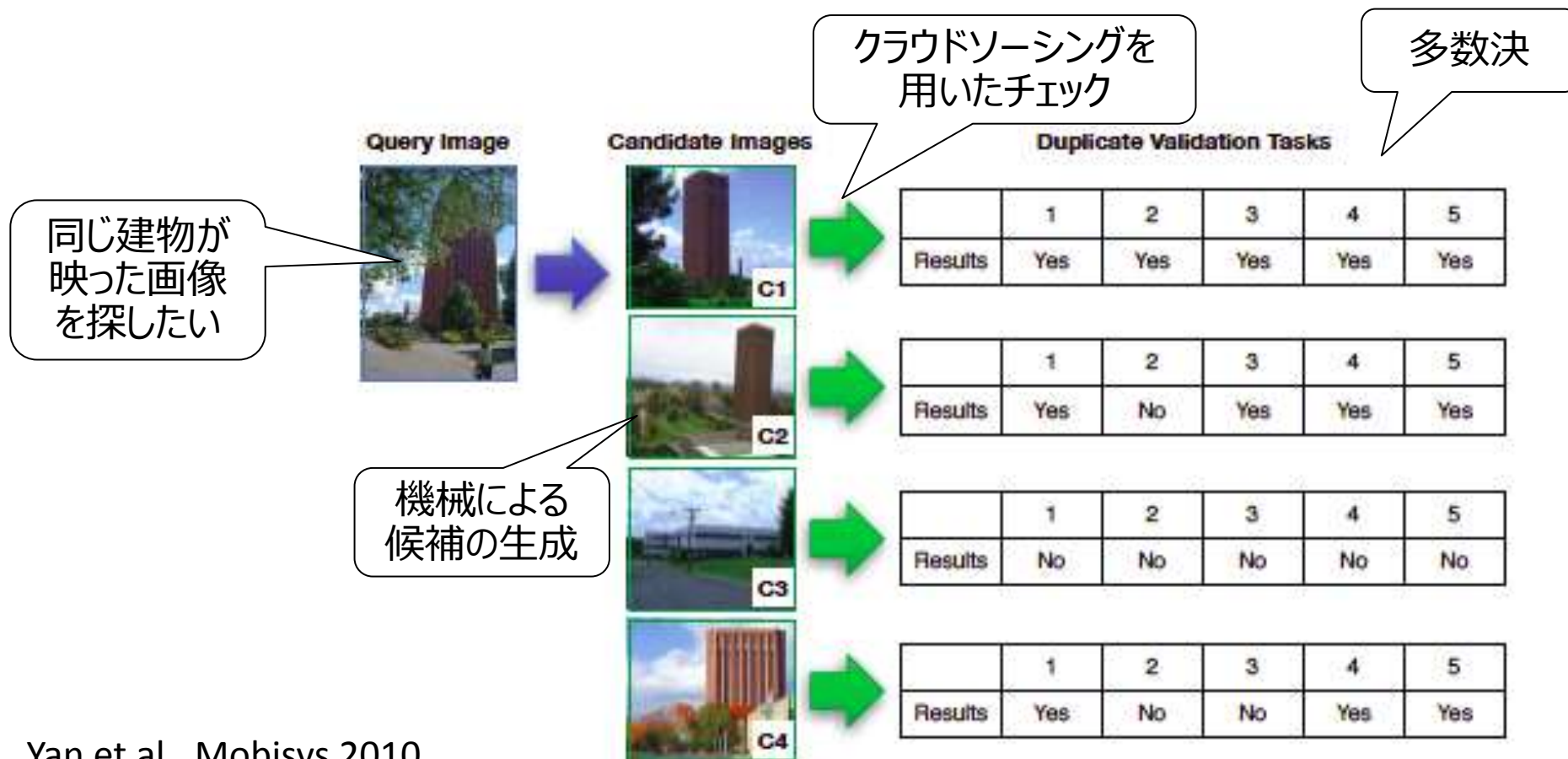


isFemale ()

- 「人間API」としてクラウドソーシングのAPIが利用される
 - AMT API: createHIT(), getAssignments(), approveAssignments(), ...
 - ヒューマンコンピューテーションアルゴリズムの労働力の供給源

