THE UNIVERSITY OF TOKYO

拡がる 機械学習の応用
～ 機械学習 ＋ {ネットワーク, クラウドソーシング} ～

鹿島久嗣
数理情報学専攻
情報理工学系研究科
東京大学
THE UNIVERSITY OF TOKYO
DEPARTMENT OF MATHEMATICAL INFORMATICS

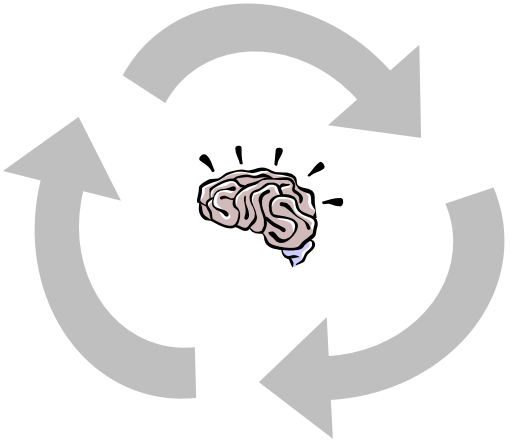
概要： 機械学習の概要を紹介したあと、この分野で最近注目されつつある話題（＋手前味噌）をご紹介します

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
   - 大量の低品質データへの挑戦

企業の研究所で10年間働いたのち、一昨年前に大学へ異動
機械学習が多くの場面で活躍できるよう頑張っています

- 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
- 2008年から東京大学情報理工学系研究科数理情報学専攻 准教授
- 基本的な研究のスタンスは
「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」
  - 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 みえてきます

- 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

「あり」のグループには鳥の名前が含まれている

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

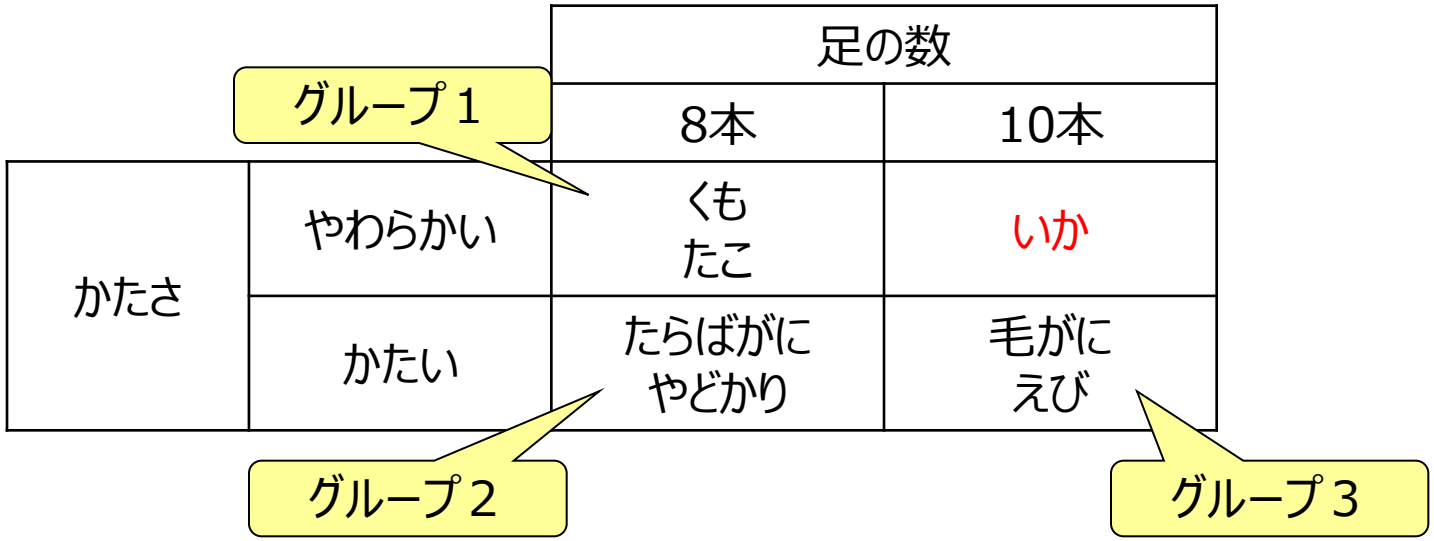
例２ なかまはずれさがし：仲間はずれはどれ？

- 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…



- 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 $\boldsymbol{x}$ が与えられたとき、対応する出力 y を予測したい
  - 入力 $\boldsymbol{x}$：「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、$y = f(\boldsymbol{x})$ となる関数 f がほしい
- しかし、ヒントなしでこれはできない…
  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わけ問題です

- 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 $\boldsymbol{x}$ が与えられたとき、これらをグループ分けしたい
  - 入力 $\boldsymbol{x}$：「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： 結局のところ、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
- 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 $\boldsymbol{x}$ として表現する
  - $\boldsymbol{x}$ を「特徴ベクトル」と呼ぶ
  - その領域を「特徴空間」と呼ぶ

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

赤みの度合い

ビタミンC含有量

「今日の気温」などでもよい

- 特徴ベクトル $\boldsymbol{x}$ 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
  一般的な解はなく、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

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

  - $\boldsymbol{x}^{(i)}$: 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

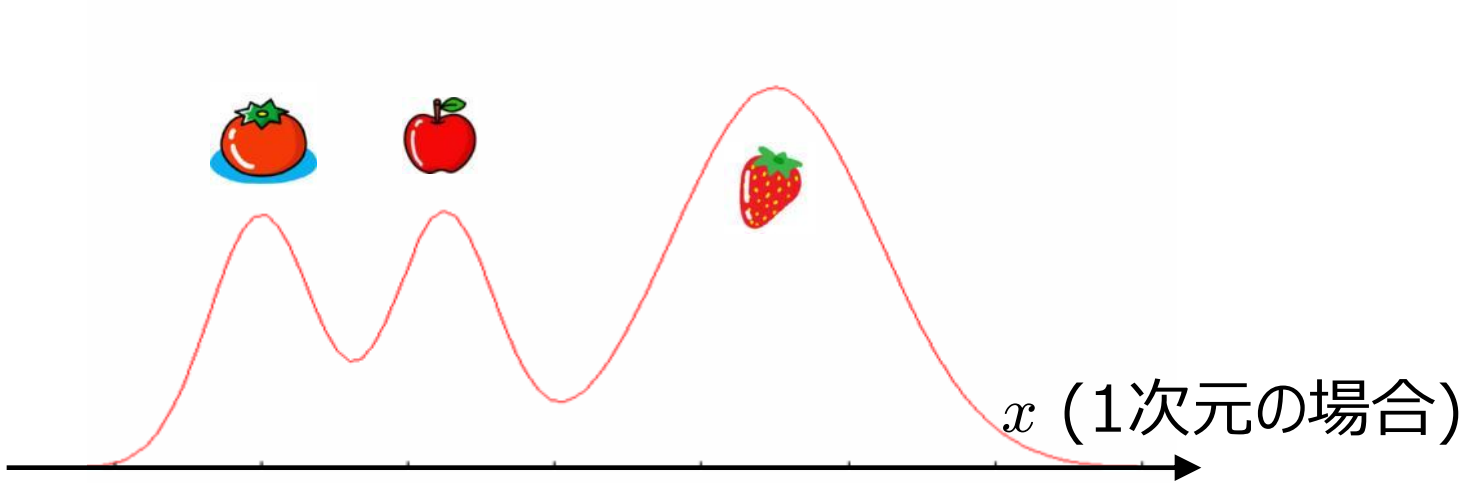
- データはN個の入力信号

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

1つめのデータ　2つめのデータ

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

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

$$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x}) = \mathrm{sign}(\, w_1 x_1 + w_2 x_2 + \ldots + w_D x_D \,)$$

  - sign()は引数が0以上なら+1、0未満なら-1を返す関数
  - $\boldsymbol{w} = (w_1, w_2, \ldots, w_D)^\top$ 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$\boldsymbol{w}$ がきまるとモデルf がきまる

$$f(\boldsymbol{x}) = \mathrm{sign}(\boldsymbol{w}^\top \boldsymbol{x}) = \mathrm{sign}(\, w_1 x_1 + w_1 x_1 + \ldots + w_D x_D \,)$$

- 訓練データから$\boldsymbol{w}$を決定するのが「学習」

$$\{(\boldsymbol{x}^{(1)}, y^{(1)}), (\boldsymbol{x}^{(2)}, y^{(2)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\} \overset{\text{学習}}{\Rightarrow} \boldsymbol{w}$$

- 基本的には、訓練データの入出力を再現できるように$\boldsymbol{w}$を調整する
  - 出力が $y = +1$ のデータについては $\boldsymbol{w}^\top \boldsymbol{x} > 0$ となるように
  - 出力が $y = -1$ のデータについては $\boldsymbol{w}^\top \boldsymbol{x} < 0$ となるように
  - まとめてかくと $y\, \boldsymbol{w}^\top \boldsymbol{x} > 0$

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？」

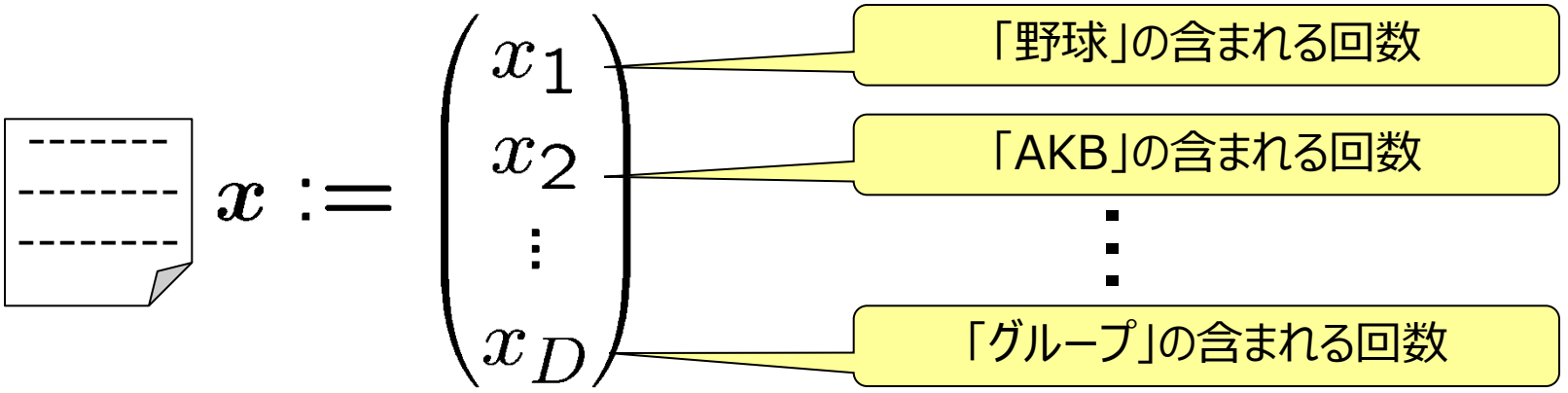
- ある顧客に、融資を行ってよいか
  - 顧客 $\boldsymbol{x}$ を、さまざまな特徴を並べたベクトルで表現
  - 融資を行ってよいか y
    - 融資を行ってよい（返済してくれる）：+1
    - 融資してはいけない（貸し倒れる）：-1
  - マーケティングの文脈では、買ってくれる(+1),買ってくれない(-1)

