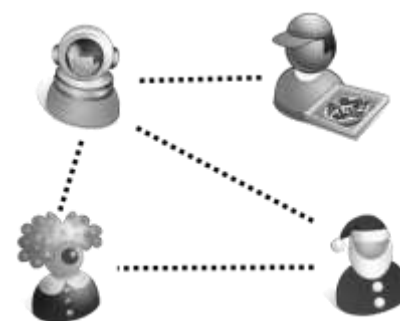


ビッグデータに立ち向かう機械学習

ネットワーク分析のための機械学習

～ 予測モデルを中心に ～



鹿島 久嗣



本チュートリアルの概要： ネットワーク構造をもつデータを扱う予測手法を概観

- ネットワーク構造をもったデータを扱う**予測問題**を、

{内部, 外部}ネットワーク × {ノード, リンク}推論

の4通りに分類

- それぞれに対する**基本的なアプローチとモデル**を解説
 - カーネル法、パタンマイニング、条件付確率場、ラベル伝播
行列／テンソル分解、マルコフネットワーク、ブロックモデル
 - 各モデルがネットワーク構造の何を仮定しているか
 - 推定方法やアルゴリズムには深入りしない

機械学習：

データにもとづく分析や予測のための道具

- データを有効に活用するためのデータ解析技術への注目
 - 機械学習、統計、データマイニング、データ工学、...
 - クイズ王に勝利したAI
 - 研究者だけではなく、産業界もその可能性に注目
 - 「データサイエンティスト」の出現
- データ解析手法の大別：
 - 発見（分析） 的な解析「いま何が起きているのか？」
 - 予測的な解析「これから何が起こるのか？」

予測的な解析： 意思決定に直接的に結びつく技術

- 予測的な解析「これから何が起こるのか？」
- 予測は直接的に競争力のある意思決定に結び付く
 - 人が病気にかかりやすいかどうかの予測 → 生命や健康
 - 顧客が商品を購入するかどうかの予測 → 経済的利益

データ解析の潮流：

個々のデータからデータ間の関係（ネットワーク）へ

- 従来：「個々のデータを対象とした解析」から
- 近年：「データの間関係の解析」へと移行しつつある
- 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：顧客と商品の間関係
 - 創薬スクリーニング：薬剤と標的の間関係
- データ間関係（ネットワーク）に注目することで、
個々のデータに注目するだけでは見えない性質が見える

ビッグデータ解析のチャレンジ:

(ネットワーク構造など) 多種多様なデータ形式の解析

■ ビッグデータの3つの「V」:

1. Volume (莫大なデータ)
2. Velocity (高速度でのデータ入出力)
3. Variety (多種多様なデータ源と種類)

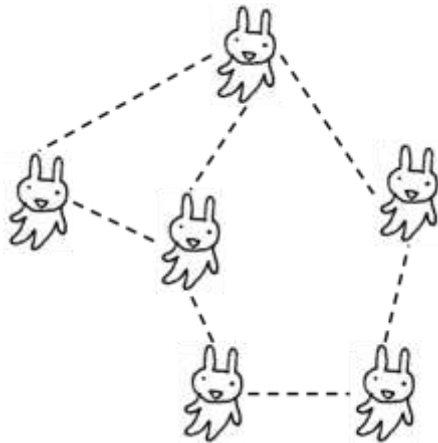
ネットワーク

- 機械学習はビッグデータ解析の主要ツール
 - ネットワーク解析技術の開発は盛ん

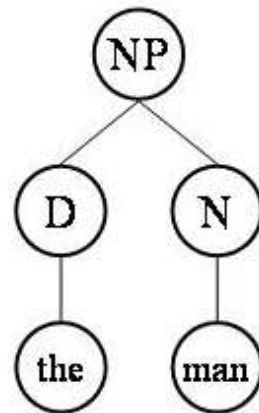
ネットワーク構造をもつデータ

ネットワーク構造をもったデータ: さまざまな分野で登場するグラフ構造データ

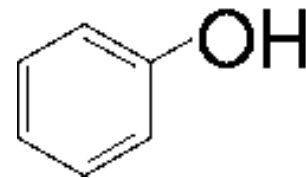
- ネットワーク構造をもったデータ
= グラフ（木、配列）によって表現されるデータ
 - 文書、構文木、Web、XML/HTML、化合物
ソーシャルネットワーク、DNA／タンパク質配列、RNA
生体ネットワーク、引用関係、企業間取引、...



ソーシャルネットワーク



構文木



化合物

AGCTCGAA...

DNA配列

ネットワークデータ解析における困難： 非ベクトル形式データの扱いは自明でない

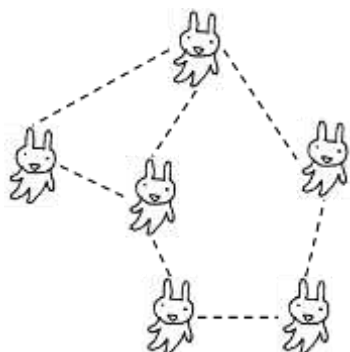
- 多くの手法ではベクトル（表）形式のデータを仮定

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	...

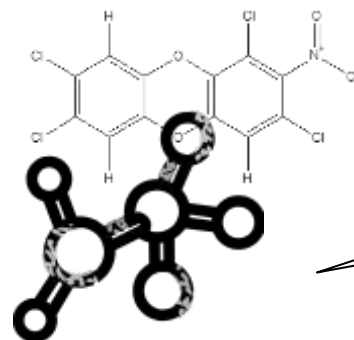
ベクトル

- 一方、非ベクトル形式データの扱いは自明ではない

— ネットワーク構造はその最たるもの



隠れた人間関係
を発見したい

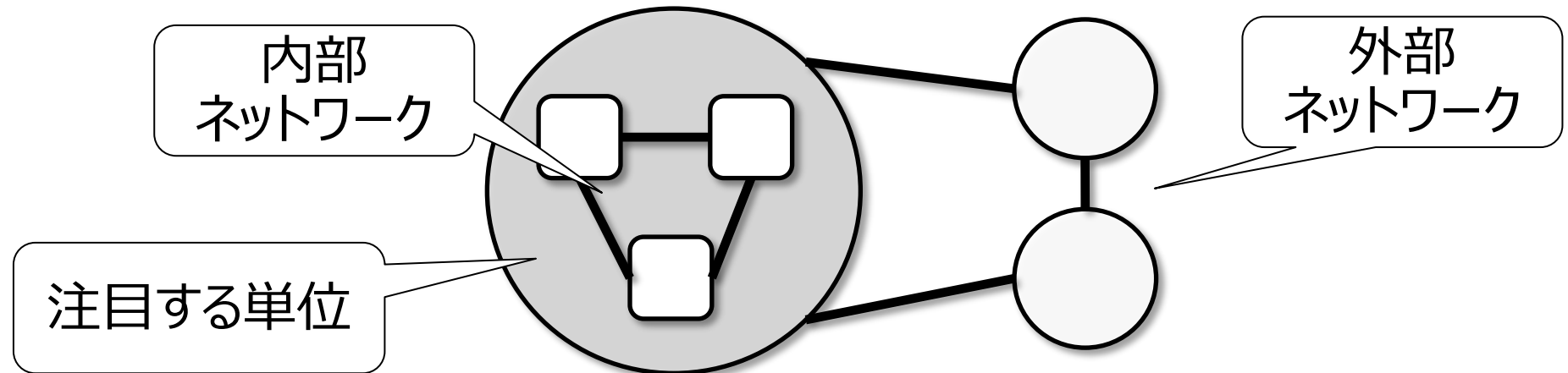


化学的特性
を知りたい

ネットワーク構造の分類：

注目するデータ単位の内外のネットワーク構造の2種類

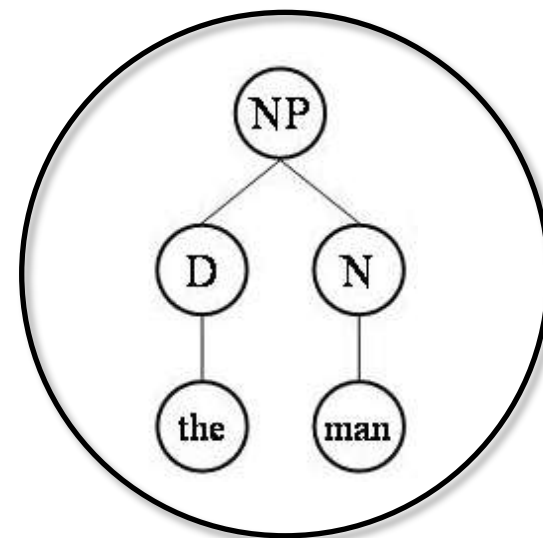
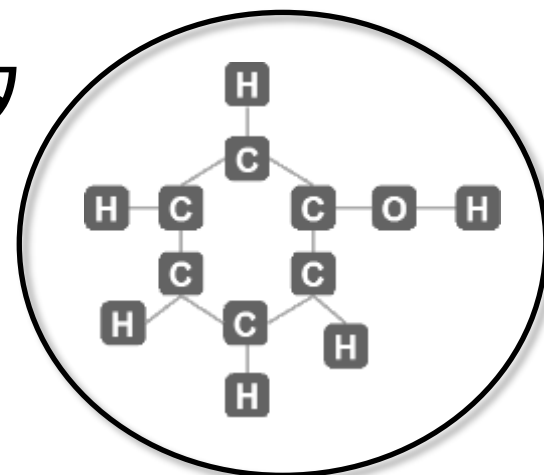
- ネットワーク構造には 2 つの種類が存在する：
 1. 内部ネットワーク：
注目するデータ単位の内側にあるネットワーク
 2. 外部ネットワーク：
注目するデータ単位の外側にあるネットワーク



内部ネットワーク：

注目するデータ単位の内側にあるネットワーク構造

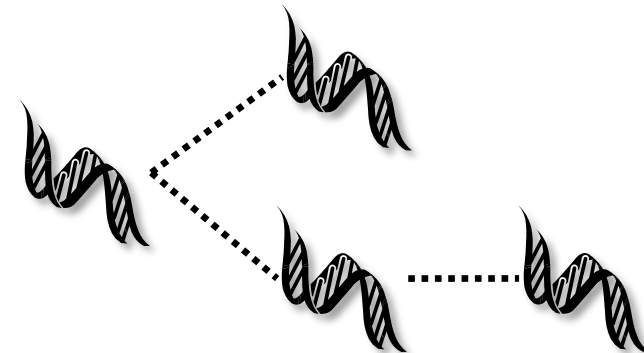
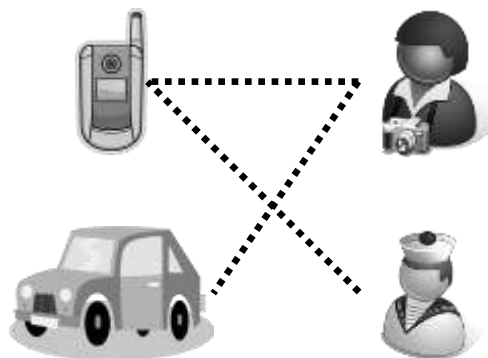
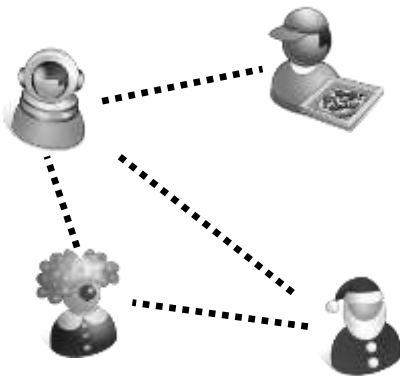
- 活性予測：化合物に注目すると
化合物は内部にグラフ構造をもったデータ
- 文書分類：文書に注目すると
文書は内部に配列構造をもったデータ
- 構文解析：文に注目すると
文は内部に木構造をもったデータ



外部ネットワーク：

注目するデータ単位の外側にあるネットワーク構造

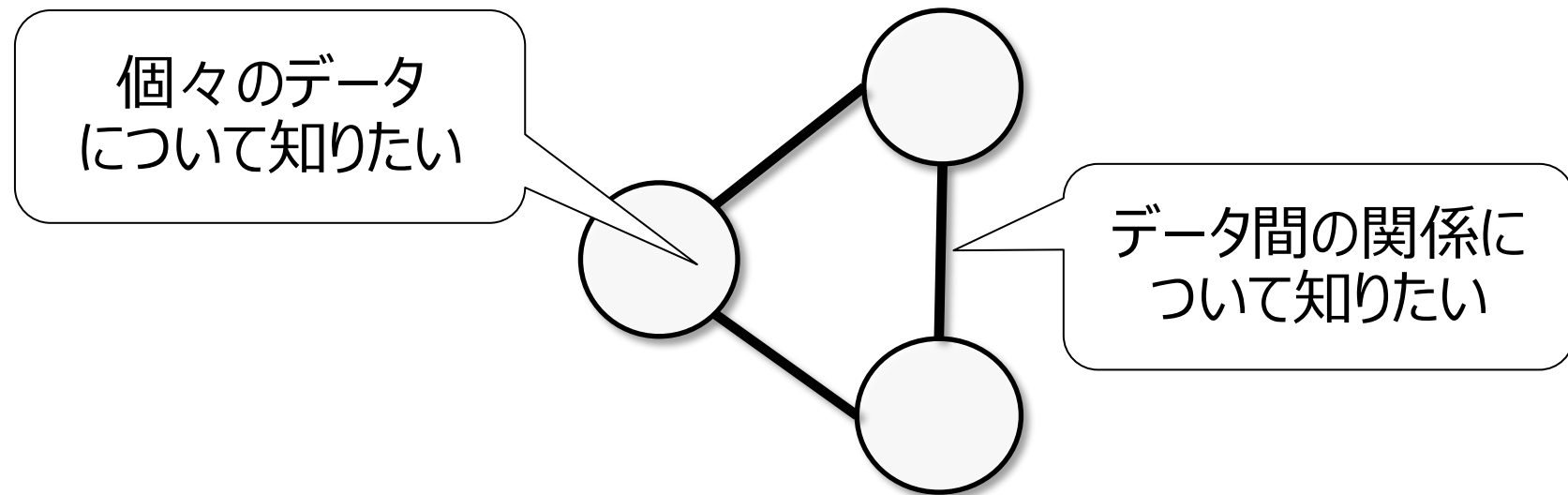
- 友人推薦：人に注目すると
ソーシャルネットワークは外部にグラフ構造をもったデータ
- 推薦システム：顧客と商品に注目すると
購買データは外部に2部グラフ構造をもったデータ
- 系統樹推定：遺伝子に注目すると
系統樹は外部に木構造をもったデータ