情報検索の例：人間による検索結果のフィルタリング／再ランキング

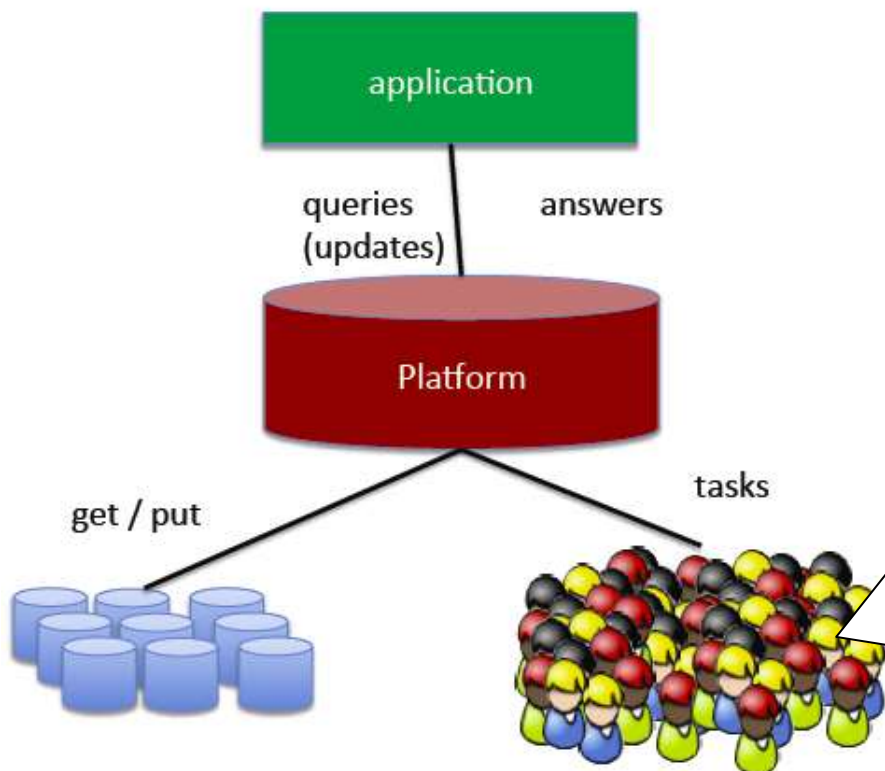
- 人間はより正確に画像を判定できるがスケールしない
- 自動的に候補を出し、人間がフィルタリングする



Yan et al., Mobisys 2010.

データベースシステムの例：人間によるデータ生成、データ比較

- (Crowd)SQL実行の際、一部の計算にクラウドソーシングサービスを利用する
- CrowdDB: データ生成や比較等を人間が行う



SQL実行の際にクラウドソーシングサービスの呼び出しがかかる



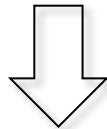
Franklin et al., DBLP 2011.