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

- x_1：過去に延滞したことがあるか？（1/0）
- x_2：リボ払い使用率
- x_D：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）



bag-of-words 表現

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

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)？

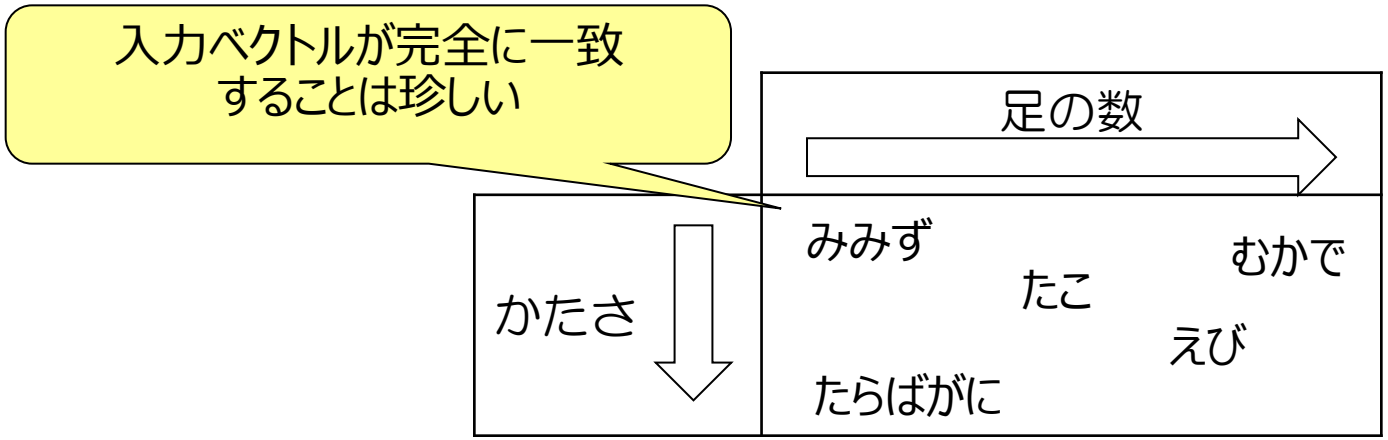
- ほか、顔画像認識や、動画認識

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$\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(N)}\}$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

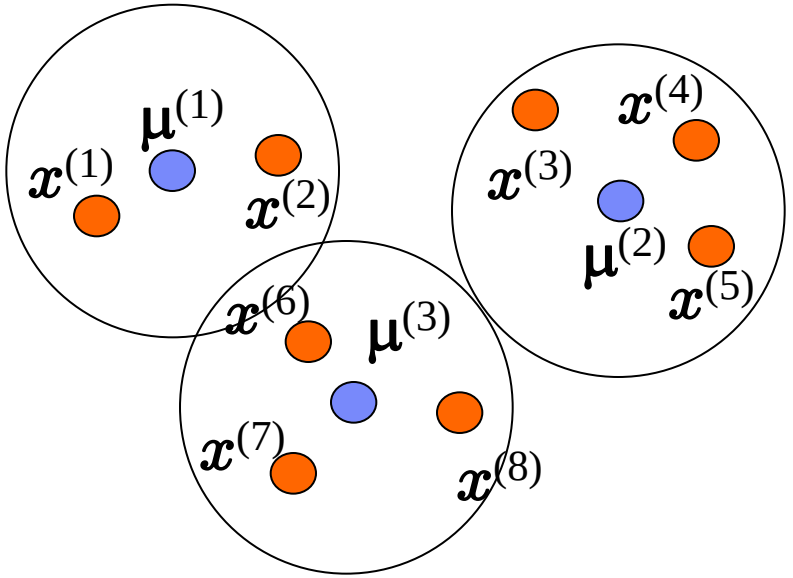
		足の数	
		8本	10本
かたさ	やわらかい	くも たこ	いか
	かたい	たらばがに やどかり	毛がに えび

- 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

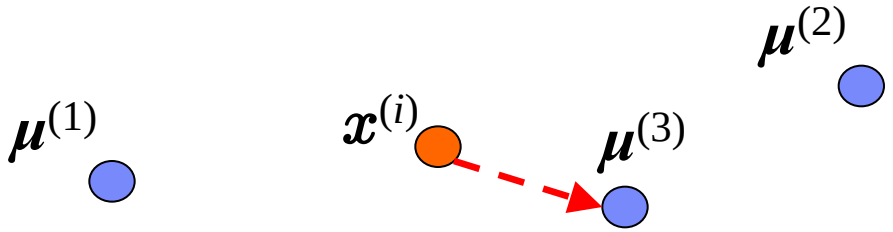


- 代表点への「近さ」（距離）はどう定義するか？
  - 距離関数 $d(\mu^{(k)}, x^{(i)})$ を目的によって適切に定義する
    - たとえばユークリッド距離 $d(\mu^{(k)}, x^{(i)}) = \|\mu^{(k)} - x^{(i)}\|_2^2$

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

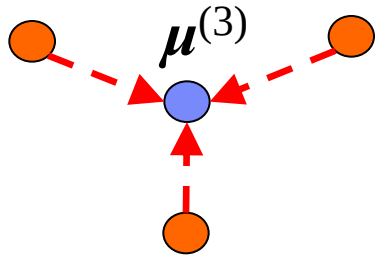
- 以下のステップを収束するまで繰り返す

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



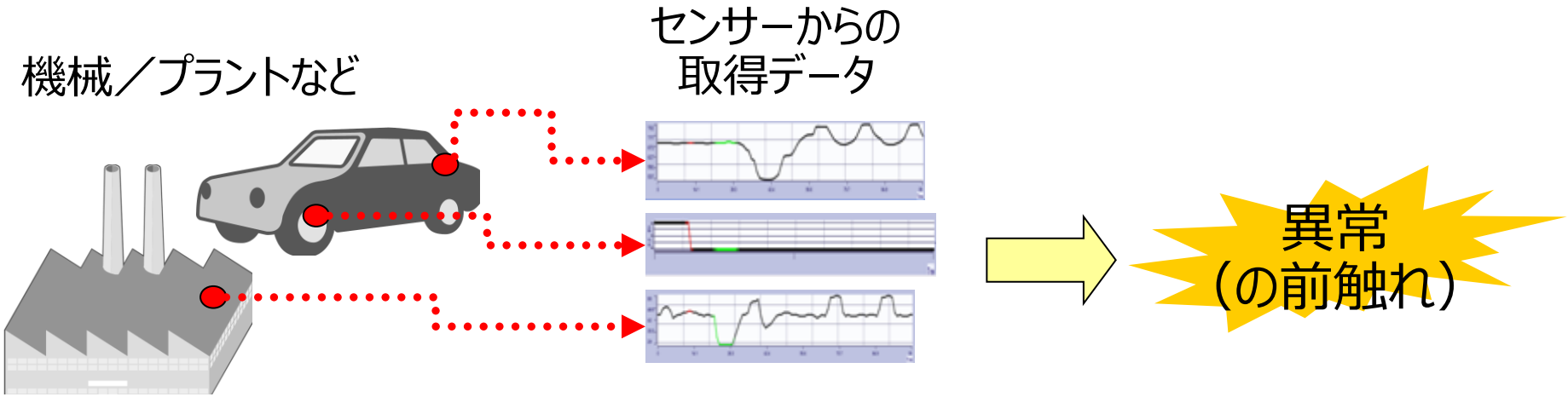
2. 各代表点に所属したデータの平均として代表点を新たに求める

（ユークリッド距離の場合）



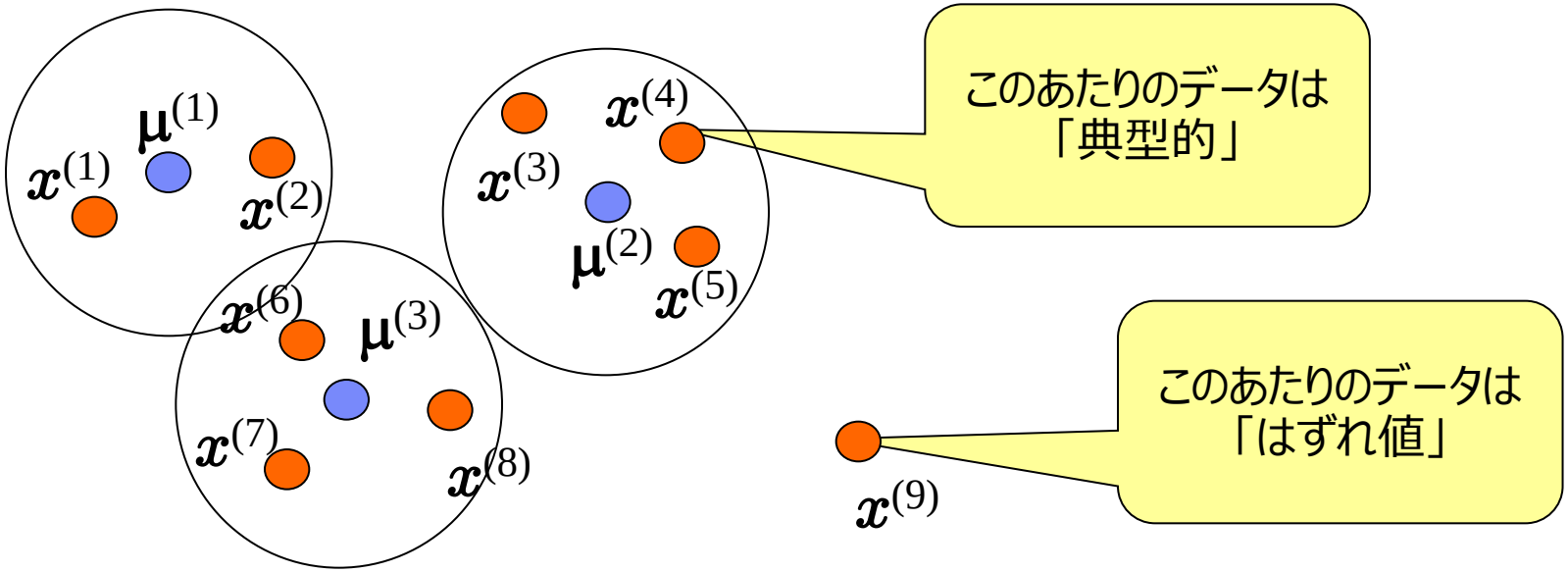
教師なし学習の応用例：異常検知
「ちょっと出かけてくるけど、ヤバそうだったら教えて」

- 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

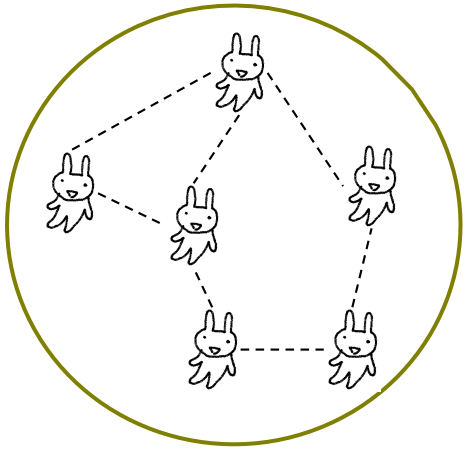
- システムの状態をベクトル x で表現し、教師無し学習によるグループ分けを行う
  - コンピュータ間の通信量、各コマンドやメッセージ頻度
  - 各センサーの計測値の平均、分散、センサー同士の相関
- 代表点から遠い x は「めったに起こらない状態」＝システム異常、不正操作、計測機器故障などの可能性がある



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、それらの間の関係に移行しつつあります

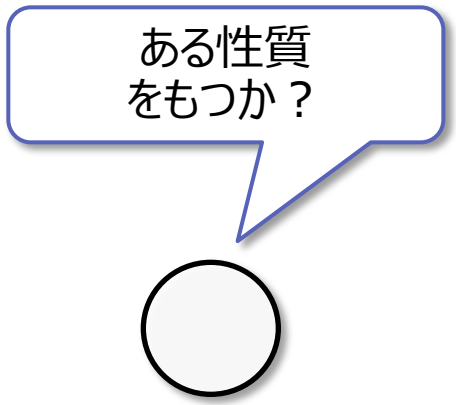
- 従来：「個々のデータを対象とした解析 」

近年：「データの間の関係の解析」

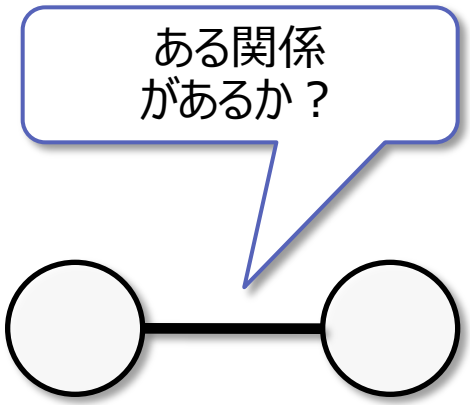
- 関係の分析は様々な領域において盛んに行われつつある
  - ソーシャルネットワーク分析：人間関係
  - オンラインショッピング：顧客と商品の間の関係
- データ間の関係に注目することで、
  個々のデータに注目しているだけでは見えない性質が見えてくることもある
  - コンピュータネットワーク上のプロセス依存関係から異常を予測
  - 複数の脳波時系列の相関関係から思考を読みとる