解析のフォーカス：

個々のデータに興味がある \Leftrightarrow 内外の関係に興味がある

- 解析のフォーカスとしても2種類ある
 1. 個々のデータの性質に興味がある
 2. データ内外の関係について興味がある



ネットワーク構造解析の世界観：

2 × 2 の 4 通りの分類

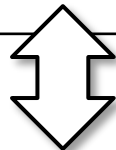
- {内部, 外部}ネットワーク
× {個々のデータ, 内外の関係}についての推論 の4通り

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォ ー カ ス	個々の データ	<ul style="list-style-type: none">• 予測• クラスタリング	<ul style="list-style-type: none">• 予測• クラスタリング• ランキング
	データ 内外の 関係	<ul style="list-style-type: none">• パタン発見• 構造予測	<ul style="list-style-type: none">• リンク予測• 構造変化解析

ネットワーク構造解析のアプローチ： それぞれの分類にそれぞれのアプローチ

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォールカス 解析の	個々のデータ	<ul style="list-style-type: none">カーネル法パタンマイニング	<ul style="list-style-type: none">ラベル伝播マルコフネットワーク行列／テンソル分解確率的ブロックモデル
	データ内外の関係	<ul style="list-style-type: none">パタンマイニング構造学習器 (HMM、CRF等)	<ul style="list-style-type: none">リンク指標ペアワイズ予測マルコフネットワーク行列／テンソル分解確率的ブロックモデル

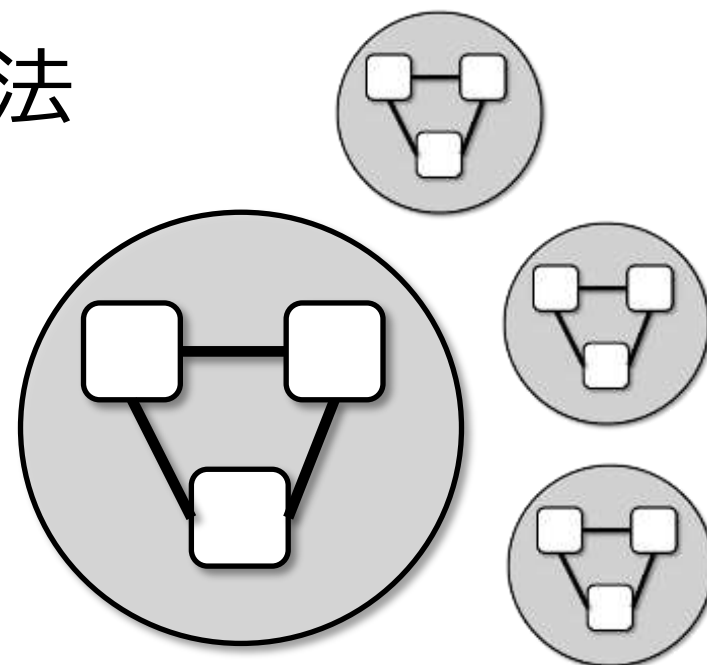
ネットワーク構造解析のためのモデル： 線形識別モデルと潜在変数モデル

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォーカス 解析の	個々のデータ	部分構造に注目した 線形モデル の拡張	ラベル伝播 「隣同志は似ている」
	データ内外の関係		 潜在変数モデル 「付き合い方の似ている同志は似ている」

内部ネットワークをもつデータの解析

内部ネットワークを持つデータの解析： グラフ分類、クラスタリング、構造予測、...

- 個々のデータにフォーカスした解析法
 - 問題： グラフ分類問題、グラフクラスタリング、...
 - アプローチ： パターンマイニング法、カーネル法
- データ内の関係にフォーカスした解析法
 - 問題： 構造予測問題
 - アプローチ： 条件付き確率場



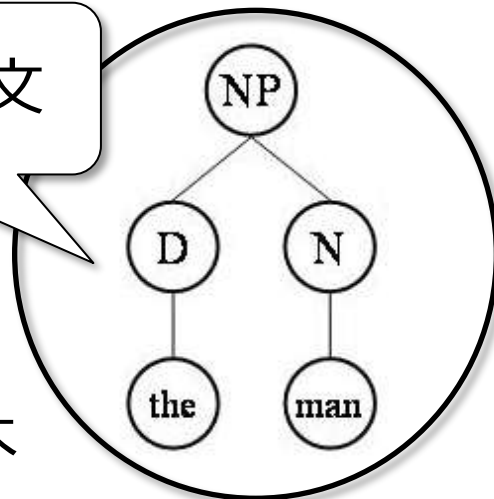
個々のデータにフォーカスした解析法

個々のデータにフォーカスした解析： 内部ネットワーク構造をもつデータの分類、クラスタリング等

- データ： 内部ネットワークをもつデータの集合
- それぞれのデータのもつ性質に興味がある
 - 分類： 各データのもつ性質を予測
 - クラスタリング： データを類似度によってグループ分け

クイズの問題文

構文解析木



分類：
質問文のジャンル

クラスタリング：
似た質問文のグループ

分類のためのモデル：

線形識別モデルはすべての基本

- 素性ベクトル：データを、その特徴量を列挙した D 次元の実数値ベクトル $\boldsymbol{x} = (x_1, x_2, \dots, x_D)^\top$ で表現
- 入力 \boldsymbol{x} に対して出力 $\{+1, -1\}$ を予測する識別モデル f

$$f(\boldsymbol{x}) = \text{sign}(w_1x_1 + w_2x_2 + \dots + w_Dx_D)$$

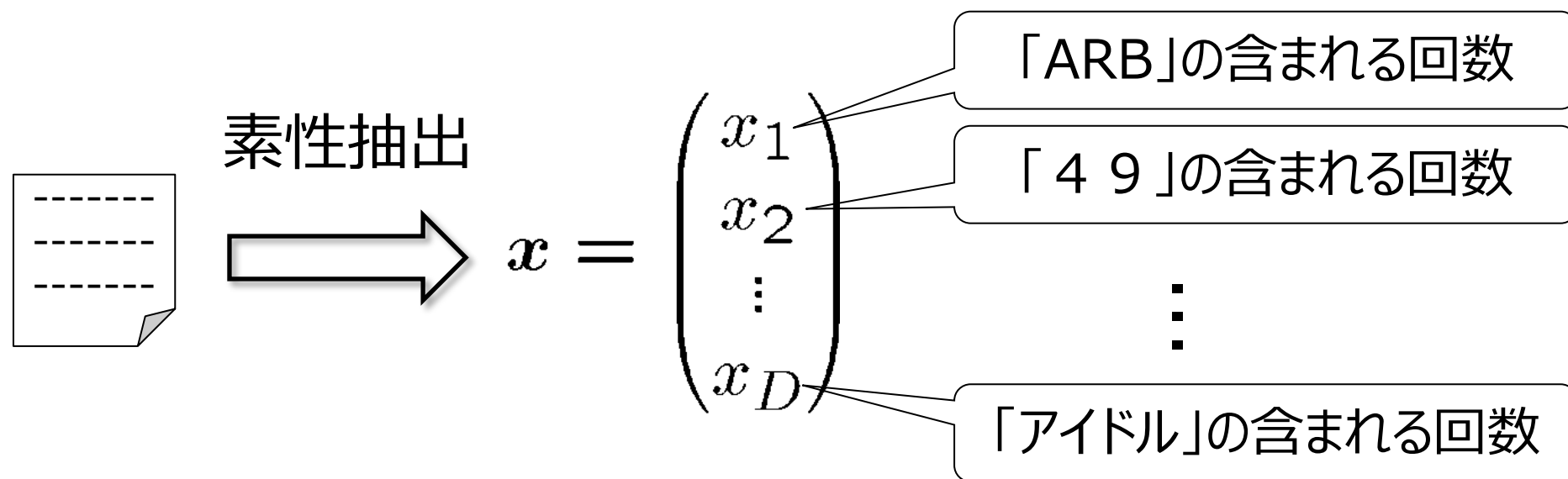
符号をとる

— モデルパラメータ $\boldsymbol{w} = (w_1, w_2, \dots, w_D)^\top$:

- w_d は x_d の出力への貢献度
- $w_d > 0$ なら出力+1に貢献 ; $w_d < 0$ なら出力-1に貢献

文書の素性ベクトル表現： 単語の入った単語袋

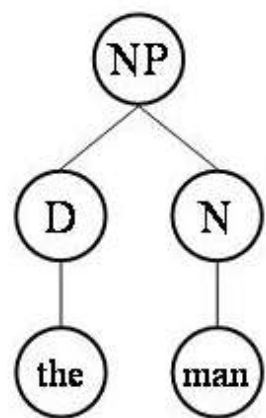
- 文書を、含まれる単語によって素性ベクトル化する



単語袋 (bag-of-words) 表現

内部ネットワークデータの素性ベクトル表現： それは自明ではない

- 内部ネットワーク構造をもったデータの素性ベクトル化
 - 部分構造を用いるのは自然だが、数が多い



素性抽出



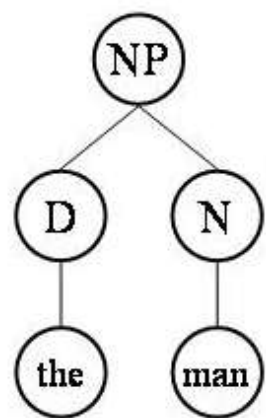
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{pmatrix}$$

素性の定義は自明でない

内部ネットワークデータの素性ベクトル表現： 部分構造を用いるには数が多い

- 部分構造を用いるのはひとつの自然な考え方
- だが、数が多い

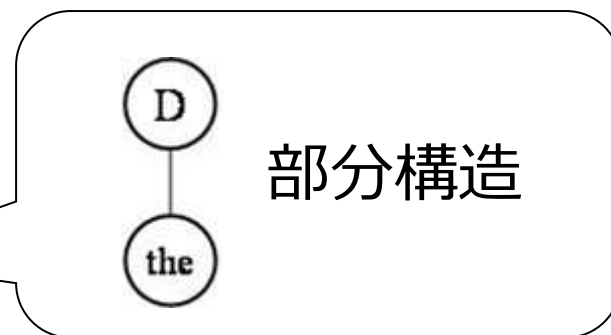
— グラフの部分構造は指数個



素性抽出



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



組み合わせ的に数が増加

部分構造袋表現

内部ネットワークデータへの2つのアプローチ： パターンマイニング法とカーネル法

- 2つのアプローチがとられる

1. パターンマイニング法

- 部分構造を「重要なもの」に限定する

2. カーネル法

- 類似度ベースのモデル

パタンマイニング法：

重要な部分構造だけを素性として用いる

- パタンマイニングは重要な部分構造だけを取り出す
 - 取り出した部分構造を素性として用いる
 - 指数的に多い候補のなかから効率的に見つける
- 「重要」の定義：
 - 頻出パタン：
データベース内に何度も現れる（構造の「単語」）
 - 相関パタン：
目標の出力と相関がある（予測精度に貢献する）

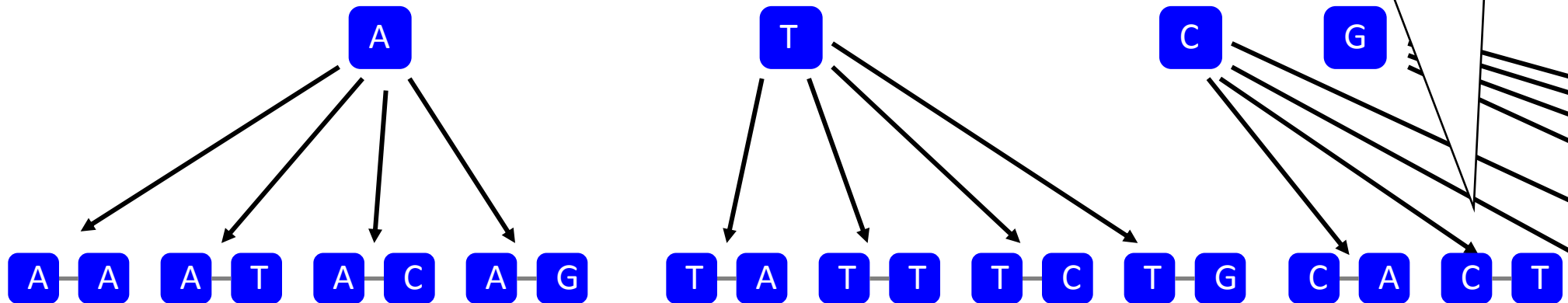
パタンマイニング法における計算の工夫： 探索空間の構造化

スキップ

- チェックする順番を工夫してムダをなくす
 - 小さい部分構造から大きい部分構造へ数える
 - 同じ部分構造を何度もチェックしないようにする

なるべく木構造となる
ような探索空間にする

各部分構造は 1 回
しかチェックされない

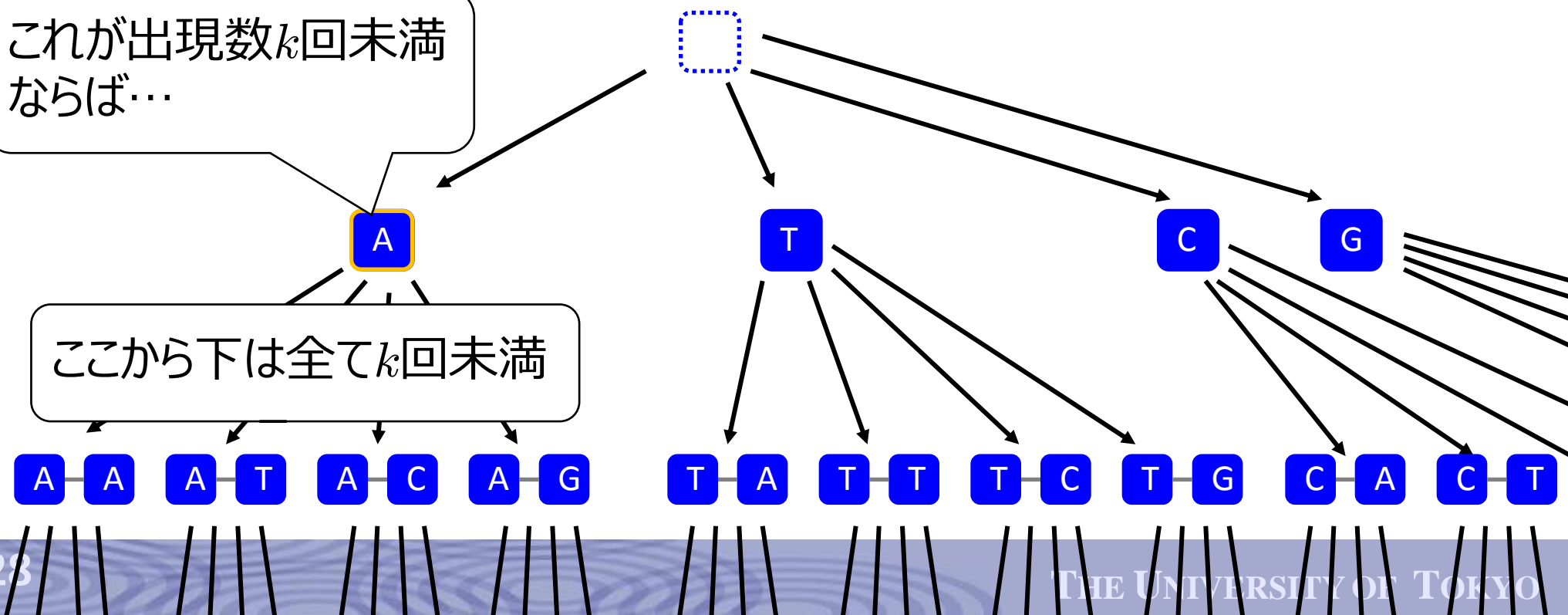


パタンマイニング法における計算の工夫： 枝刈りによる探索空間の縮小

スキップ°

- 枝刈りによって重要な部分構造を効率的に探索
- 頻出パタンの場合：
出現回数の非増加性を用いて探索空間を狭める

これが出現数 k 回未満
ならば...