ヒューマンコンピュータシヨンの最適化： 人間の不確定性を扱うための機械学習

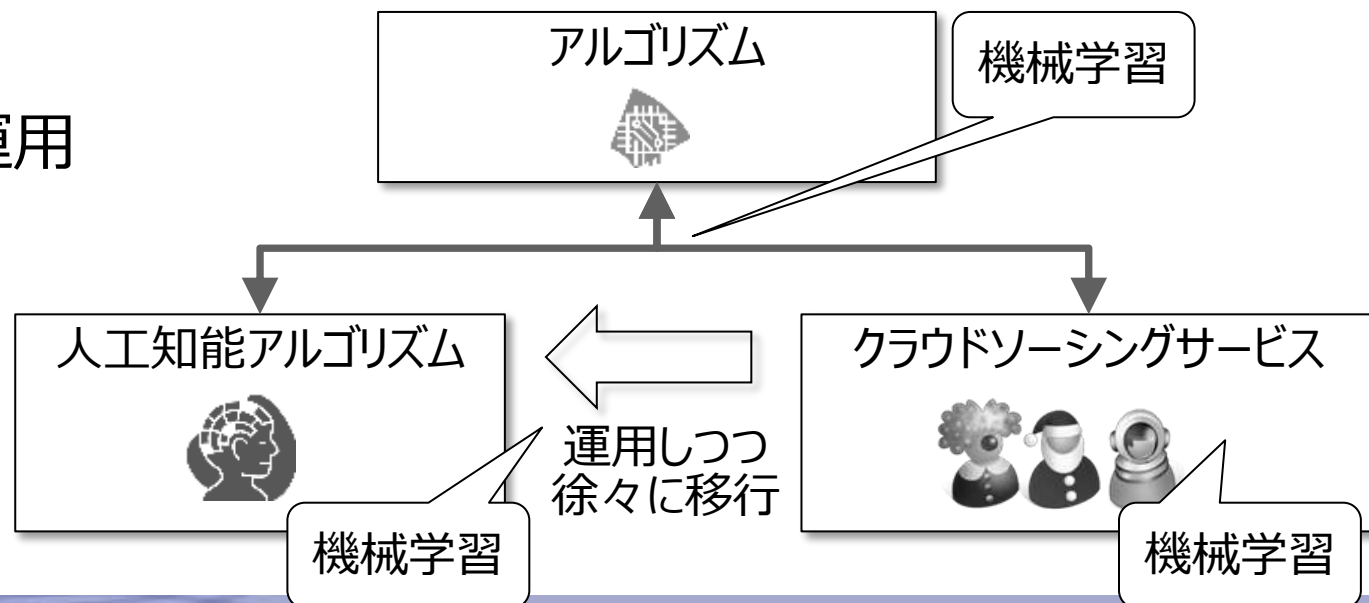
- 現状ではさまざまな試行錯誤の段階
- （主に人間まわりの）最適化が課題：
 - 金銭コスト：安く動かしたい
 - 速度：速く動かしたい
 - 精度：正しい答えを得たい
 - 安定性：いつでも使えて、同じ答えを得たい
 - 安全性：情報を守りたい
- 機械学習は（人間回りの）不確実性に対処するための有効な手段

機械学習の役割： クラウドソーシング／ヒューマンコンピューテーションのための機械学習へ

- 機械学習のためのクラウドソーシング／ヒューマンコンピューテーション
 - 機械学習のデータ収集のために人間を使う



- クラウドソーシング／ヒューマンコンピューテーションのための機械学習
 - クラウドソーシングと機械の併用
 - フローコントロール
 - クラウドソーシング運用



技術的課題：品質管理と人的資源の効果的利用

- ワーカーと成果物の品質管理問題
 - 信頼度の高いワーカーと結果の特定
- ヒューマンコンピュテーションアルゴリズムの効率的実行
 - フローコントロール
- 人的資源の効果的利用
 - タスク⇒人の割り当て
 - 検索(PULL)⇔推薦(PUSH)
 - スケジューリング
 - リアルタイム性
 - モチベートする仕組み（プライシング、教育、...）
- セキュリティ／プライバシー

- 機械学習を用いるためのラベル付きデータ収集のために、クラウドソーシングサービスが用いられつつある
- クラウドソーシングでは、成果物（データ）の質が課題
 - 複数のワーカーが生成したデータから学習を行うための手法が盛んに研究されている
- 人間と機械の得意分野を認識し、両者を合わせて用いる計算パラダイムとしてヒューマンコンピューテーションが認識されつつある
 - 人間の不確実性を扱うための機械学習は重要な技術となる

機械学習界隈で近ごろ注目の話題を紹介しました

1. 機械学習概論

- データからの予測と発見
- 教師つき学習 と 教師なし学習

2. ネットワークと機械学習

- 個々のデータから、データ間の関係へ
- 行列やテンソルを用いた分析

3. 機械学習とクラウドソーシング、ヒューマンコンピューテーション

- クラウドソーシングを利用した機械学習
- ヒューマンコンピューテーションによる機械と人間の協調問題解決

レポート：以下のいずれかについてまとめる（7月4日締切）

1. インパクトのある実問題に対して機械学習が適用された例を調べてまとめよ
2. 推薦アルゴリズムを実際に使ってみて得られた知見を報告せよ
3. ヒューマンコンピューテーションの試みの例を調べてまとめよ

注意： 参考にした文献、Webサイト等の情報を明記すること

- 締切：7/4(水)中に鹿島のメールボックス（工6号館1階）