関係データとは ものごとの関係を表現したデータ です

- 通常のデータ解析では、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



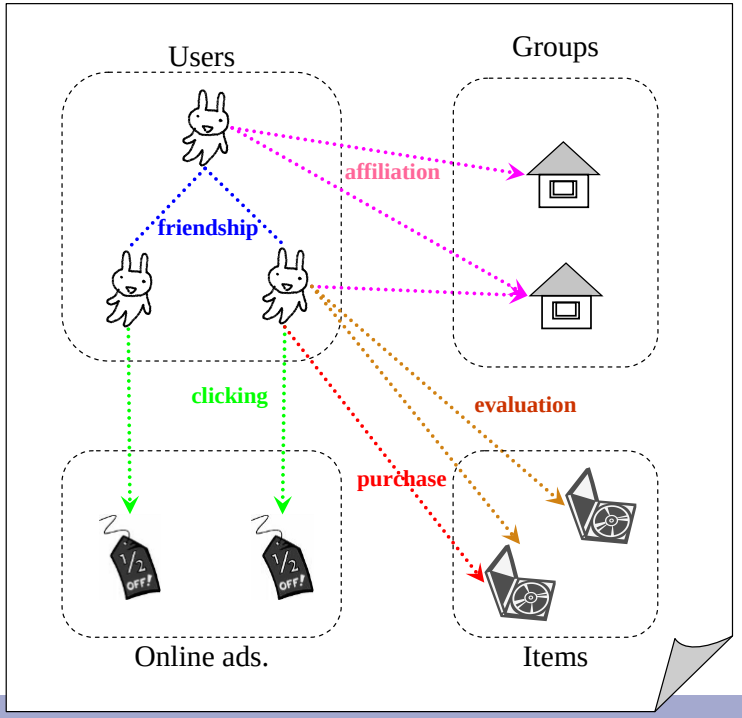
単一データ
についての予測



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

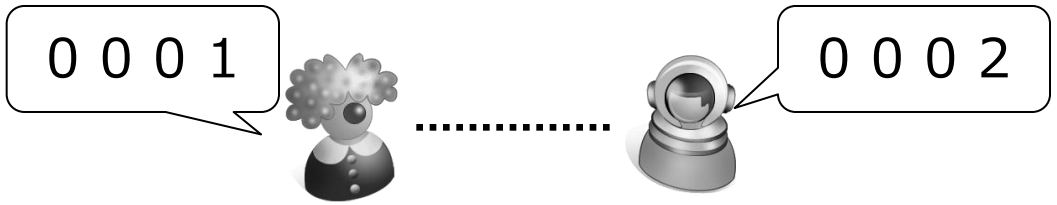
- 予測
  - 推薦システム（協調フィルタリング）
    - 顧客と商品との間の関係（購買、評価）を予測
      - 例： Netflix challenge
  - SNSの友人推薦
  - 新規薬剤候補の探索
- 発見
  - 顧客セグメンテーションの発見
  - 協調するタンパク質グループの発見
  - 例外の発見

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	○○	４０代	男性	東京都	…
0002	××	３０代	女性	大阪府	…