カーネル法： 類似度ベースのモデル

- 構造データ s に対して出力 $\{+1, -1\}$ を予測するためのカーネル識別モデル f は カーネル関数 k を用いて

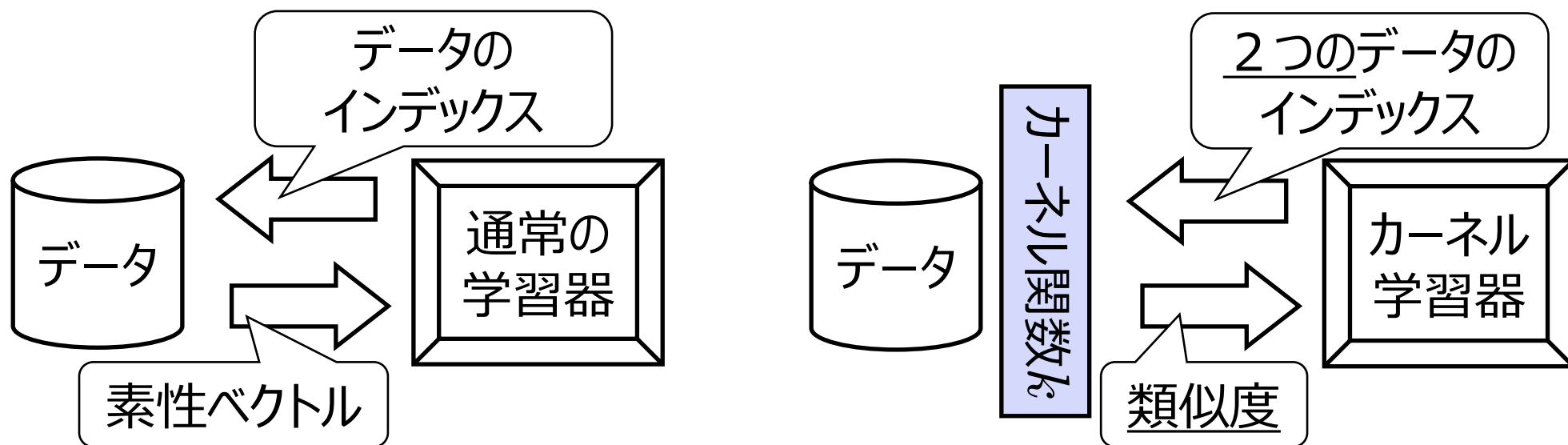
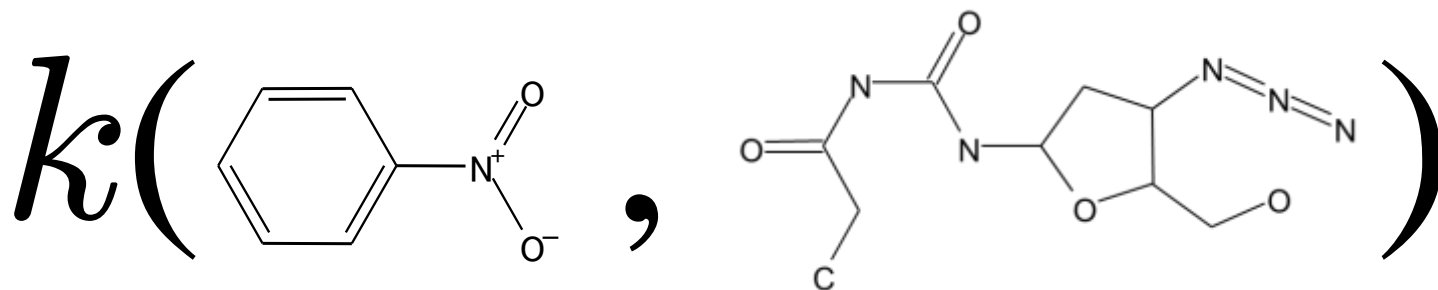
$$f(s) = \text{sign}(\alpha_1 k(s, s_1) + \cdots + \alpha_N k(s, s_N))$$

- カーネル関数 $k(s, s')$: 2つのデータ s と s' の類似度
- モデルパラメータ : $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_N)^\top$ (N 次元)
 - α_i は i 番目のデータとの類似度の出力への貢献度
 - $\alpha_i > 0$ なら出力 $+1$ に貢献 ; $\alpha_i < 0$ なら出力 -1 に貢献

カーネル法のよいところ：

カーネル関数があれば素性の次元数を回避できる

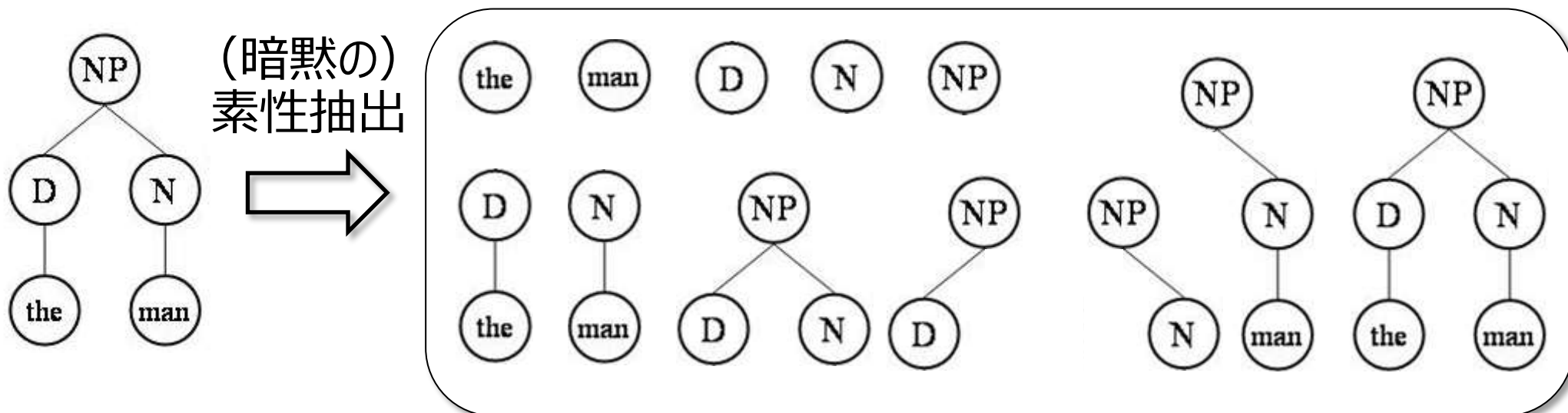
- カーネル法は高次元の素性ベクトルを陽に構成せずとも2つの構造データのカーネル関数だけあれば動く



カーネル関数の例：

構文木に対するカーネル関数は 共通部分構造の個数

- 構文木を部分構造に（心の中で）分解する



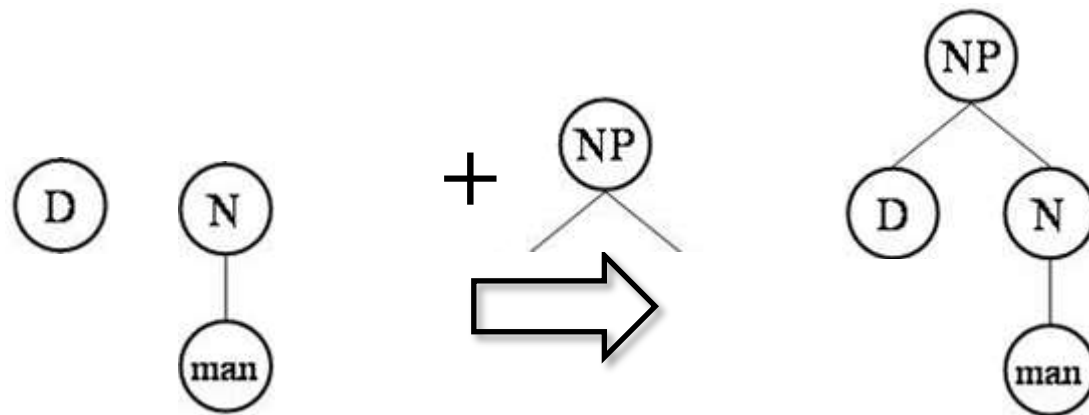
- 2つの木のカーネルは共通の部分構造数

$$k(\text{木}, \text{木}') = \text{共通の部分構造数}$$

構文木カーネルの計算：

動的計画法で効率的に計算可能共通部分構造の個数

- 素朴な計算は、部分構造の数に比例した時間がかかる
 - 部分構造数は非常に多い
- 部分構造の再帰性に着目した動的計画法



- 計算量は2つの木のサイズの積
(部分構造の数よりもずっと少ない)

様々なカーネル関数：

各種内部ネットワーク構造に対するカーネル関数がある

- 様々なカーネル関数：

- グラフ、ハイパーグラフ、順序木、無順序木、配列、...

- ポイントは、部分構造のクラスとアルゴリズムの兼ね合い

- グラフカーネルは パスを部分構造にとることで計算可能

- 最近では線形時間で計算可能なものが主流

- ハッシュ関数を用いた高速化

- 接尾辞配列等の効率よいデータ構造

カーネル法の問題点：

学習後のモデル適用時の計算量がデータ数に依存

- カーネル法のモデルの予測値の計算はデータ数 N に比例した計算量がかかる

$$f(s) = \text{sign}(\alpha_1 k(s, s_1) + \cdots + \alpha_N k(s, s_N))$$

- 予測時の高速化（ N に非依存化）は実用上重要：
 - パタンマイニングを用いて、学習後のモデルから通常の線形モデルを抽出
 - 接尾辞配列等のデータ構造を用いて、データを圧縮

まとめ: 内部ネットワークをもつデータに注目した解析では パタンマイニング法とカーネル法 が用いられる

1. パタンマイニング法

- 部分構造を「重要なもの」に限定して素性を構成
- 探索空間を構造づけ、枝刈りによって効率的に列挙
- 連続値ラベルには少し弱い（離散化が前提）

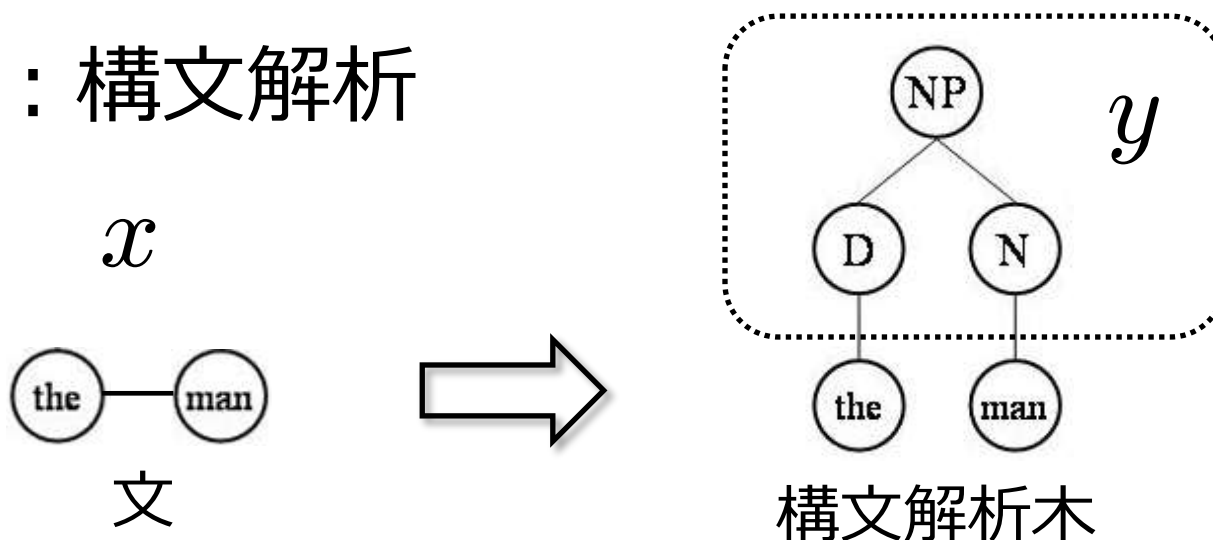
2. カーネル法

- 効率的に計算できる類似度をデザインすることが重要
- グラフ、木、配列など様々な構造に対するカーネル関数
- モデル適用時の高速化が応用上のカギ

個々のデータ内の関係に注目した解析法

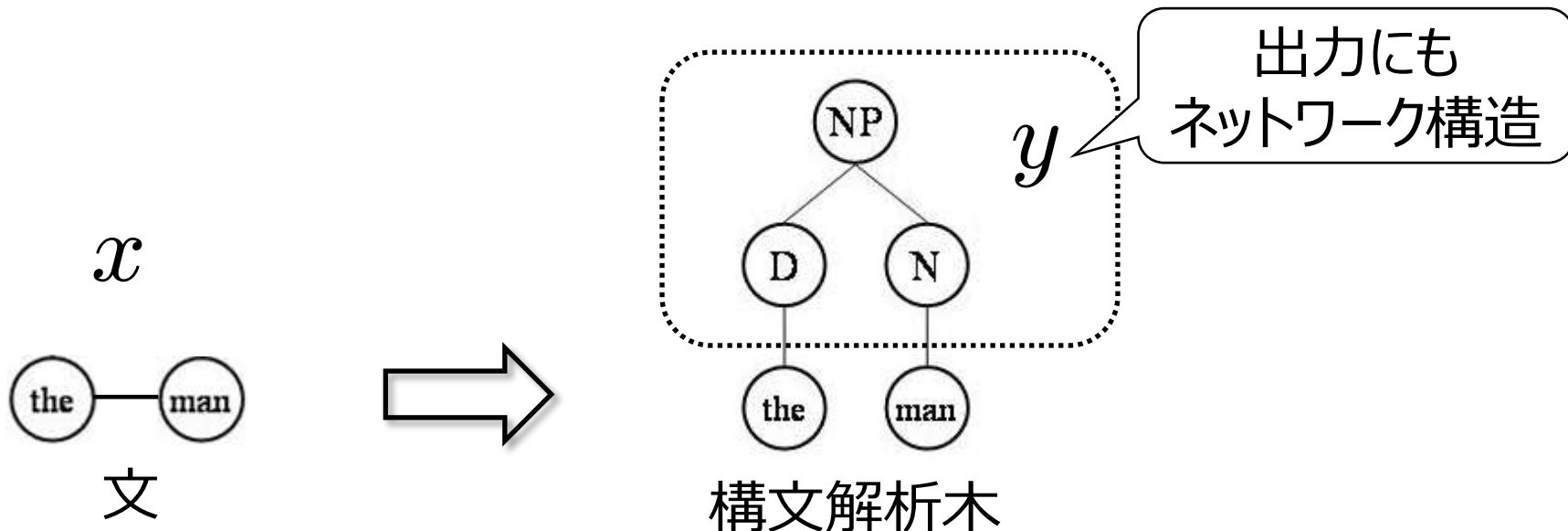
データ内の構造にフォーカスした解析： 内部ネットワークのパターン発見と構造予測

- データ： 内部ネットワークをもつデータの集合
- それぞれのデータのもつ内部構造に興味がある
 - パターン発見： 特徴的な部分構造を発見
 - 構造予測： 各データ内のネットワーク構造を予測
- 例： 構文解析