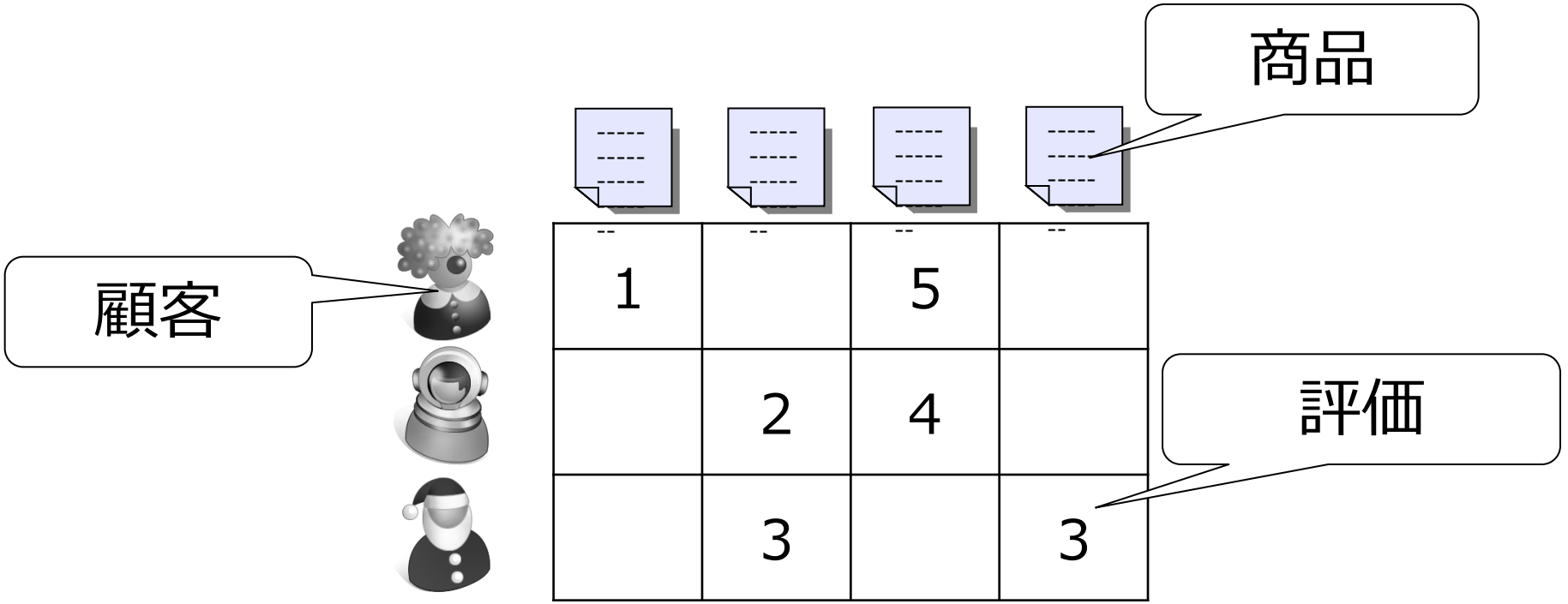
- 関係データはこれらの間の関係を表す



- 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

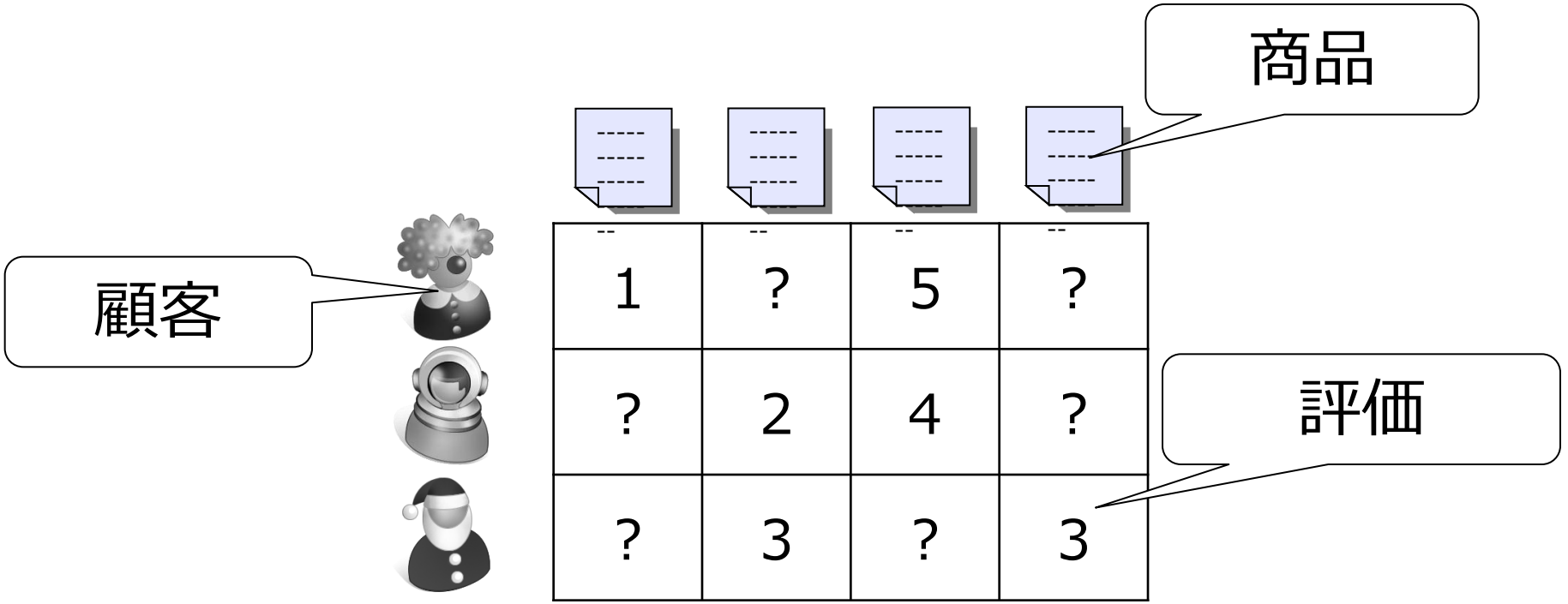


行列データの解析手法

- 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

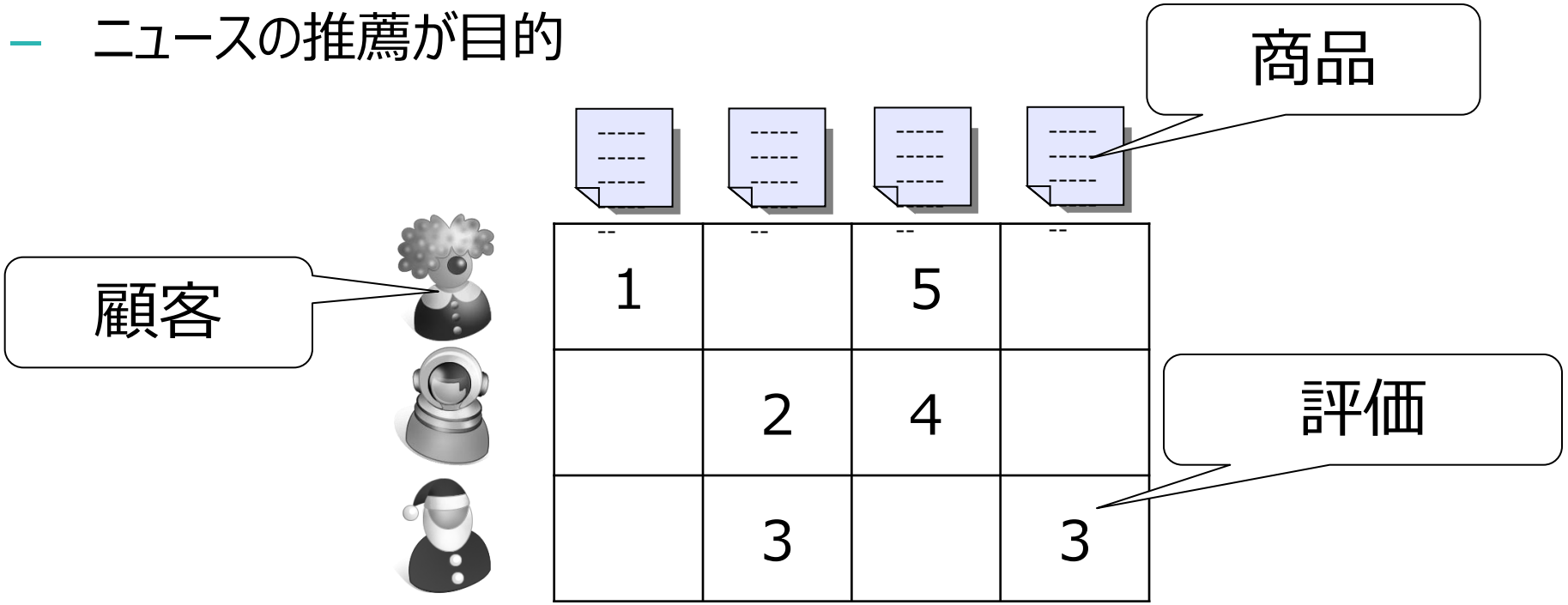
行列の補完問題は、行列の観測部分をもとに、未知の部分を予測する問題です

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



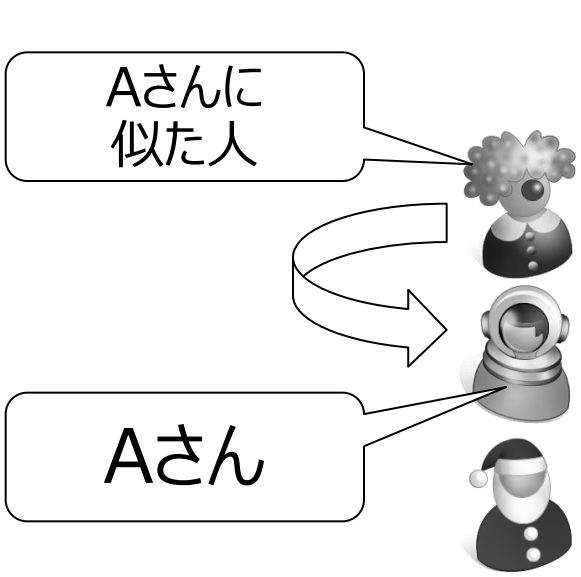
GroupLens：協調フィルタリングの初等的手法

- 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から、わかっていない部分を予測したい
- 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



1	2	5	5
	2	4	5 ?
	5		3

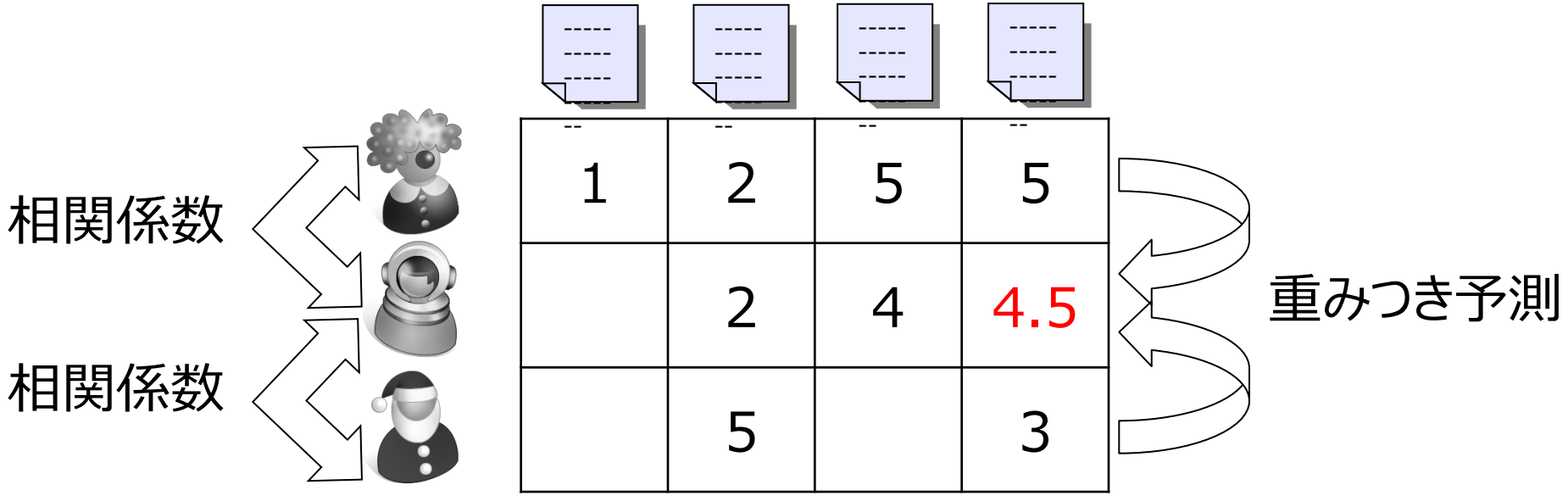
知りたい要素

「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

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

- 同様に、商品間の類似度を用いることも可能

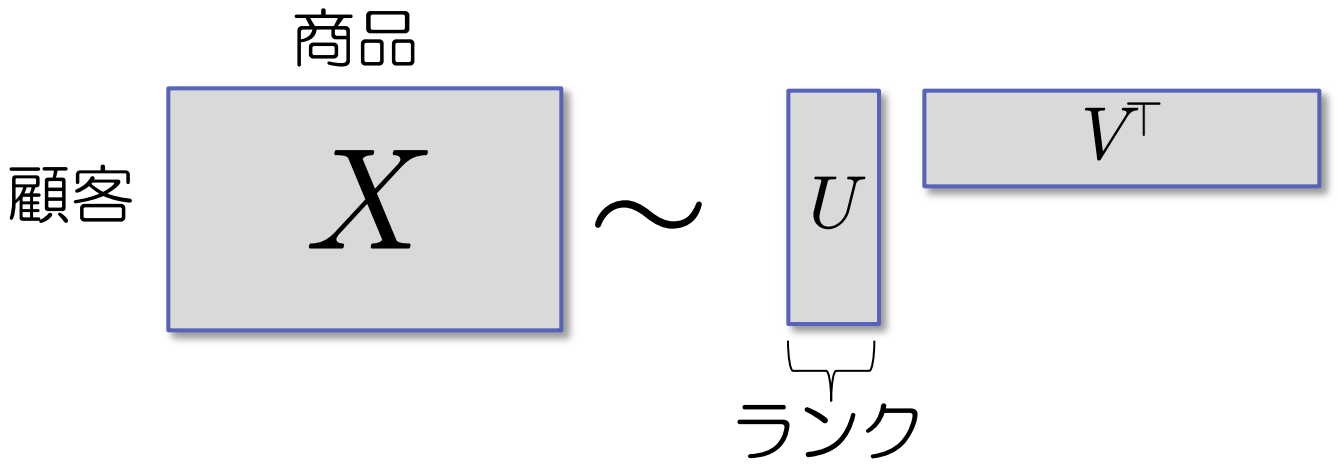


協調フィルタリングの初等的手法は行列の低ランク性を暗に仮定しています

- 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（⇒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

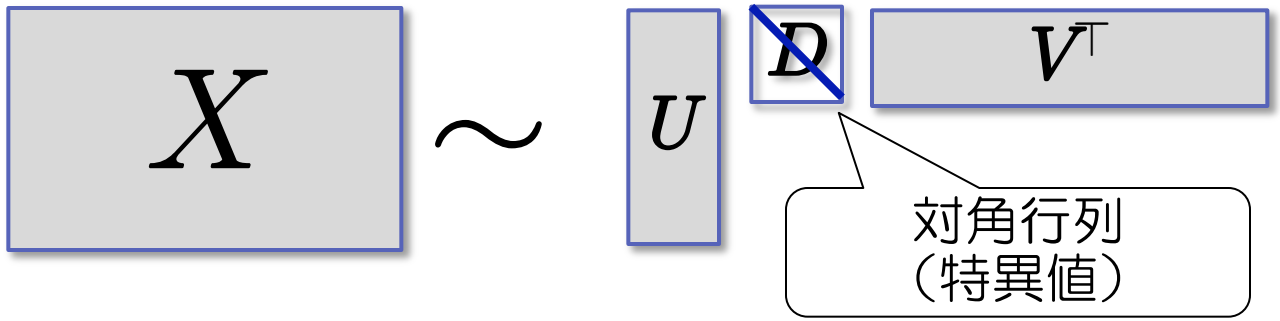


$$\text{minimize}_{Y} \|X - Y\|_{\text{F}}^{2} \text{ s.t. } \text{rank}(Y) \leq k$$

- 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 $X = UV^\top$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

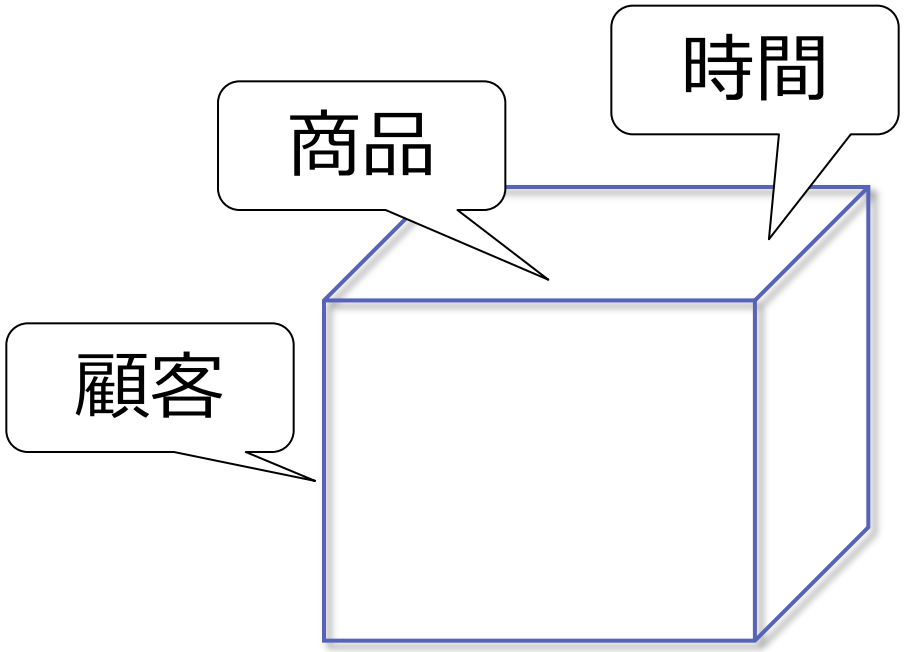


  - 制約： $U^\top U = I \quad V^\top V = I$

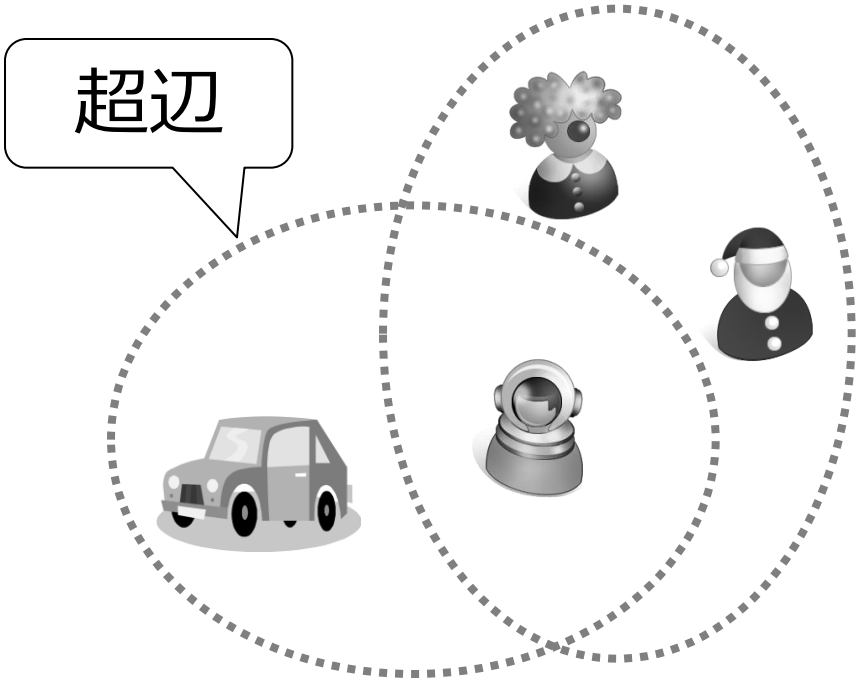
- $X^\top X$ の固有値問題になる
  - 固有値を大きい方からk個とる

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



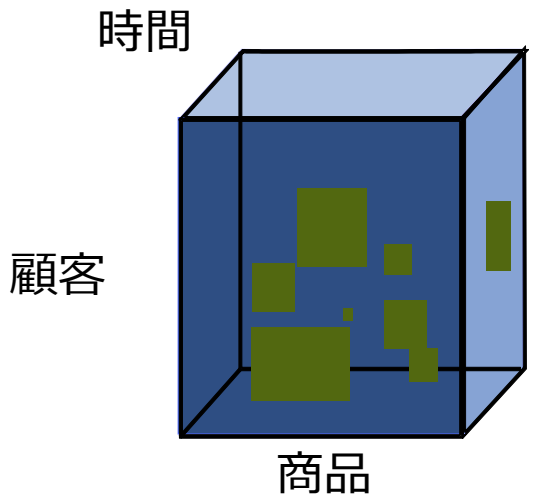
多次元配列



ハイパーグラフ

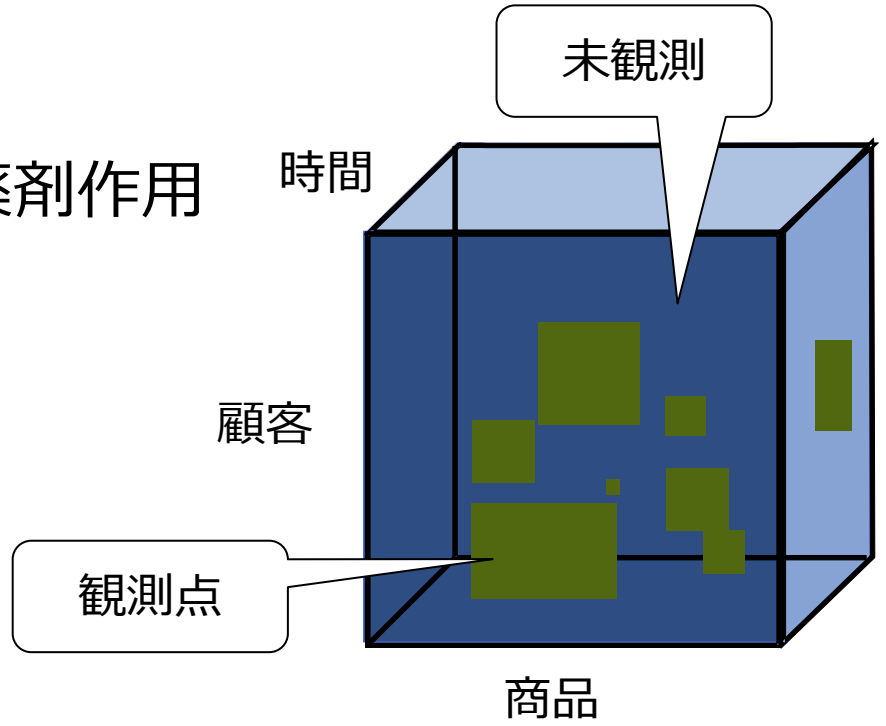
テンソル（多次元配列）は行列よりも一般的な関係の表現です

- 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



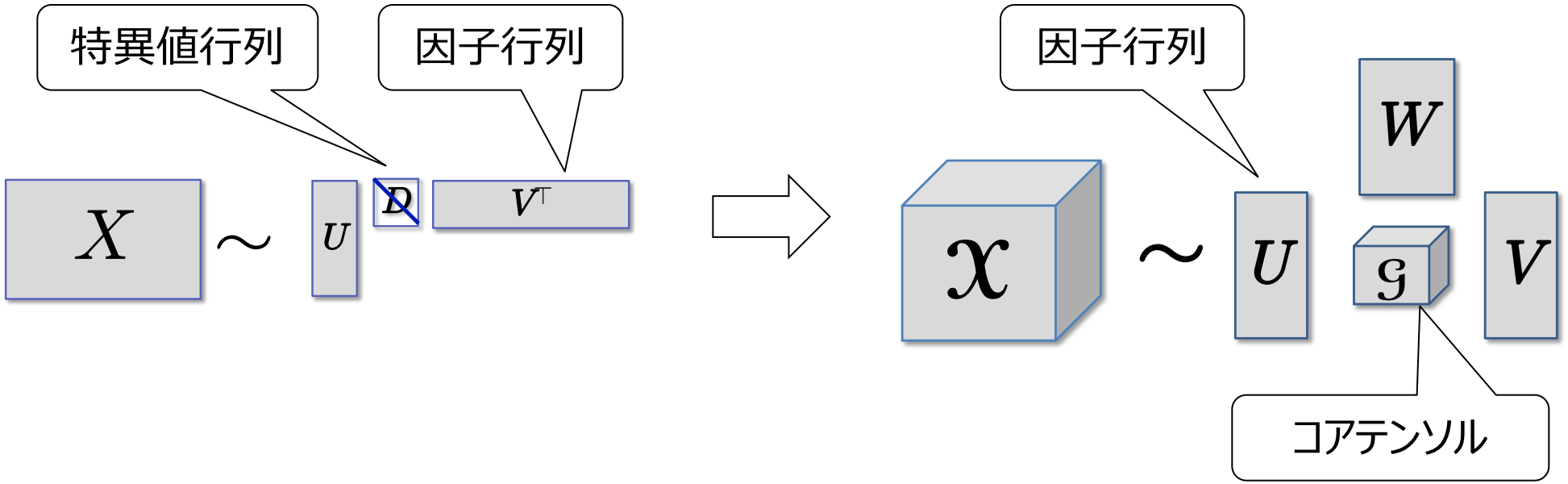
テンソル補完問題：
より高次の関係の予測問題を扱います

- 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



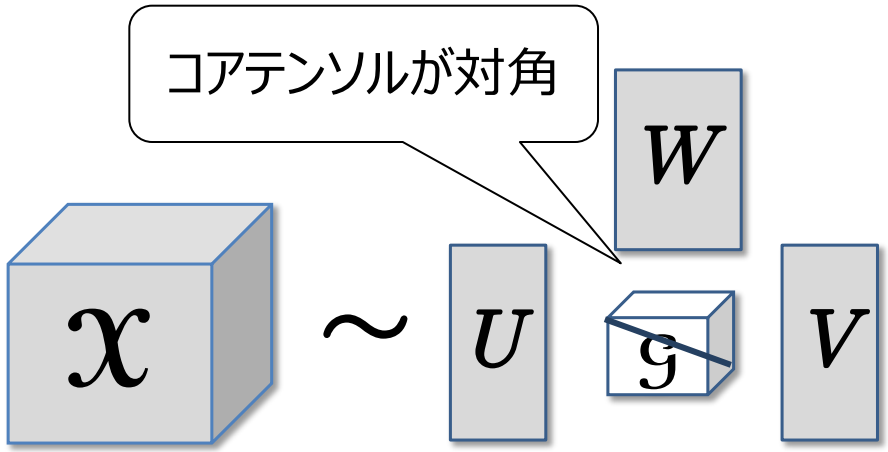
行列分解は多次元配列（テンソル）の低ラング分解に一般化されます

- 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

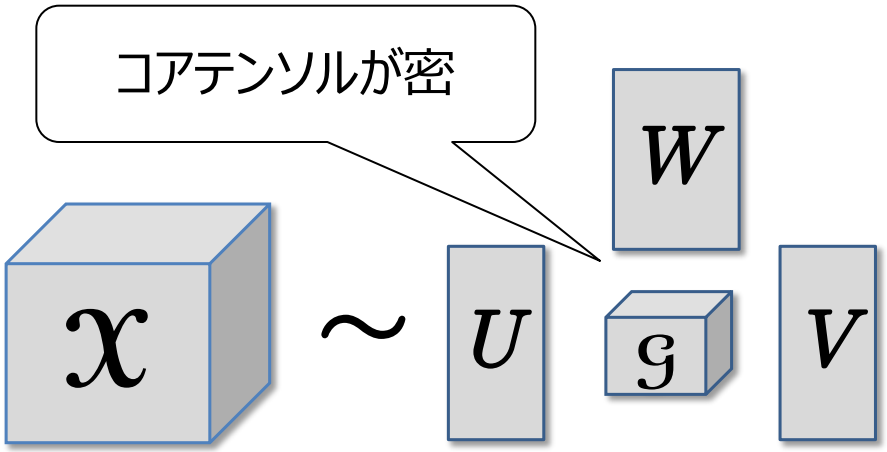


テンソル分解のタイプ：CP分解とTucker分解

- 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）



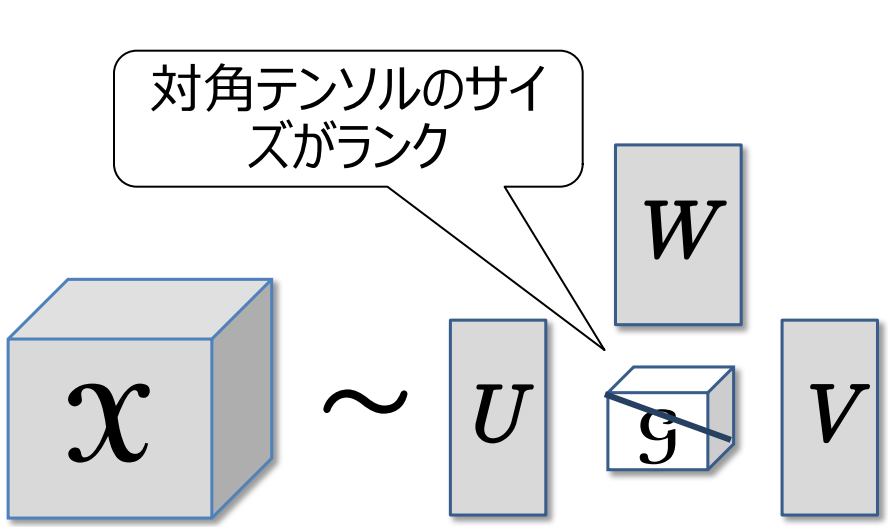
CP分解



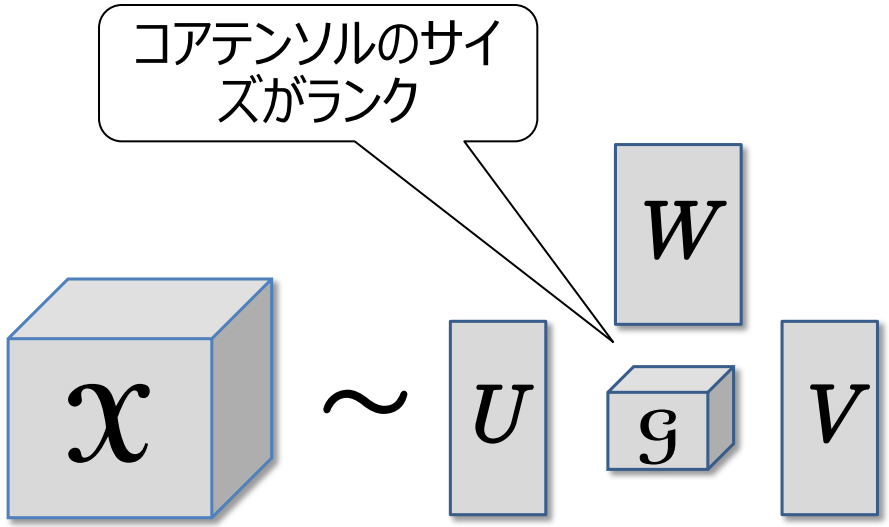
Tucker分解

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



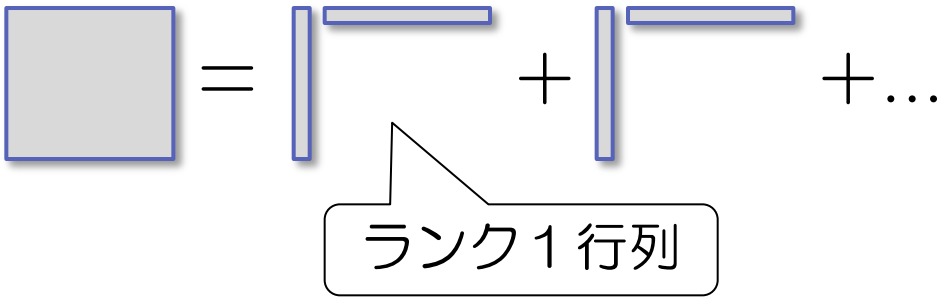
CP分解



Tucker分解

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

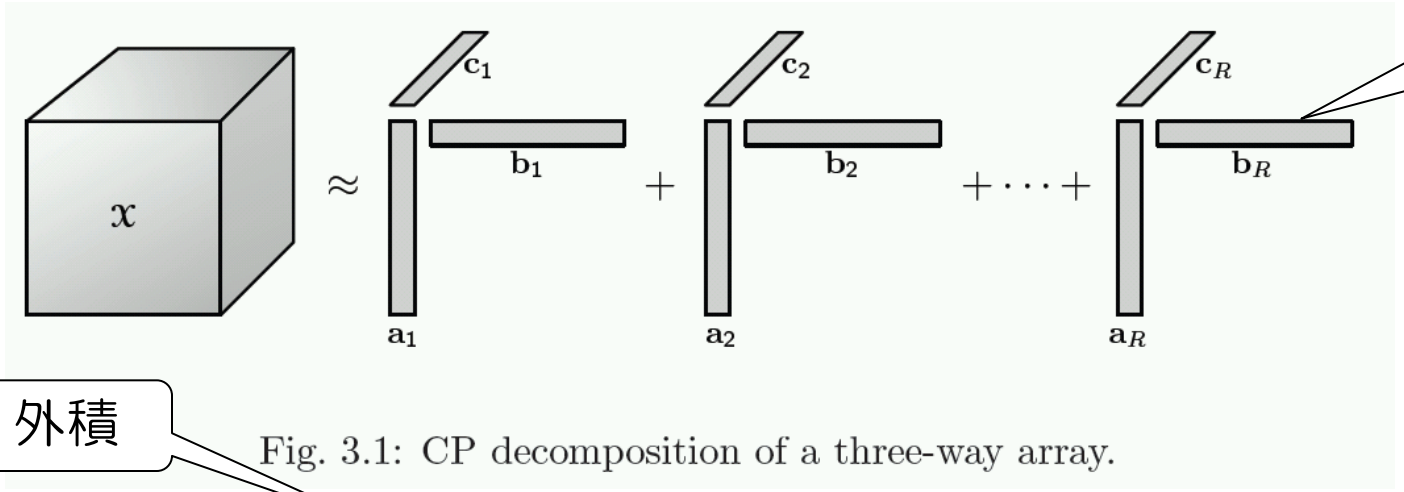


Fig. 3.1: CP decomposition of a three-way array.