構造予測問題の難しいところ： モデルの出力候補が指数的に多い

- 前述の線形モデルと異なり、出力 y が一次元でない
 - 出力も内部ネットワーク構造をもつ
- 分類モデルの直接適用は、
指数的に多いクラスのカテゴリ分類問題となる



構造予測のための代表的モデル： 条件付き確率場をはじめとするさまざまなモデル

- 構造予測の代表的モデル：
 - 条件付き確率場（CRF）：確率モデル
 - 構造化パーセプトロン、構造化SVM等

構造予測モデルの考え方：

線形識別モデルによって入出力の組の正しさを識別

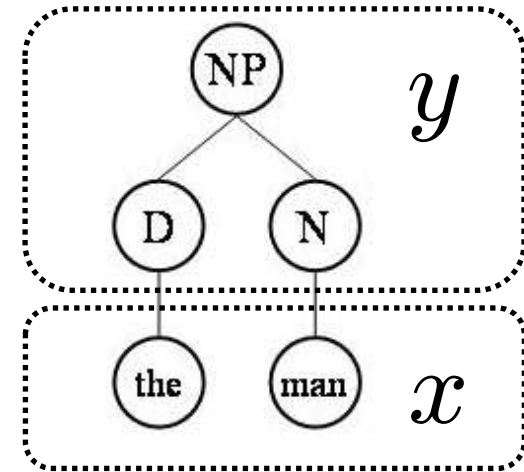
- 素性ベクトル：入出力合わせた素性を列挙した
 D 次元の実数値ベクトル $\phi = (\phi_1, \phi_2, \dots, \phi_D)^\top$ で表現
- 入力 x に対して出力 y を予測するモデル f

$$f(x) = \underset{y}{\operatorname{argmax}} (w_1 \phi_1(x, y) + \dots + w_D \phi_D(x, y))$$

最大となる
 y を返す

入出力の親和度の高さを表す

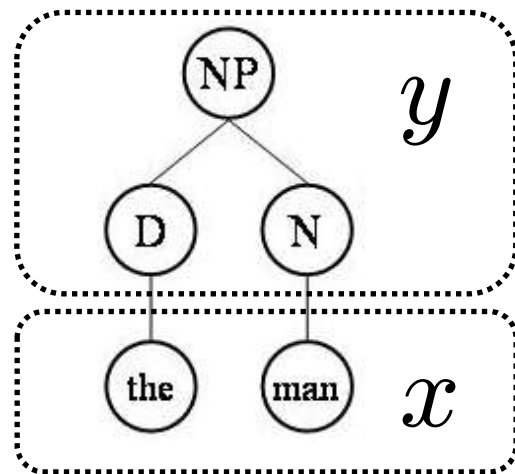
$\phi(x, y)$ ← 素性抽出



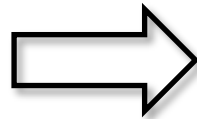
構造予測は困難：

素性を絞ることで、動的計画法等を適用し効率化を図る

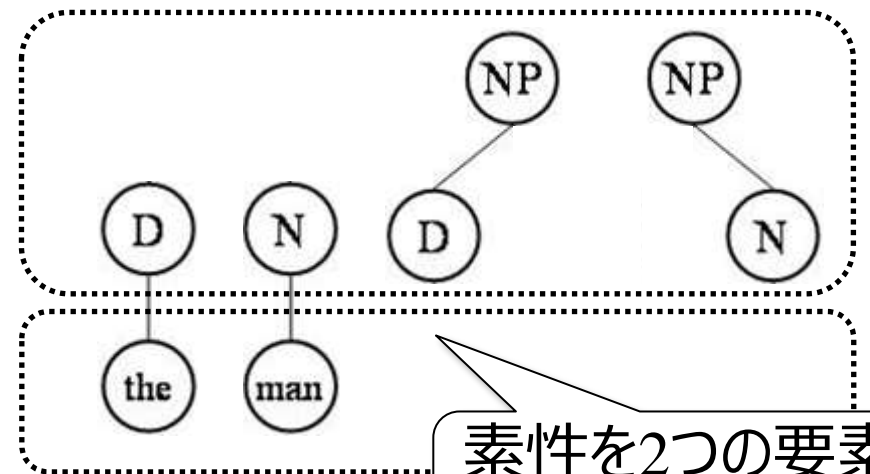
- 指数個の予測候補のなかから 最も入力に適合する 出力を見つけるのは非常に困難
- 素性を2つの要素の組に絞ることで、動的計画法が適用できる場合が多い



素性抽出



$\phi(x, y)$



素性を2つの要素の組に絞る

まとめ:

内部ネットワークの構造にフォーカスした解析

- 構造予測問題：出力も構造をもつ予測問題
- 条件付き確率場（CRF）をはじめとする種々のモデル
 - 入出力の組に対する素性を定義することで線形識別モデルに帰着する
 - 「入出力が組として正しいか？」の識別
 - 素性を絞ることで計算の効率化を図る

外部ネットワークをもつデータの解析

ネットワーク構造解析の世界観：

2 × 2 の 4 通りの分類

- {内部, 外部}ネットワーク
× {個々のデータ, 内外の関係}についての推論 の4通り

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォークス 解析の	個々のデータ	<ul style="list-style-type: none">• 予測• クラスタリング• 構造ラベリング	<ul style="list-style-type: none">• 予測• クラスタリング• ランキング
	データ内外の関係	<ul style="list-style-type: none">• パタン発見• 構造予測	<ul style="list-style-type: none">• リンク予測• 構造変化解析

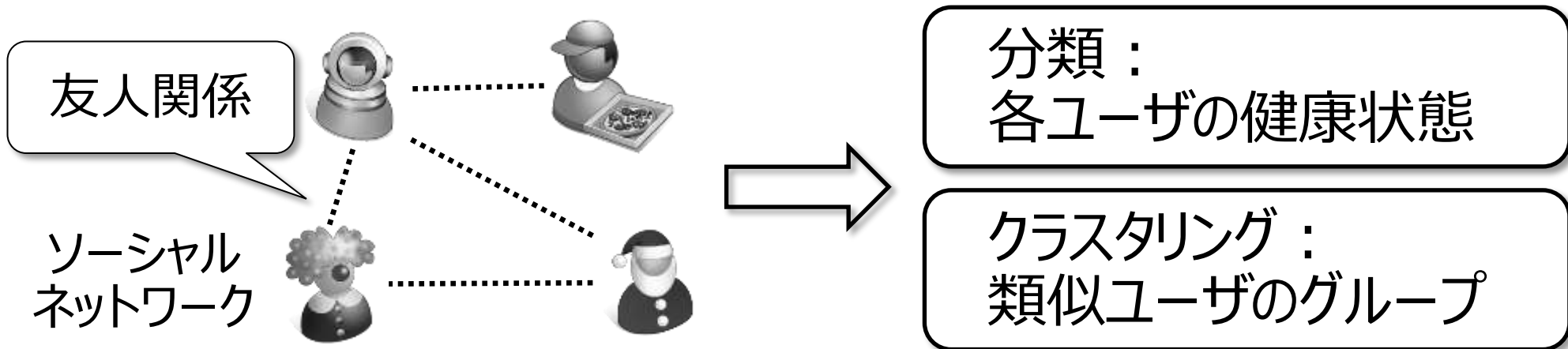
外部ネットワークを持つデータの解析： ノード分類／クラスタリング、リンク予測、構造変化解析、...

- 個々のデータにフォーカスした解析法
 - 問題： ノード分類、ノードクラスタリング、ランキング
 - アプローチ： ラベル伝播法（ノード分類）
- 関係データにフォーカスした解析法
 - 問題： リンク予測、構造変化解析
 - アプローチ： リンク指標、ペアワイズ予測（リンク予測）
- 両者に共通する解析法： マルコフネットワーク
行列／テンソル分解、確率的ブロックモデル

個々のデータにフォーカスした解析法

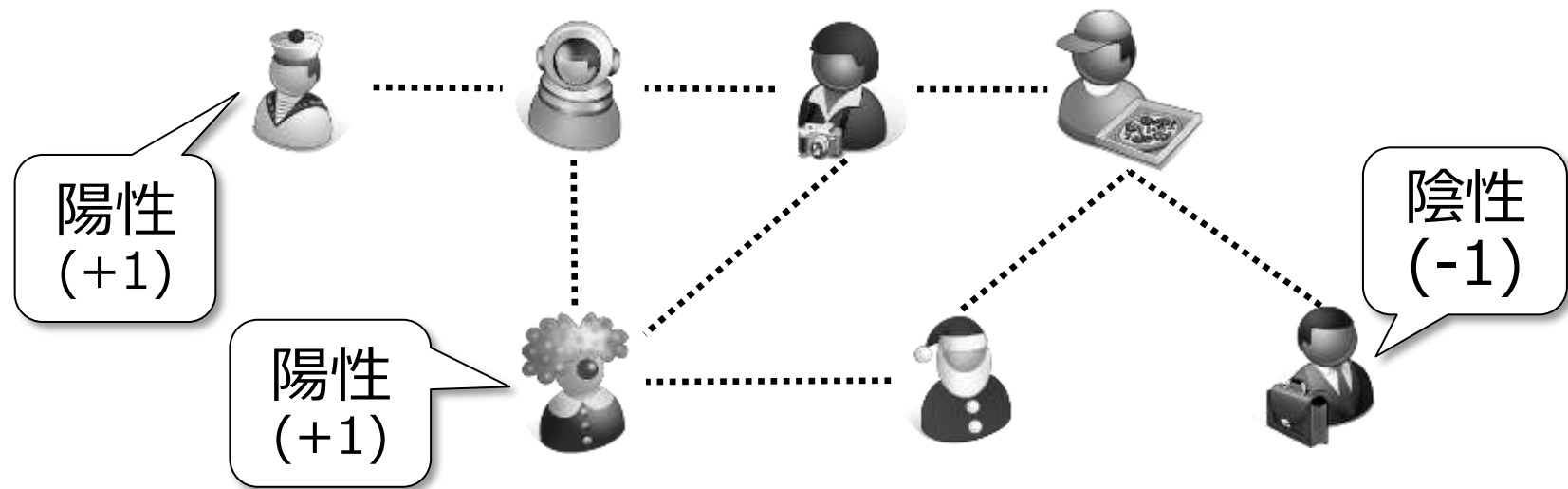
個々のデータにフォーカスした解析： 外部ネットワーク構造をもつデータの分類、クラスタリング等

- データ： 外部ネットワークをもつデータの集合
- それぞれのデータのもつ性質に興味がある
 - 分類： 各データのもつ性質を予測
 - クラスタリング： データをグループ分け
 - ランキング： 重要度でランク付け



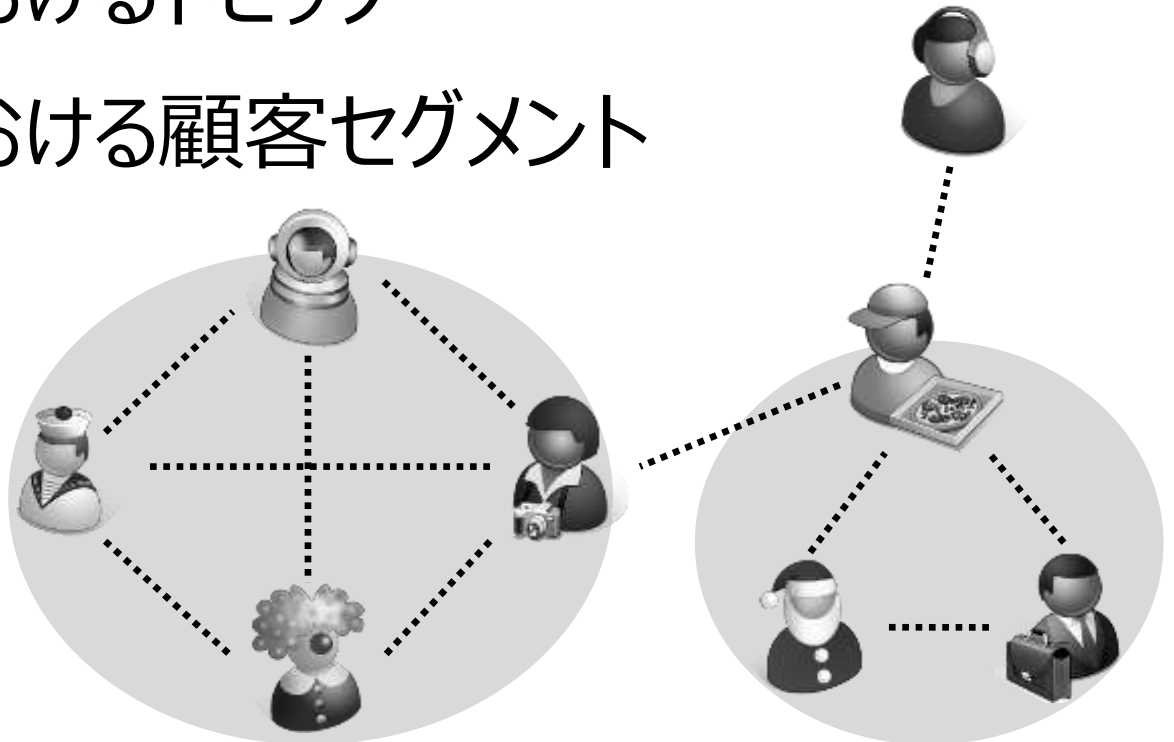
ノード分類問題： ネットワーク上で各ノードのもつ性質を予測

- ネットワーク上のノード分類問題：
 - 入力：いくつかのデータについてのラベル $\{+1, -1\}$
 - 出力：残りのデータについてのラベル予測
- たとえば、ある種の感染症の検査結果