外積

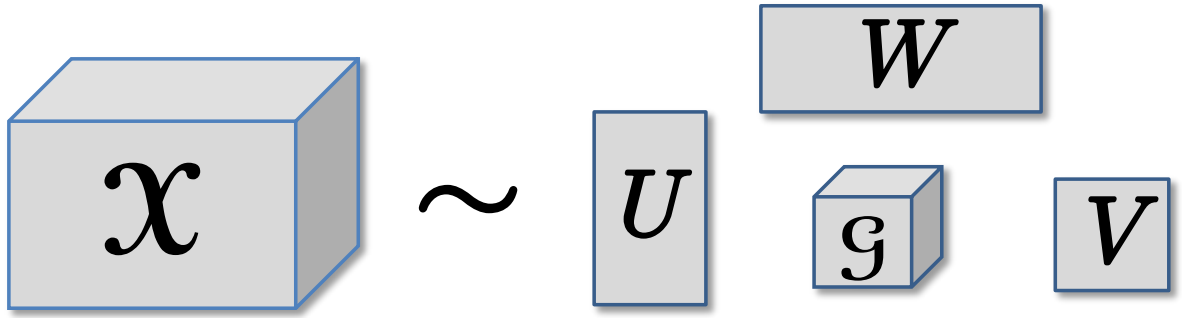
$$\mathcal{X} \sim \sum_r \lambda_r \boldsymbol{a}_r \circ \boldsymbol{b}_r \circ \boldsymbol{c}_r \qquad x_{ijk} = \sum_r \lambda_r a_{ri} b_{rj} c_{rk}$$

* 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

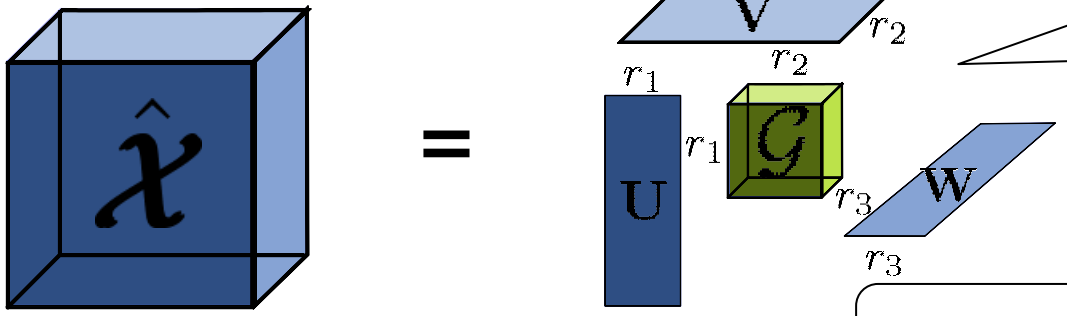
- ソーシャルネットワーク分析 (人×人×時間)
- Webリンク解析　（Webページ×Webページ×アンカーテキスト）
- タグ推薦　（人×Webページ×タグ）
- 画像認識　（画像×人×向き×明るさ×…）
- 脳波解析（場所×場所×時間）

タグ推薦タスクへの応用例 (Rendle *et al.* (2010))

削除

高次関係の予測ではデータの疎性が課題です

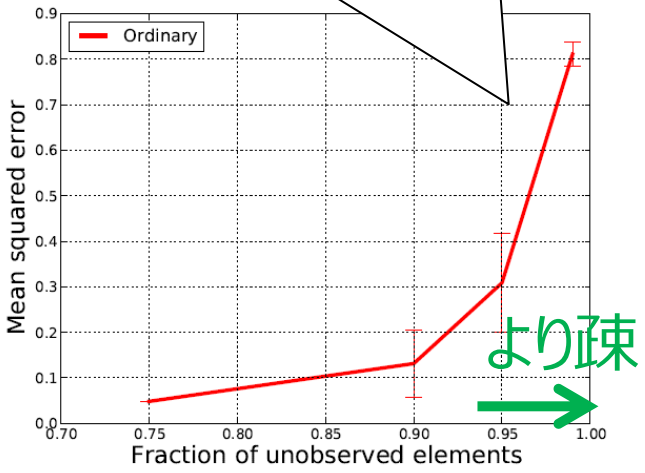
- テンソルの分析では低ラング性の仮定を行うのが通常
  - Tucker分解など



低ランク分解

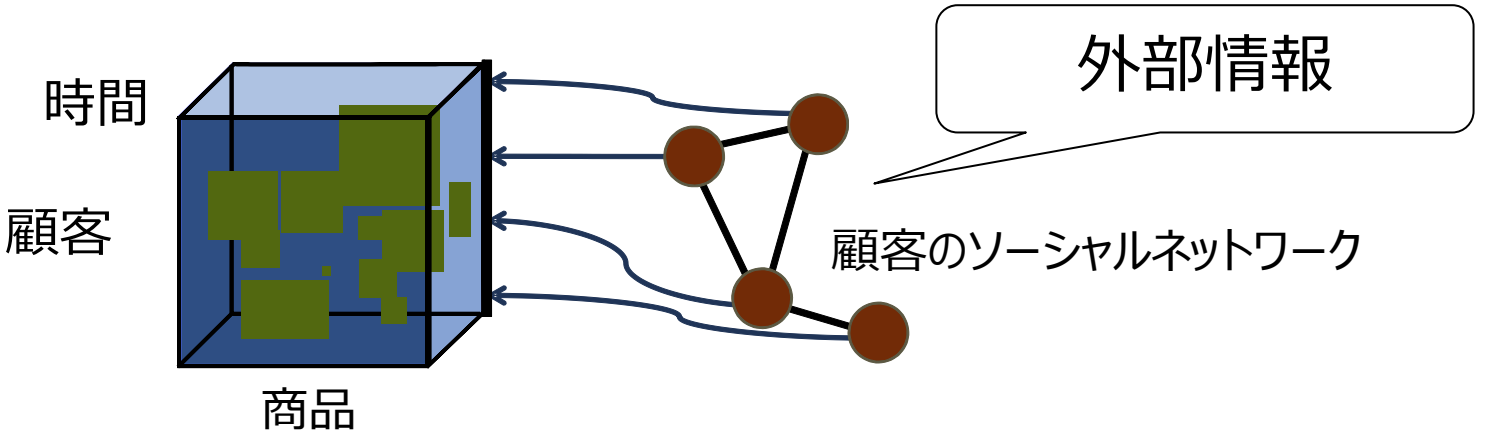
- 課題：疎なデータの予測
  - 観測部分が少ないときに、
    予測精度が著しく悪化してしまう
  - 可能な関係の数は組み合わせ的に増加する
- 低ランクの仮定だけでは足りない！

予測精度の悪化



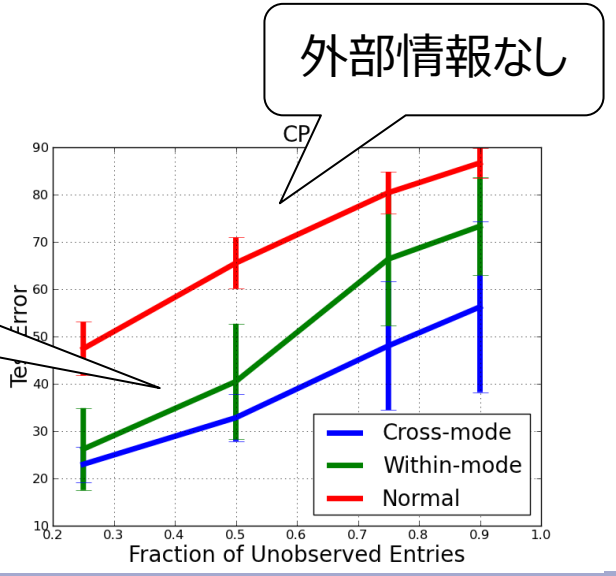
疎性への取り組み：低ランク性の仮定だけでは足りないので、併せて外部情報を利用します

- 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

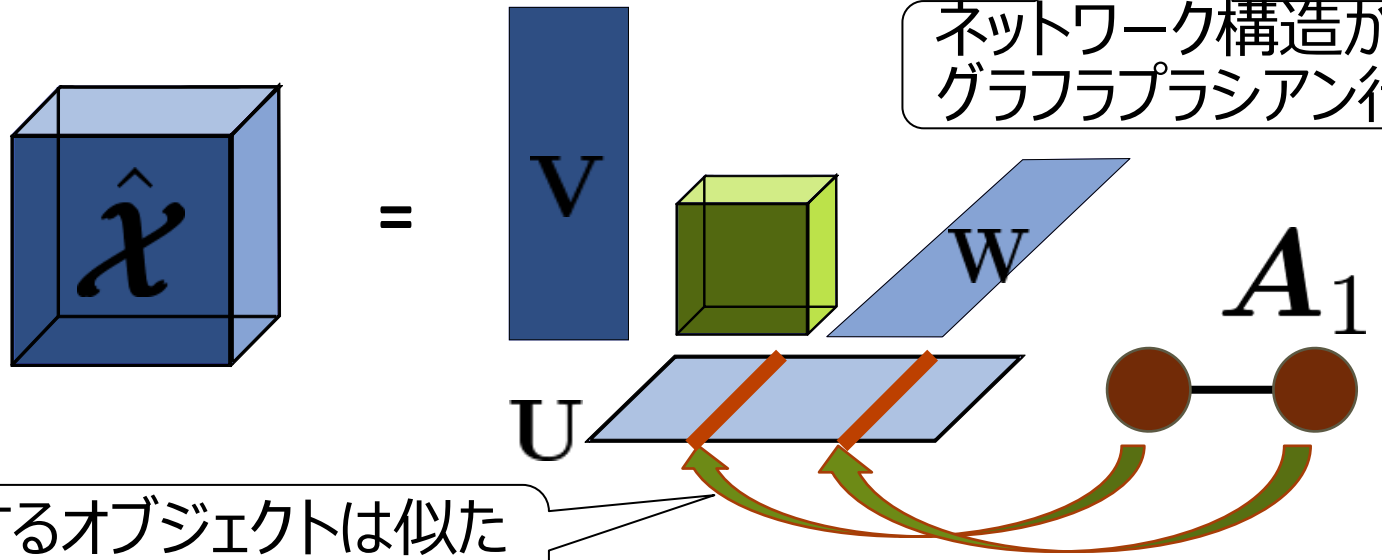
- ネットワーク正則化： 外部情報として与えられる関係情報を推論のガイドに用いる （最適化問題の目的関数に導入）
  - 隣り合ったデータが振る舞いをするように働く