ノードクラスタリング： ネットワークを密な部分に分解する

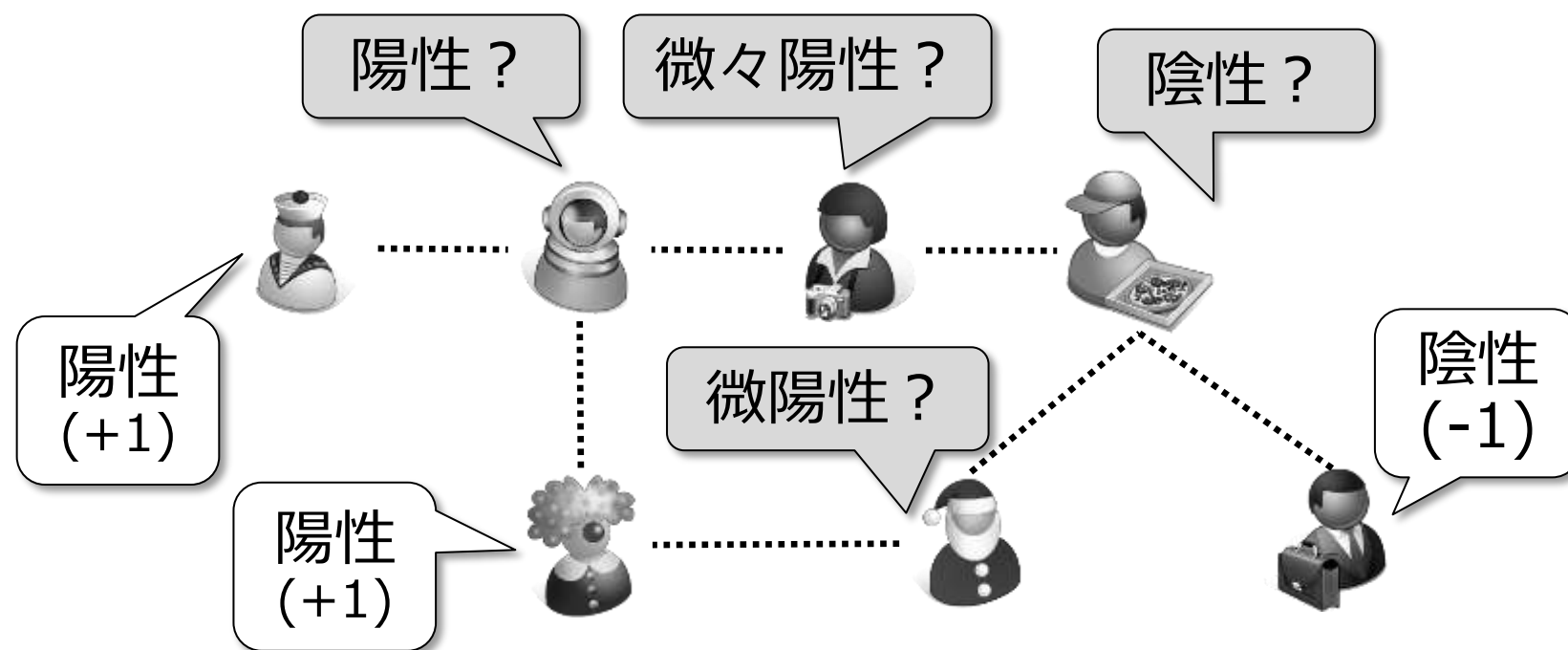
- 密な部分：ネットワークにおいて重要な意味をもつ塊
 - ソーシャルネットワークにおけるコミュニティ
 - 文献ネットワークにおけるトピック
 - 購買ネットワークにおける顧客セグメント
- 例外／異常の発見



ラベル伝播法によるノード分類：

「となり同志は似ている」をもとにノード分類問題を解く

- 仮定：「リンク＝似ている」をもとに未知ラベルを予測
- 隣り合うデータのラベルは同じ可能性が高い



ラベル伝播法の定式化： 最適化問題として定式化する

- 隣同志の予測値が近くなるように予測を最適化

$$\underset{p}{\text{minimize}} \underbrace{\sum_i (p_i - t_i)^2}_{\text{予測値を正解に近づける}} + \lambda \underbrace{\sum_{i,j} A_{i,j} (p_i - p_j)^2}_{\text{隣同志の予測値が近くなるようにする}}$$

予測値を正解に
近づける

隣同志の予測値が
近くなるようにする

- $A_{i,j} \in \{0, 1\}$: ノード i とノード j の隣接関係
- $p_i \in [-1, 1]$: ノード i に対する予測値
- t_i : ノード i の予測の目標値($\{+1, -1, 0(\text{ラベルなし})\}$)
- λ : 2つの項のバランスをとる定数

マルコフネットワーク:

「隣同志は似ている」とは限らない場合のモデル

■ 線形モデルに帰着

予測値

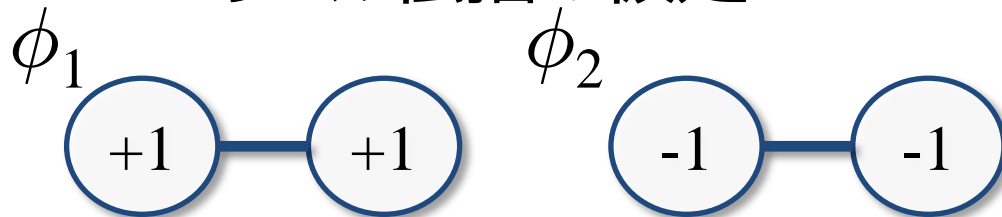
既知ラベル

$$f(t) = \operatorname{argmax}_p (w_1 \phi_1(p, t) + \dots + w_D \phi_D(p, t))$$

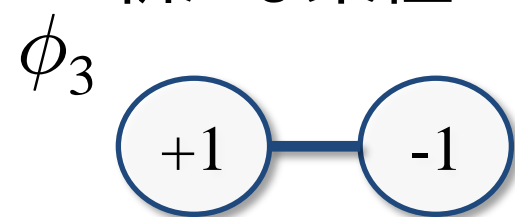
— $p_i \in \{-1, 1\}$: ノード i に対する予測値

— $\phi_d(p, t)$: 隣接ラベルの組み合わせ素性

ラベル伝播の仮定



新たな素性



まとめ: 外部ネットワーク上の推論

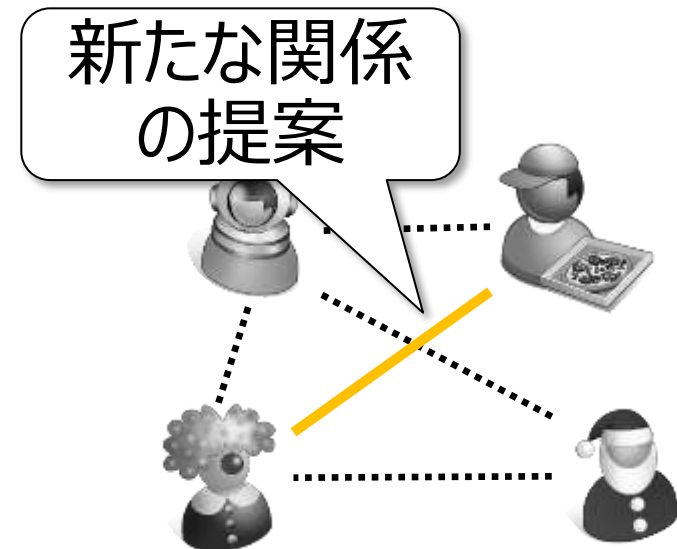
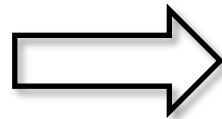
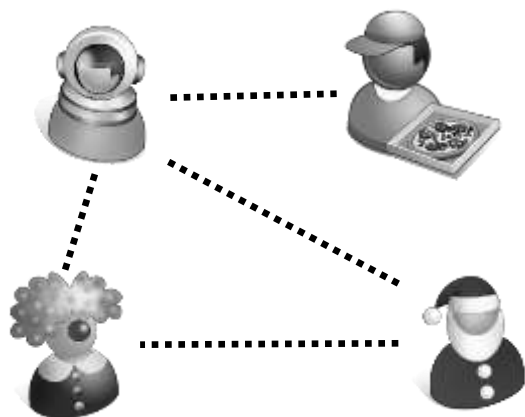
- ネットワーク上の個々のノードに注目した解析
 - ノード分類、クラスタリング、ランキング、...
- 共通する仮定：「隣り合うノードは似ている」
 - ノード分類手法：ラベル伝播法
- より一般的なモデル：マルコフネットワーク

データ間の関係に注目した解析

データ間の関係にフォーカスした解析： 外部ネットワークの構造予測（リンク予測）

- リンク予測： 外部ネットワークの構造予測問題
 - ソーシャルネットワークにおけるつながり推薦
 - オンラインショッピングにおける購買予測
 - 生体分子間の相互作用予測

ソーシャルネットワーク



リンク予測のための手法： リンク指標とペアワイズ予測

■ リンク指標

- 2つのノード間のリンクの張られやすさの指標
- 「学習」ではない

■ ペアワイズ予測

- リンク予測問題を線形モデルに帰着

リンク指標：

ネットワーク上の2つのノードの親和度を表す指標

- 複雑ネットワーク理論などに基づくさまざまなリンク指標
(大きいほどリンクの確度が高い)

— 共通隣接ノード数 (CN) :

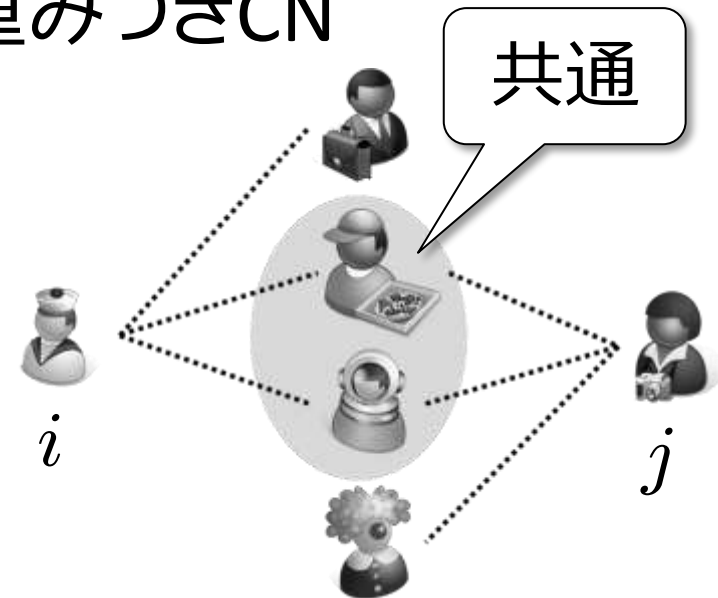
2つのノードが共通にもつ隣接ノードの数

- Adamic/Adar、Jaccard係数：重みつきCN
- Katz：長距離CN

— 優先的選択：隣接ノード数の積

$$A/A = \sum_{k \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(k)|}$$

隣接ノード
集合



補足：リンク指標間の関係

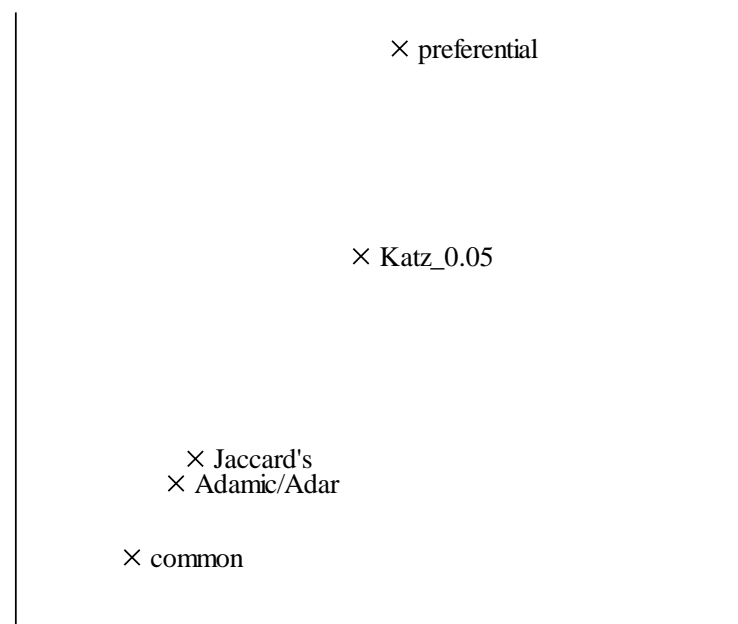
リンク指標間の相関係数の2次元視覚化

■ 各リンク指標間の相関係数

	common	Jaccard's	Adamic/Adar	preferential	Katz _{0.05}
common	1	0.92	0.94	0.31	0.61
Jaccard's	0.92	1	0.97	0.53	0.75
Adamic/Adar	0.94	0.97	1	0.49	0.70
preferential	0.31	0.53	0.49	1	0.84
Katz _{0.05}	0.61	0.75	0.70	0.84	1

— データ：生体ネットワーク

■ 2次元での視覚化 (多次元尺度構成法)



ペアワイズ予測モデル： 線形モデルの2データの組への拡張

- 2つのノードの素性ベクトル x と x' が利用できる場合
- 組み合わせ素性を用いた線形モデル

行列パラメータをもつ

$$f(x, x') = \text{sign} \left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}^\top \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} \begin{bmatrix} x'_1 \\ x'_2 \\ x'_3 \end{bmatrix} \right)$$

$$= \text{sign} \sum_{i,j} w_{i,j} x_i x'_j$$

組み合わせ素性

ペアワイズ予測モデルの解釈： リンク指標のパラメトライズと解釈可能

- ペアワイズ予測をベクトルと行列で書くと

$$f(x, x') = \text{sign}(x^\top W x')$$

- リンクの強さ $x^\top W x'$
- 素性 x_i をノード i が隣接ノードである(1)か否(0)かとするれば
(= 隣接行列の列を素性ベクトルにする)
 - $W = I$ のとき共通隣接ノード数に一致
 - $W = 1$ のとき優先的選択に一致

ペアワイズ予測モデルにおける工夫： 行列パラメータの低ランク化

- 行列パラメータ（組み合わせ素性）はパラメータ数が多い

$$f(x, x') = \text{sign}(x^\top W x')$$

- W の低ランク性を仮定し実効パラメータ数を減らす

$$W = U V^\top$$

パラメータ数 減少

—素性のグループ化の効果

- 次元圧縮

$$\tilde{x} = U^\top x$$

補足：トレースノルム

非凸なランク制約にかわる凸制約

- ランク制約は凸集合を与えない
- トレースノルム：特異値の和
 - トレースノルム制約は凸集合を与える
 - 特異値の L_1 ノルム制約
 - 凸最適化問題を得るために用いられる

特異値の和

$$\text{rank}(\mathbf{Y}) \leq k \iff \|\mathbf{Y}\|_{\text{trace}} = \sum_i \sigma_i(\mathbf{Y}) \leq c$$

マルコフネットワーク: リンク予測のモデルとしても適用可能

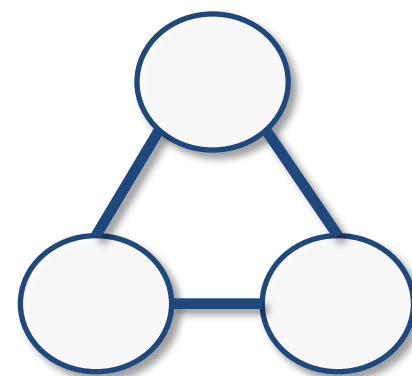
■ リンク予測のためのマルコフネットワーク

$$f(l) = \operatorname{argmax}_p (w_1 \phi_1(p, l) + \dots + w_D \phi_D(p, l))$$

予測リンク

既知リンク

— 素性 ϕ_d : ネットワーク構造のテンプレート



共通隣接ノード指標に対応するような素性

- リンク指標：ネットワーク理論等に基づくノードペアの親和性を測る指標
 - 共通隣接ノード数など
- ペアワイズ分類
 - ノードペアに対する線形モデル
 - 行列パラメータをもつ
 - 低ランク性の仮定によってパラメータを圧縮
- より一般的なモデルとしてマルコフネットワークも

潜在変数モデル:

ノードの潜在的な状態を仮定するモデル

- 前述のモデルは局所的な情報に基づくモデル
 - 2つのノードないし局所的な構造を用いる
- もう少し大局的な情報を見るのが潜在変数モデル
 - 連続的な潜在変数 (行列分解、テンソル分解)
 - 離散的な潜在変数 (確率的ブロックモデル)

ネットワークデータの表現： ネットワーク構造は行列として表現できる

- リンク（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
- 重みつきグラフの場合も




行列データの解析手法： 協調フィルタリングを出発点に潜在変数モデルへ

- 行列の補完問題
- 協調フィルタリングの初等的手法：GroupLens
- 潜在変数モデル
 - 行列の低ランク分解：連続潜在変数
 - 確率的ブロックモデル：離散潜在変数

行列の補完問題：

観測された要素を手掛かりに 未観測要素を予測する

- 行列の見える部分を手掛かりに、
見えていない部分（）を予測する
- 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

顧客



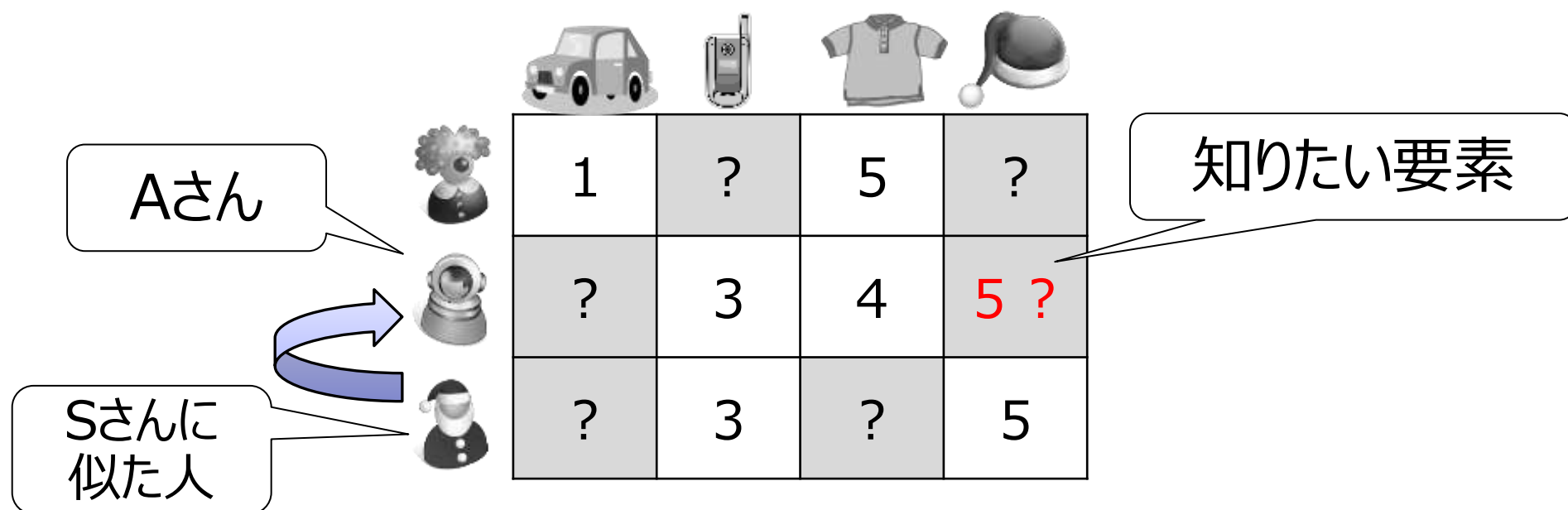


		   	商品	
	1	?	5	?
	?	2	4	?
	?	3	?	5

評価

協調フィルタリングの初期の手法 GroupLens : 自分と似た顧客のデータを用いて予測

- GroupLens : 初期の（ニュース推薦）アルゴリズム
- 予測したい顧客と似た顧客の評価を用いて予測を行う



GroupLensの予測手法：

顧客類似度を相関係数で測り、重みづけ予測








- 2人の顧客の類似度：（観測部分の）相関係数

- 相関係数で重みつき予測

人ごとの平均値

相関係数

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

					
相関係数		1	?	5	3
		?	3	4	4.5
		?	3	?	5

重みつき予測

行列補完における低ランク性の仮定：

GroupLensの予測手法は暗に行列の低ランク性を仮定？

- GroupLensの仮定：
行列の各行が、別の行の重みつき線形和によって表せる
(線形従属)
⇒ 行列がフルランクではない (≡低い)
- 低ランク性の仮定は行列の穴埋めに有効？
 - データよりもパラメータが多い状況では
事前知識を用いて解に制約を設ける必要
 - 低ランク性の仮定は 実質パラメータ数を減らす

行列分解： 行列の低ランク性を仮定

- 低ランク性：行列が2つの（薄い）行列の積で書ける

顧客 X 商品 = U V^T } ランク k

パラメータ数が減少

- U と 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）のグループ

行列の低ランク分解の例： 特異値分解

- 与えられた行列を低ランク行列で近似

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

- 特異値分解

— $X \sim U \begin{matrix} \text{ } \end{matrix} \begin{matrix} \text{ } \end{matrix} V^\top$

対角行列（特異値）

- 制約： $U^\top U = I \quad V^\top V = I$
- $X^\top X$ の固有値を大きい方から k 個とるのが最適

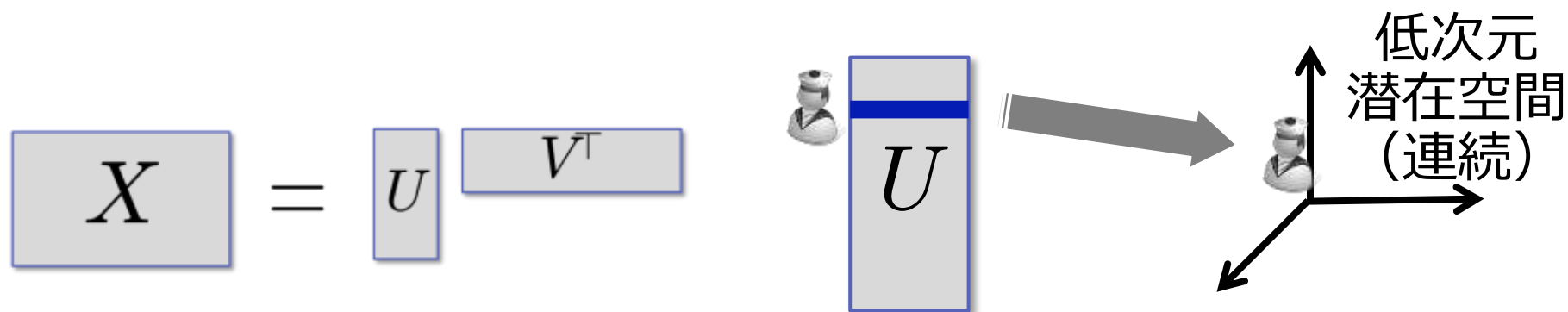
部分観測の場合の解法： EM、勾配法、トレースノルム制約

- 特異値分解は全要素が観測されていることが前提
- 未観測要素がある場合（補完問題）には：
 - ムリヤリ適用：適当に埋めてから分解
 - EM的繰り返し：穴埋め⇒分解 を繰り返す
- 大規模データの場合：
観測部分のみを用いて確率的勾配法
- 凸最適化として解く場合：トレースノルム制約

確率的ブロックモデル：

低ランク分解の離散潜在変数版？

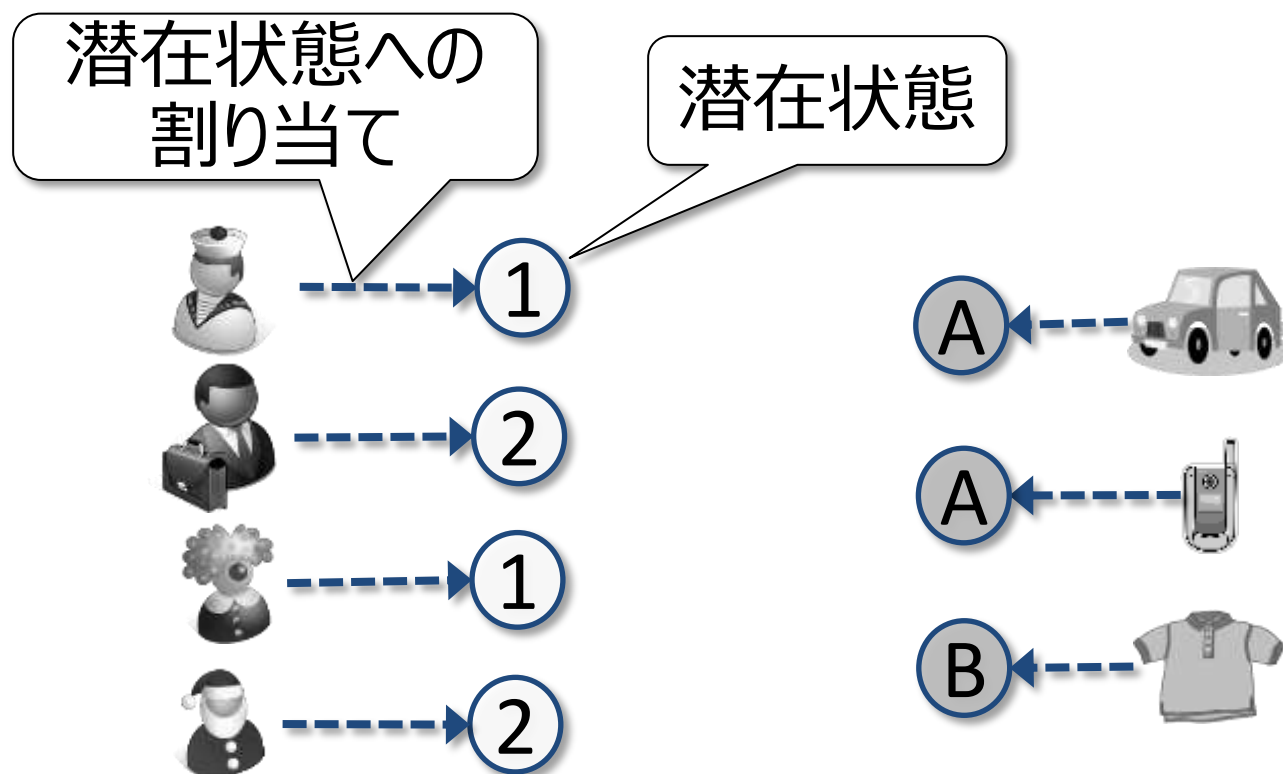
- 低ランク分解では、低次元の連続的な値を持つ潜在空間でノードを表現
 - 各ノードには k 次元の実数値が割り当てられる



- 離散潜在変数の場合： 確率的ブロックモデル
 - 各ノードには離散的な潜在状態が割り当てられる

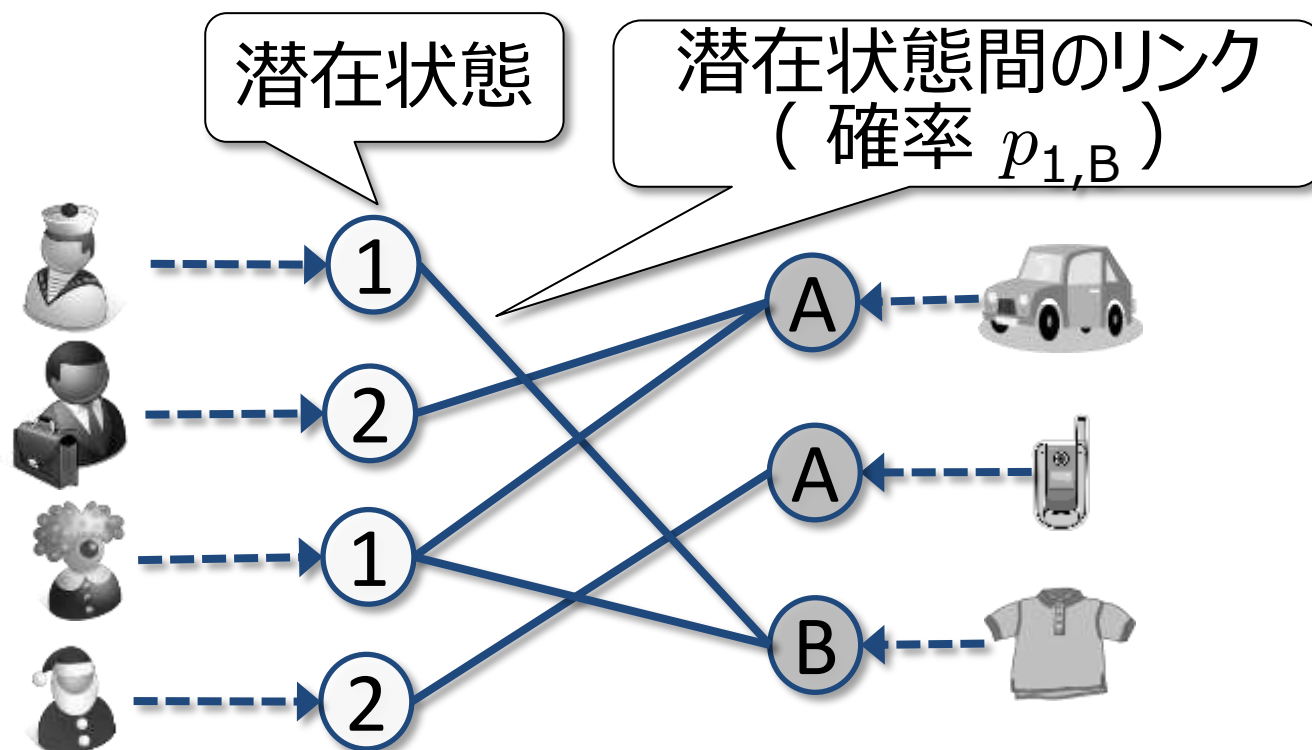
確率的ブロックモデルにおける潜在変数： 各ノードは離散的な潜在状態に割り当てられる