近似誤差の項

ネットワーク正則化項

$$\min \|\mathcal{X} - \hat{\mathcal{X}}\|_F^2 + \mathrm{tr}\left(U^\top L_1 U\right)$$

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」

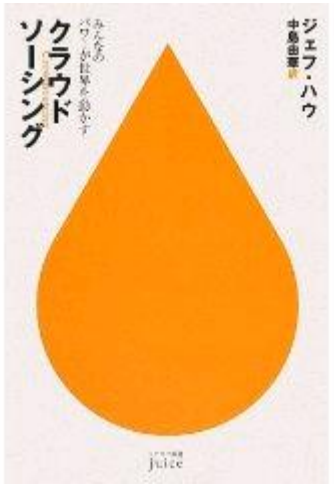
まとめ

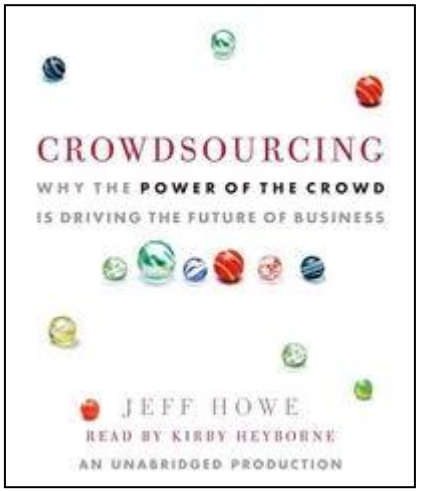
- データ解析の興味の対象は、単一のデータから、データ間の関係へ
- データ間の関係は、行列やテンソルで表現される
- 低ランク分解を中心とした分析手法が用いられる
- データが少ない場合でも、補助情報をうまく使うと予測精度が向上する

クラウドソーシングと機械学習

クラウドソーシングとは、不特定多数に仕事を依頼する仕組みです

- Jeff Howe によって名づけられた「（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」一般をさす
  - いわゆるアウトソーシングでの相手が不特定多数
  - 報酬の有無、公募形式の有無、などさまざまな場合がありうる
- 例）
  - P&Gのとりくみ：課題を広く一般に公開し、解決策を募る
  - Amazon MechanicalTurk：
    クラウドソーシングのプラットフォーム





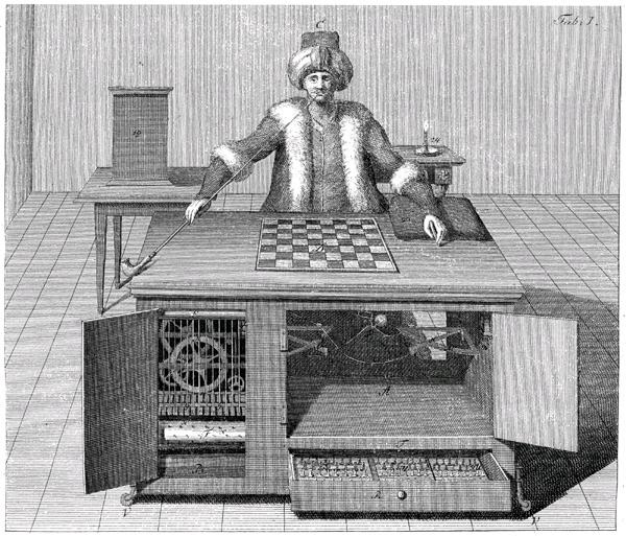
Howe , J. / Crowdsourcing (2004)

Mechanical Turk：2005年に米Amazonが開始したクラウドソーシングのプラットフォーム

- 世界中にいるワーカー（Turker）に簡単な作業を、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け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

≠ クラウドコンピューティング



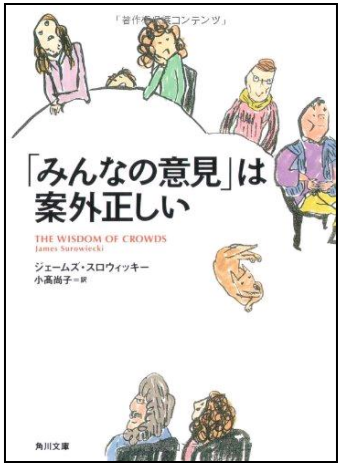
http://ja.wikipedia.org/wiki/チェス

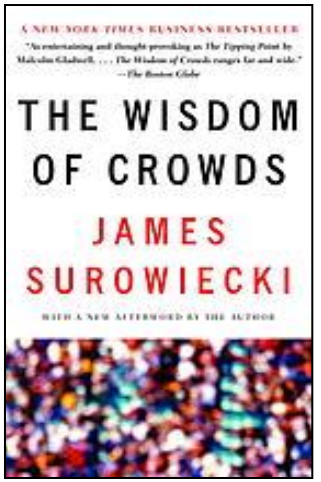
ITの世界でもクラウドソーシングが用いられつつあります

- 人間をセンサーやプロセッサとしてもちいる （アクチュエータとしては？）
  - 交通監視システム
  - 放射線量測定
- データアノテーションに用いる
  - 自然言語処理、コンピュータビジョン、バイオ
  - 機械学習の正解データ作成
- ヒューマン・コンピュテーション： 人間を計算資源の一部として（おもに機械には不得意だが人間には簡単な）タスクを（非明示的に）行わせる
  - ESP game ： ゲームを通じて画像のタグ付
  - 「人間をその一部に使う」プログラミング言語なども

ワーカーの品質管理問題は クラウドソーシングにおける本質的課題です

- クラウドソーシングで得られる成果物の品質には大きなばらつきがある
  - ワーカーごとに能力が大きく異なる
  - 「スパム」ワーカーの存在
  - 「群衆の叡智」の前提をくずす、非独立性
    - 「多くの非専門家の意見を統合すると、専門家単体の意見よりも正しい」
    - メディアからの影響、ソーシャルネットワーク
- ワーカーの品質管理が重要課題
  - タスクの分割
  - タスクの割り当て
  - ワーカーをいかにモチベートするか
  - 結果の統合





Surowiecki , J. / The Wisdom of Crowds (2004)

クラウドソーシング＋機械学習には2種類あります：クラウドソーシング{を用いた, のための} 機械学習

- 機械学習とクラウドソーシングのかかわりを考えたとき２つの可能性がある

1. クラウドソーシングを用いた機械学習
   - データ収集や解析にクラウドソーシングを利用
2. クラウドソーシングのための機械学習
   - クラウドソーシング（システム／サービス）がよりうまく働くためのデータ解析

- 現状では、前者がほとんど

機械学習においてもクラウドソーシングが用いられつつありますがやはり得られるデータの品質は課題です

- 2000年代に入ってからインターネット経由で安価な労働力を調達できるクラウドソーシング（特にAmazon MechanicalTurk）が教師つき学習のラベルづけを安価に行う方法として用いられている
  - とくに自然言語処理、画像処理における利用が盛ん
    - 自然言語処理におけるAMT利用
      - テキストに対して感情を振る、単語が似ているかどうかなど５つのタスク
    - コンピュータビジョンにおけるAMT利用
      - 領域分割や身体のマーカー付けなどの４タスク
- やはり、得られるラベルの品質のばらつきが指摘されている
  - 同じデータに対して複数人でラベル付けすることで品質を上げる

Cheap and Fast—But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks.
Snow, R. et al., In EMNLP, 2008.
Utility data annotation with Amazon Mechanical Turk.
Sorokin, A. and Forsyth, D. In CVPR workshop on Internet Vision, 2008.

複数の（信頼できない）情報から 真実を導くための手段として機械学習のテクニックが注目されています

- 1970年代後半に複数の医者の診断を統合する文脈で始まった「専門家の意見の統合」問題
  - 単純な多数決よりも良い判断を下したい
- 多数の素人から集められたラベルの信頼度を上げるための手段として、意見統合のワザが再注目されている
  - 「専門家の意見の統合」から「群衆の意見の統合」へ
- そして今、さらに「群衆の意見からの学習（複数の教師からの学習）」へと形をかえ、機械学習分野でひそかに盛り上がりつつある
  - ラベル推定からモデル推定へ
  - たとえば、機械学習のトップ会議のひとつNIPSでは「crowd」をタイトルに含むものが5本（＝多い）