- 各ノードは潜在状態のうちのいずれかに割り当てられる
 - 多項分布に従って確率的に割り当てる



確率的ブロックモデルにおけるリンクの生成： 潜在状態の組に応じた確率でリンクが張られる

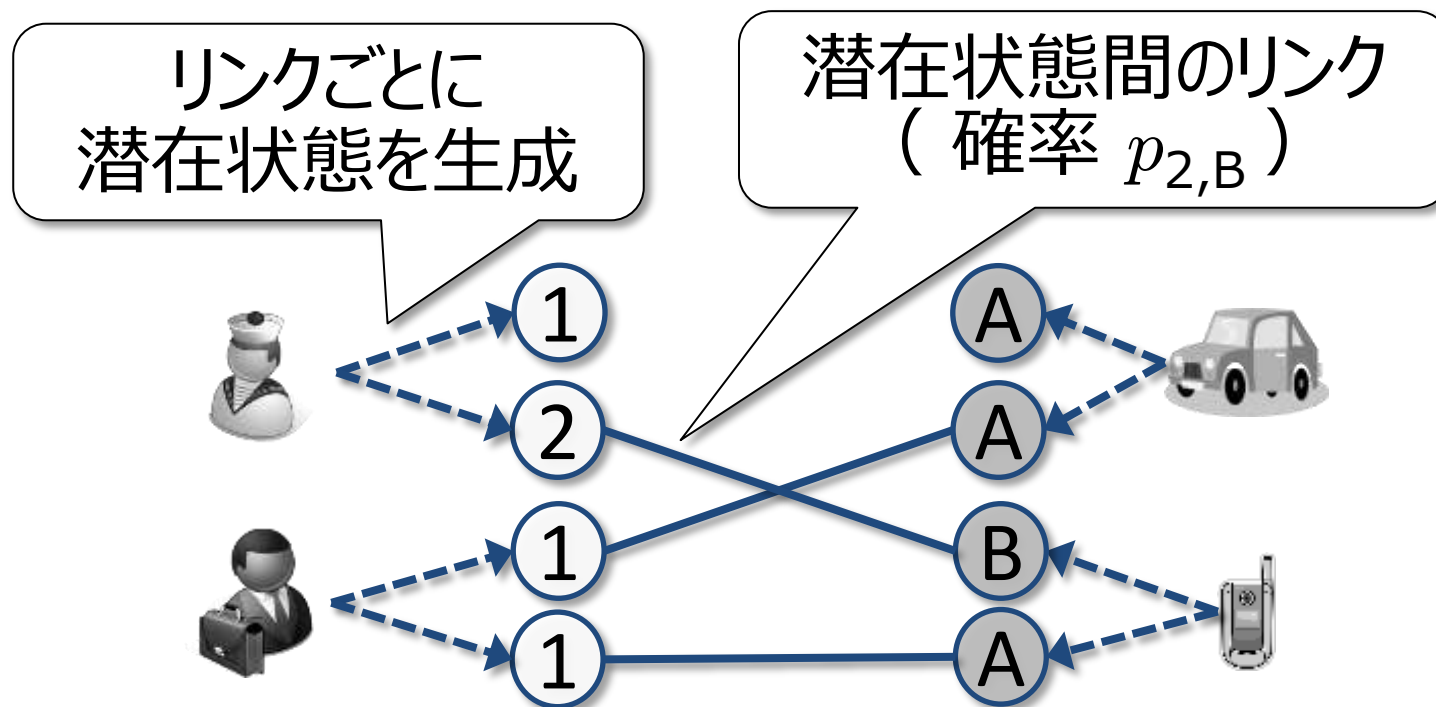
- 2つの潜在状態の組に応じた確率でリンクが張られる
 - ベルヌーイ分布に従ってリンクを生成



混合メンバシップモデル： リンクごとにノードの役割が変わるようなモデル

■ 混合メンバシップモデル

- 各ノードがリンクごとに別々の潜在状態をとりうる
- 「関わりかたによって役割をかえる」



まとめ:

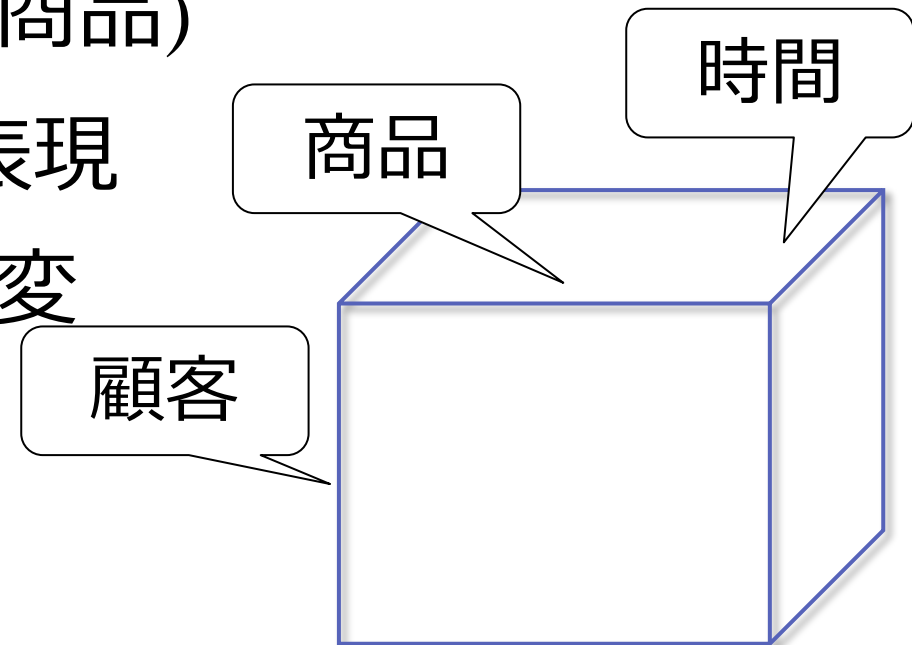
外部ネットワーク構造の潜在変数モデル

- 潜在変数モデル：
 - ネットワークの大局的構造を捉える
 - ノードの潜在的情報を仮定
- GroupLens：初期の協調フィルタリング手法
 - 行／列の類似度を用いて重みつき予測
- 行列分解：低ランク性を仮定した連続潜在変数モデル
- 離散潜在変数モデル：確率的ブロックモデル、混合メンバシップモデル

テンソル（多次元配列）：

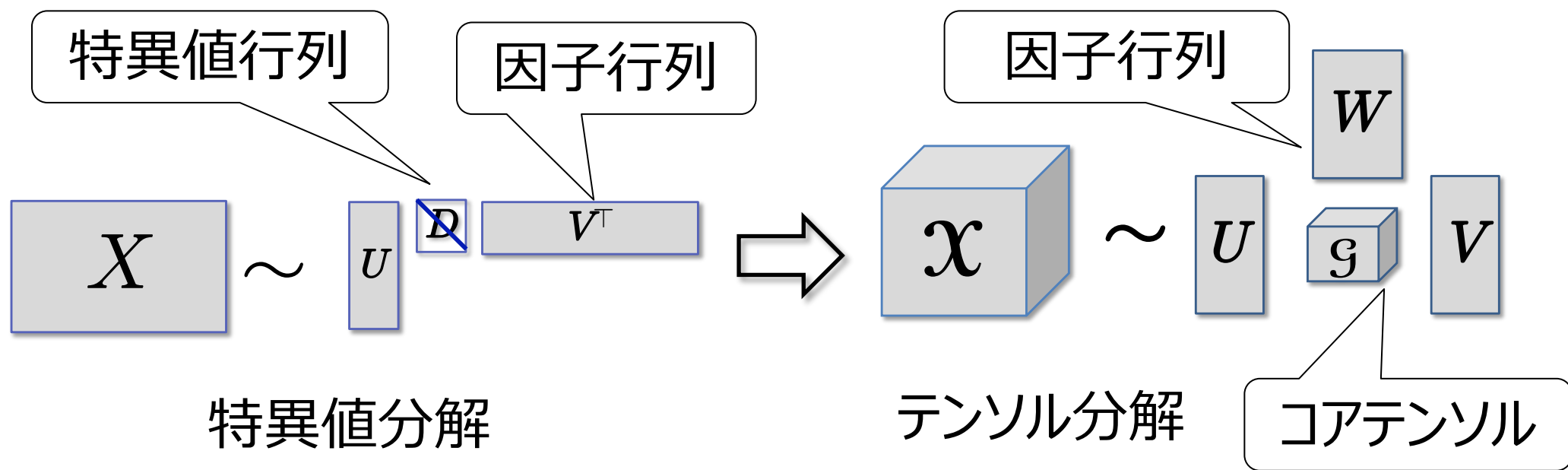
多ノードの関係の表現

- テンソル：行列の多次元拡張
- 複数オブジェクト間の複雑な関係を表現できる
 - 関係の時間的变化：（顧客, 商品, 時間）
 - 関係の種類：（顧客, 行動, 商品）
- ハイパーグラフはより一般的な表現
 - 関係に参加するノード数が可変



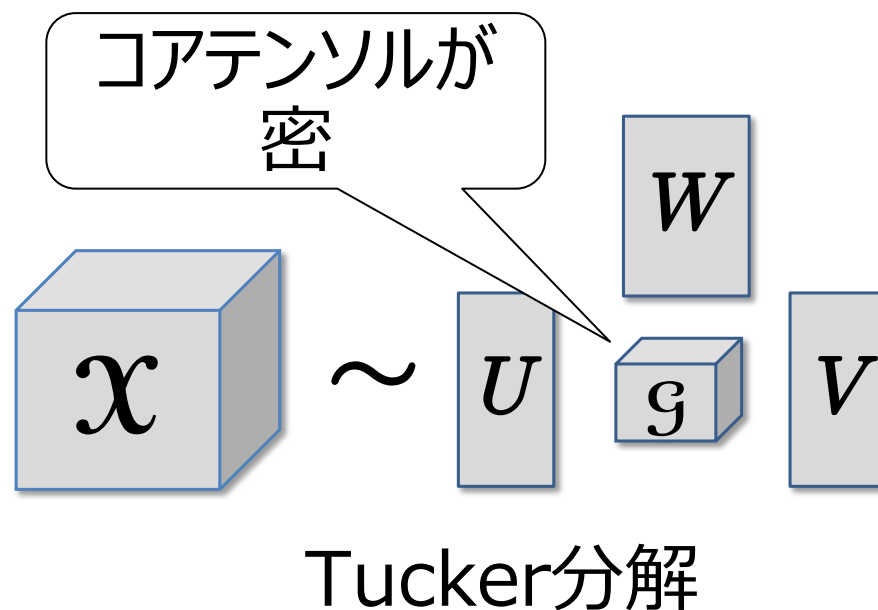
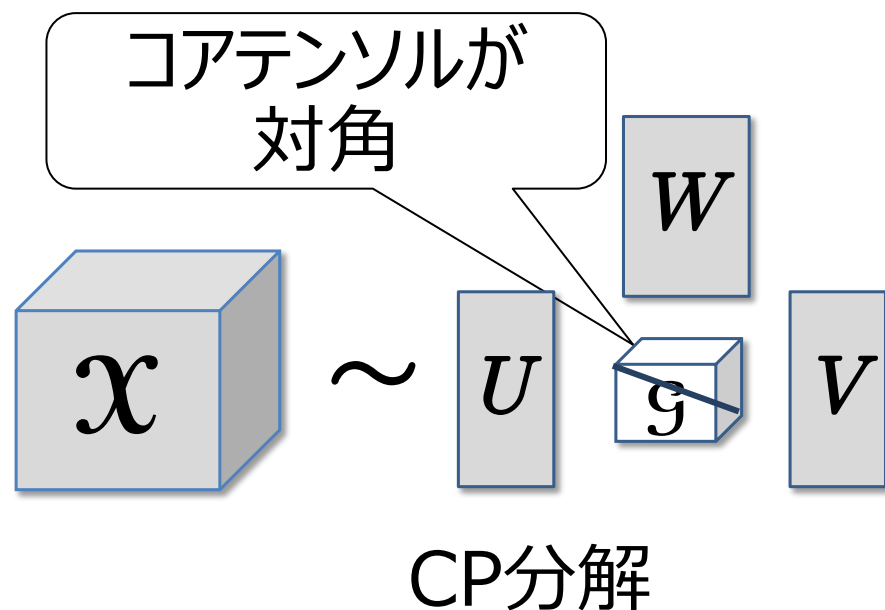
テンソル分解： 行列の低ランク分解の一般化

- 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人気



典型的なテンソル分解： CP分解とTucker分解

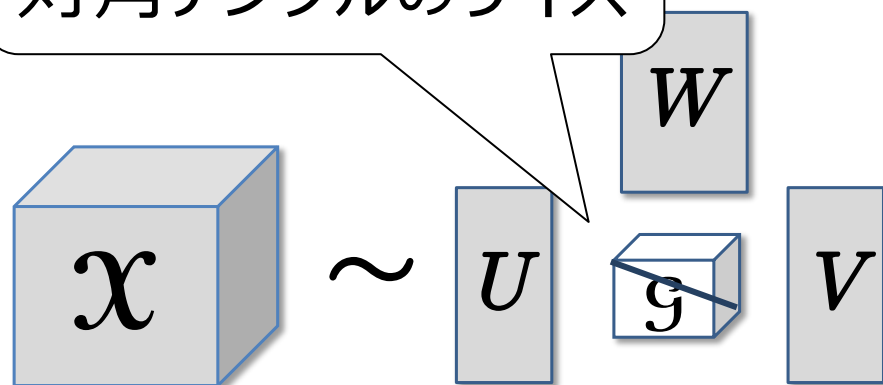
- CP分解：特異値分解の自然な拡張
(コアテンソルが対角；正方)
- Tucker分解：よりコンパクトな表現
(みっちりコア；各モードの次数が異なる)