機械学習における典型的設定：
各ワーカーがいくつかの問題に対して回答を行い、そこから真実を推定

- 各データに対し、複数のワーカーがラベルをつける
  - ただし、その信頼度は低い
- 目的：
  - 真実の答え
  - もしくは、予測モデル

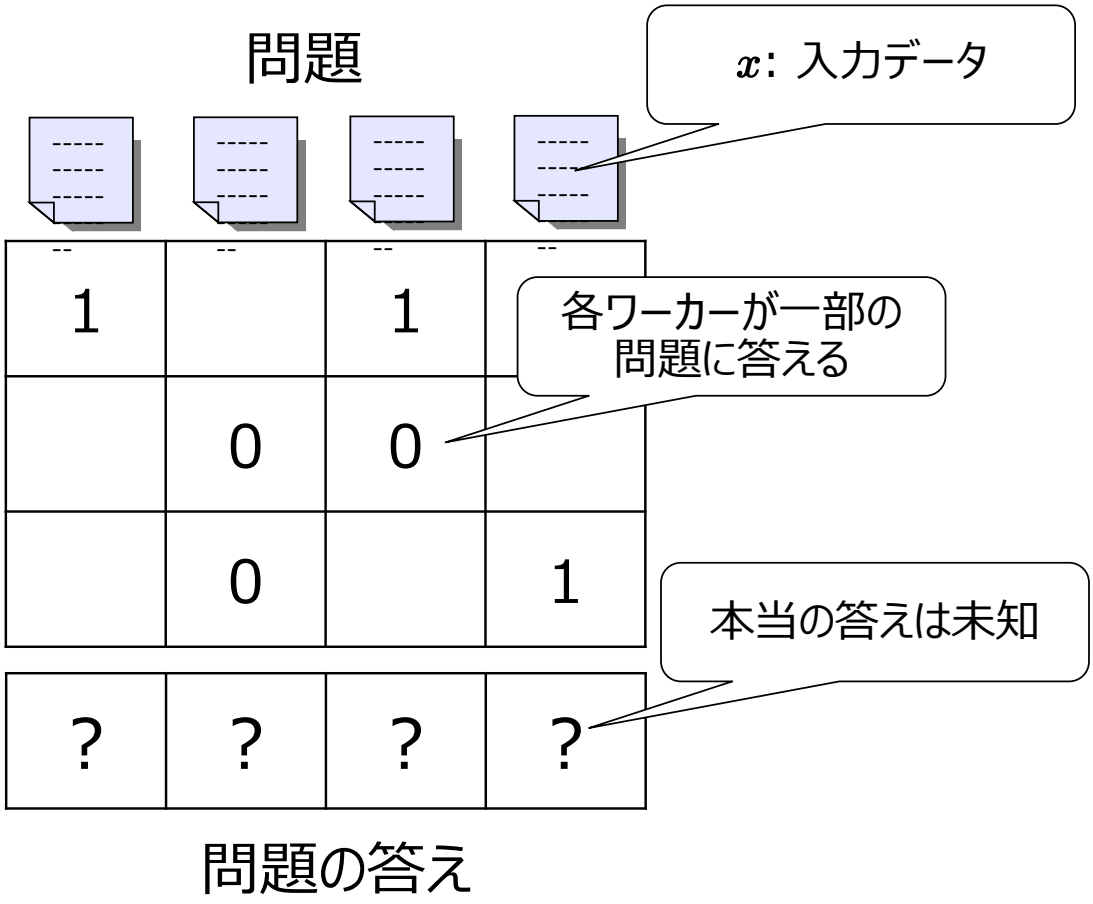
    （本当はこっち）
- 教師つき学習でいうと、訓練データが

  （入力, 出力）

  ↓

  （入力, 出力,ワーカーID）

  となった



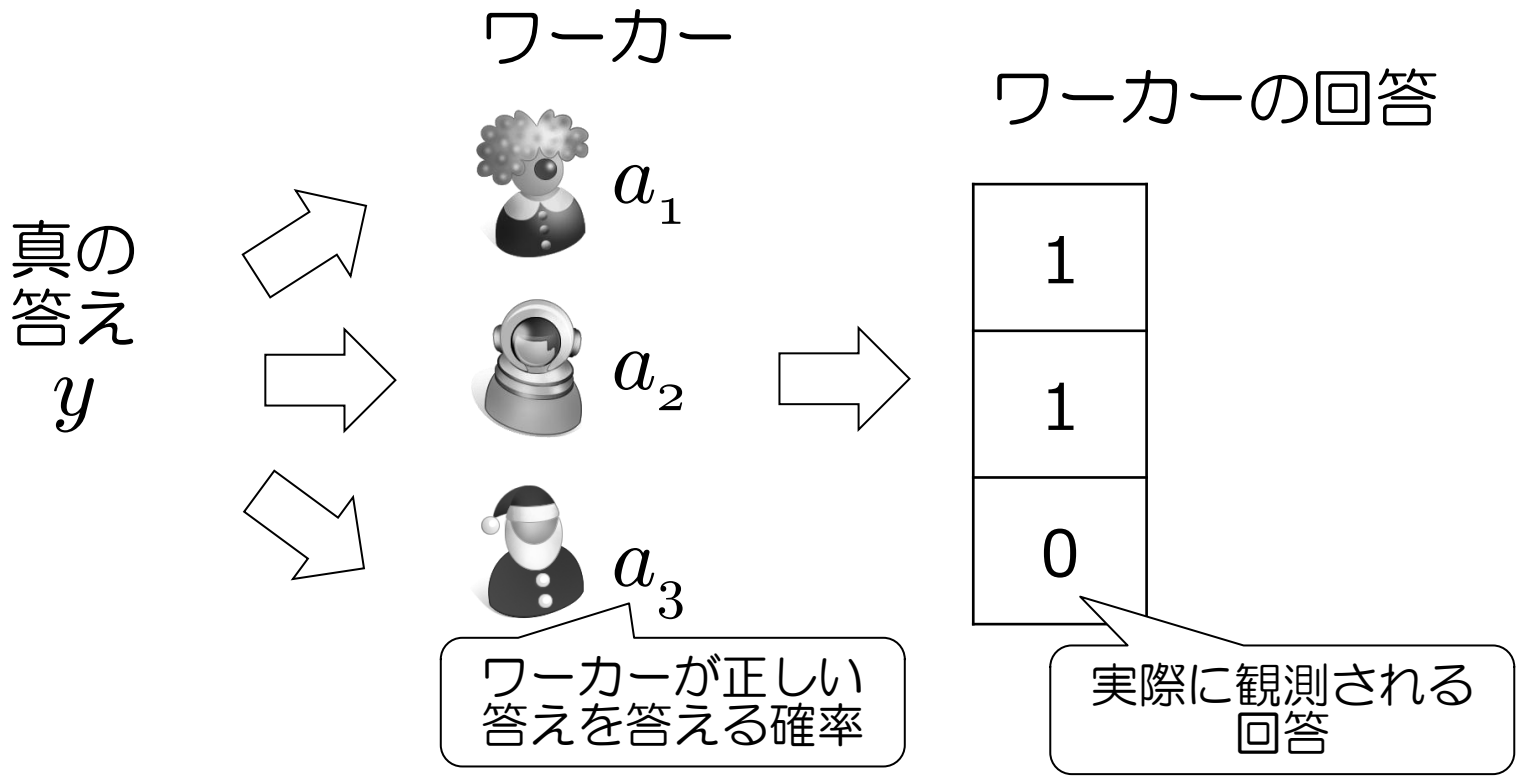
Dawid&Skene (1979)の 先駆的な研究：各ワーカーの能力をモデル化し、真実を推定します

- この話題では、おそらくほぼ最古の先駆的研究
- 患者の麻酔適合度を5人の麻酔医が4段階で診断したものから、真実の答え（麻酔適合度）を当てるのが目的
- 問題が１つしかないのであれば、多数決以上の結果は望みにくい。
  一方で、複数の問題があるなら、つねに多数派に入っているワーカーは信頼度が高そうだ
- ワーカーの能力をモデル化する
  - ワーカーが真の答えと異なる回答をする確率をパラメータ化
  - 真の答えも未知なので、真の答えを潜在変数とし、パラメータとともに推定
- 最近でもこの方法のバリエーションがいくつも提案されている

Maximum Likelihood Estimation of Observer Error-rated using the EM Algorithm
Dawid, A.P., Skene, A.M., Applied Statistics, 28(1), pp.20-28, 1979

Dawid&Skeneの先駆的な研究：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です

- 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



最近の発展：
モデルの直接推定、繰り返し発注、問題難易度のモデル、問題選択

- 最近、４つの大きな発展：
  1. モデルの直接推定 （ Raykar et al. (2010) ）
  2. 繰り返しソーシング （ Donmez et al. (2009) ）
     - 繰り返し仕事を発注するような状況において、有能なワーカーを素早く見つけ、彼らに仕事をまかせる（発注コスト／質を安くする）
  3. 「問題の難しさ」の導入 (Whitehill et al.(2009))
     - ワーカーの能力だけでなく、個々の問題の難しさをモデルに導入
  4. 問題の選択 (Yan et al.(2011))
     - 誰にどの問題を解かせれば予測性能が上がるかを決定する

Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit. Raykar, V.C . et al., In ICML, 2009
Efficiently Learning the Accuracy of Labeling Sources for Selective Sampling. Donmez, P. et al., In KDD, 2009.
Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise. Whitehill, J. et al., In NIPS, 2009.
Active Learning from Crowds, Yan, Y. et al., In ICML, 2011.

モデルの直接推定により、データをより有効活用できるようになりました

- 機械学習におけるクラウドソーシング利用の本来の目的は、予測モデル
- 前述の方法は、まず正しいラベルを推定して、これを用いて機械学習を行う
  - 正しいラベルの推定には入力情報は用いない
  - ひとつのデータに複数のラベルがついている必要がある
- モデルの直接推定 （ Raykar et al. (2010) ）
  - 直接モデルパラメータを学習する
  - ひとつのデータにひとつのラベルでも適用可能
- 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っい、真実の答えを推定する
  2. （推定した）真実の答えを用いてモデルを推定する
  3. モデルの出力に近いワーカーの信頼度を上げる

Supervised Learning from Multiple Experts: Whom to trust when everyone lies a bit.
Raykar, V.C. et al., In ICML, 2009

現存するアプローチの問題点：
タスクの均一性と最適化問題としての不安定性

- 現存するアプローチの２つの問題点
  1. 単一のタスクのみを対象としている
     - 発注側は複数のタスクを同時に依頼したい
     - 一方、ワーカー側も複数のタスクに同時に取り組む
     - 現状では、タスクはすべて独立に扱われる
  2. 実は、真実のラベルを経由する方法は、問題の構造としてはスジが悪い
     - いわゆる凸最適化（最適解が求まるタイプの問題）になっていない

⇨

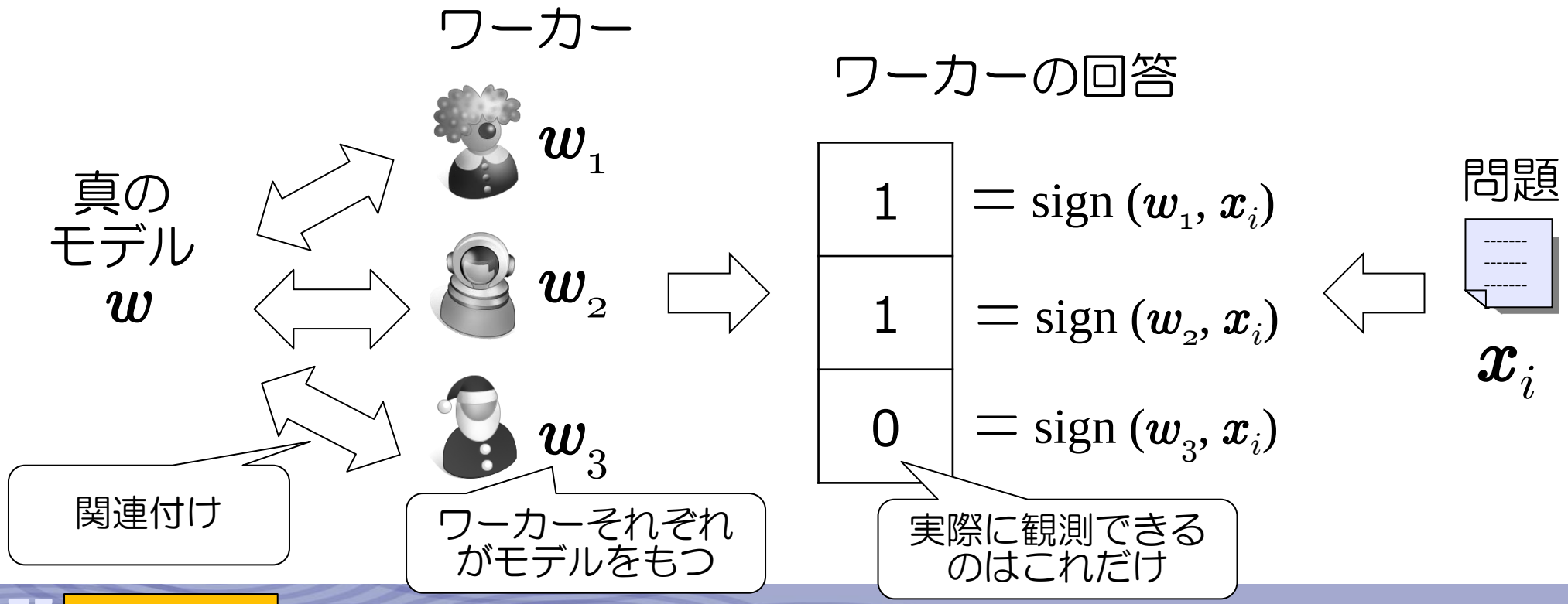
– 最適解が求まるような定式化にしたい

IBISワークショップ
ポスター奨励賞

Convex formulations of multi-task learning from crowds. Kajino, H. & Kashima, H., 2011. (submitted)

ワーカーのパーソナルモデルを導入することで、真の答えの推定ステップを省き、凸最適化問題を導きます

- 個々のワーカーが別々のモデルをもち、これを用いてラベル付を行うとする
- 真のモデルと個々のワーカーモデルは関連があるとする（互いに引き寄せられる）としてすべてのモデルを同時推定
- 我々のモデルでは、真の答えの推定ステップが入らない（凸最適化）



削除

削除

まとめ

- 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
- 個々のワーカーのモデルを導入することで、より振る舞いのよい定式化が可能になる

機械学習界で近ごろ注目の話題（＋手前味噌）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
   - 大量の低品質データへの挑戦
   - 低品質ラベルを対処するためのワーカーのモデル化