テンソル分解のランク： 分解のタイプごとにランクの定義が異なる

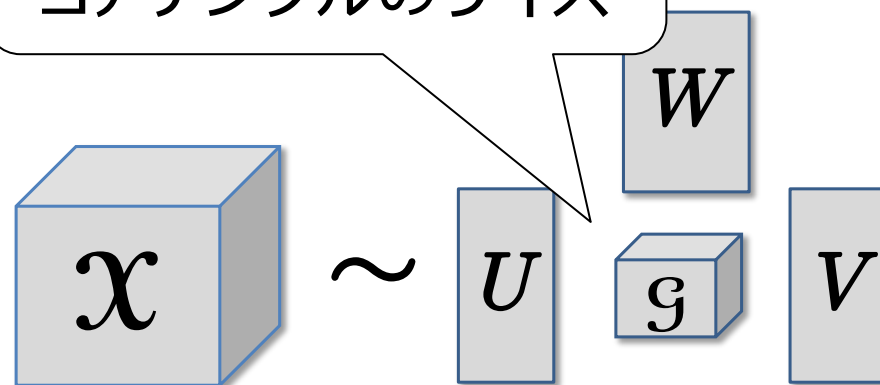
- 行列のランク：特異値分解の非零の特異値数
- テンソル分解のランクは分解のタイプによって決まる
 - CP分解、Tucker分解 それぞれランクの定義がある

CPのランク＝
対角テンソルのサイズ



CP分解

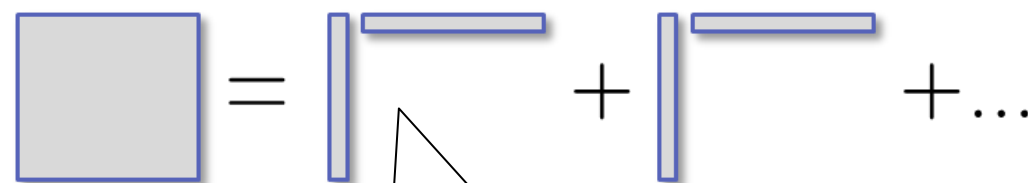
Tuckerのランク＝
コアテンソルのサイズ



Tucker分解

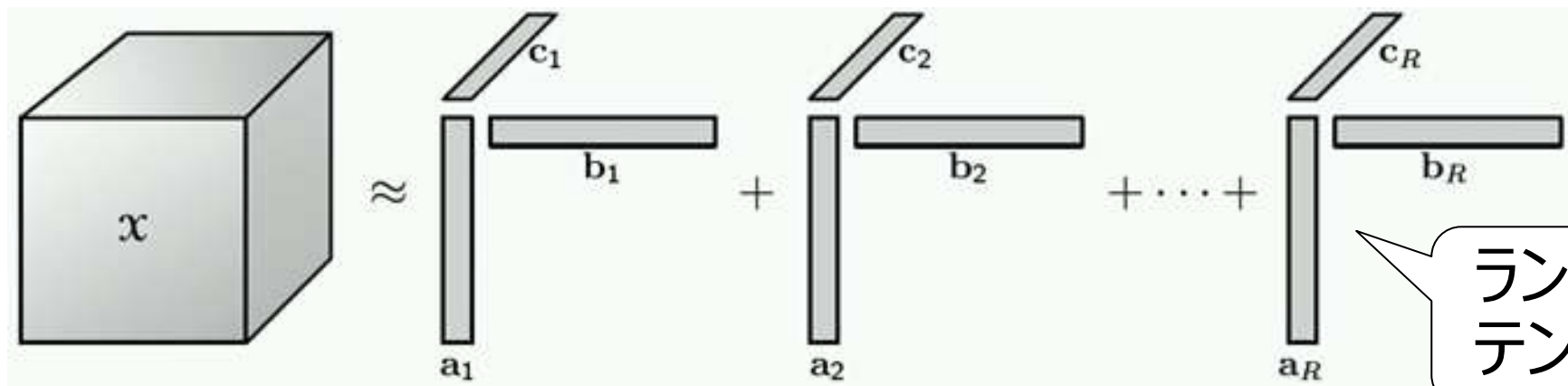
CP分解： ランク1テンソルの和

- 行列：ランク1行列の和



ランク 1 行列

- CP分解：ランク1テンソルの和



ランク 1
テンソル

外積

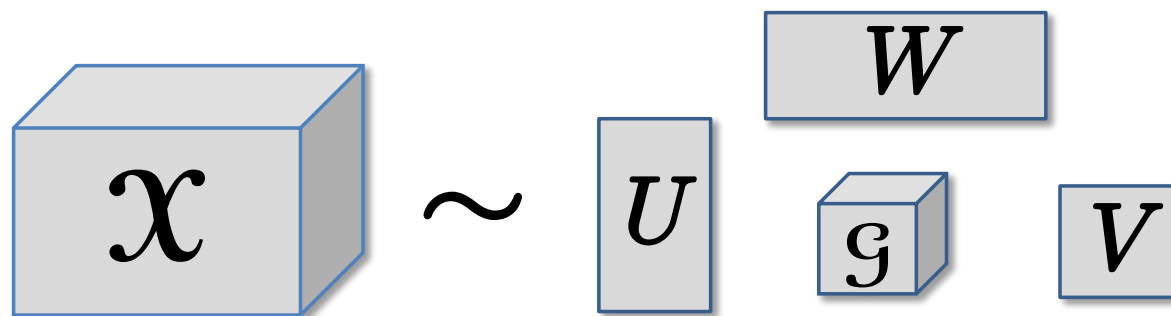
$$\mathcal{X} \sim \sum_r \lambda_r \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \quad (x_{ijk} = \sum_r \lambda_r a_{ri} b_{rj} c_{rk})$$

Tucker分解： コアテンソルと複数の因子行列に分解

- Tucker分解：コアテンソル＋因子行列

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

- モード積 \times_k を使って定義される



- 多くの場合因子行列の列ベクトルが正規直交性を仮定
- CP分解は特殊ケース：コアテンソルが対角



応用事例：

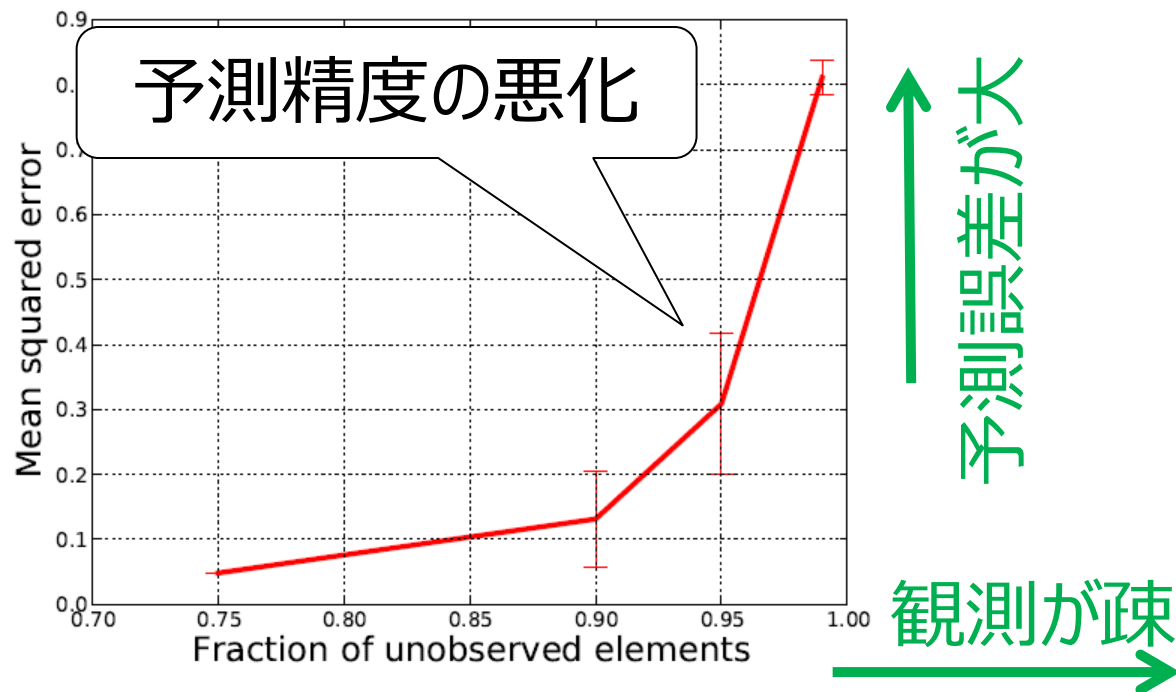
ネットワークの時間変化、3 項関係の予測 など

- ソーシャルネットワーク分析 (人×人×時間)
 - 人間関係の時間的移り変わりを解析
- タグ推薦 (人×Webページ×タグ)
 - Webページにつけるタグを推薦
 - 人によってタグのつけ方には個性がある
- Webリンク解析
(Webページ×Webページ×アンカーテキスト)
- 画像認識 (画像×人×向き×明るさ×...)

テンソル分解の課題：

多項関係の予測ではデータの疎性対処が課題

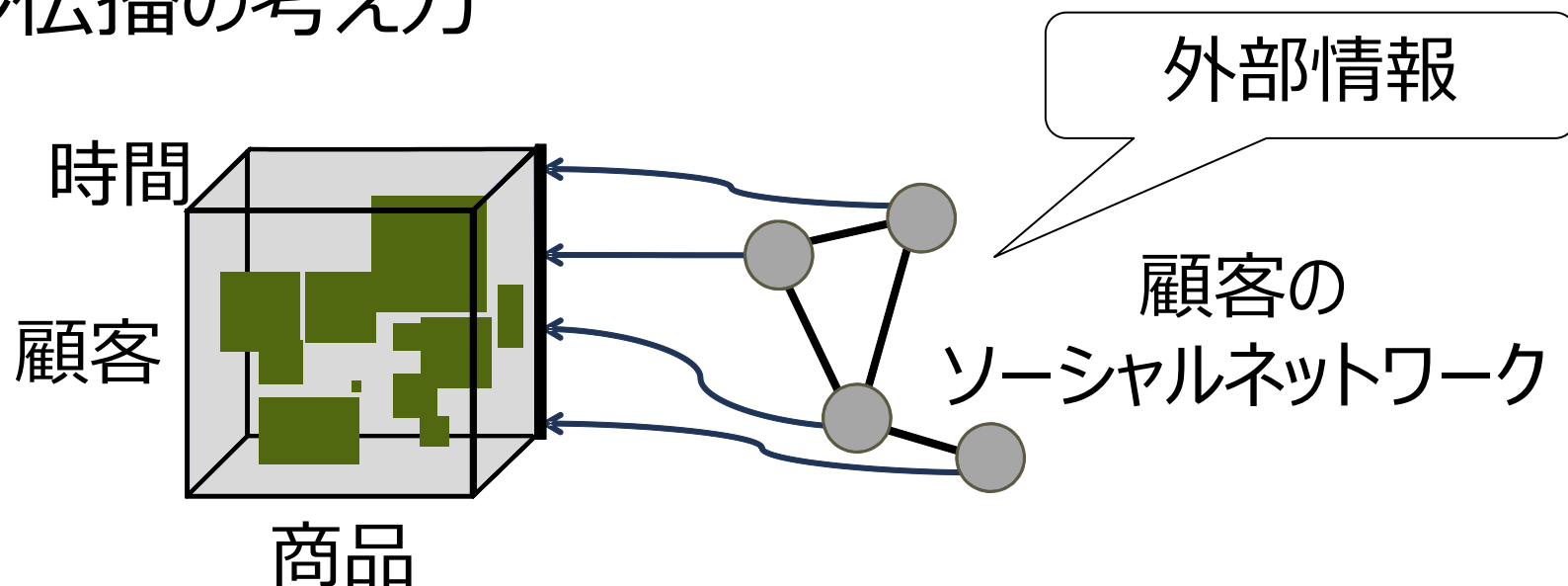
- 観測が疎なときには、補完の予測精度が著しく悪化
 - 可能な関係の数は組み合わせ的に増加
- 低ランクの仮定だけでは足りない！



疎性への対処： 低ランク性＋外部情報の利用

- 多くの場合、データ間の関係が外部情報として利用可能
- 仮定：「隣同志は振る舞いが似ている」

— ラベル伝播の考え方



Narita, Hayashi, Tomioka & Kashima: Tensor Factorization Using Auxiliary Information
In ECML PKDD 2011 (*won the Best Student Paper Award*), Data Mining & Knowledge Discovery, 2012

補助ネットワークの利用：

「隣同志は似ている」を用いた予測の補助

- 外部情報として与えられる関係情報を推論のガイドに

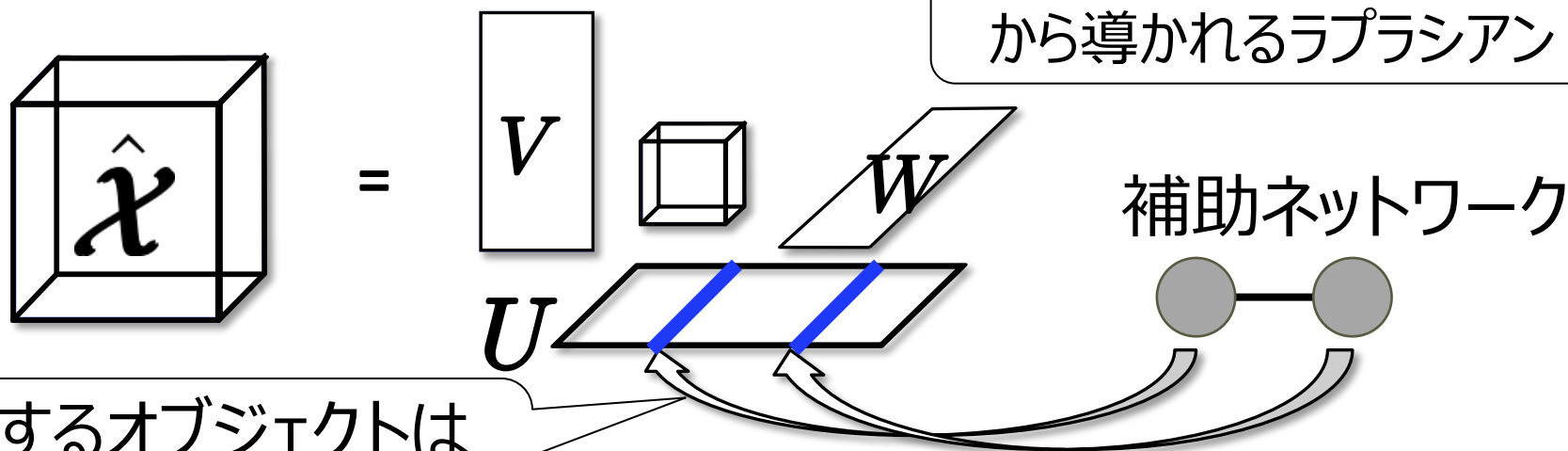
— 隣り合うオブジェクトが似た振る舞いをするように働く

誤差項

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

補助項

補助ネットワーク
から導かれるラプラシアン



「隣接するオブジェクトは
似た振る舞いをするべし」

まとめ:

テンソル分解による外部ネットワーク構造解析

- テンソル（多次元配列）：3項以上の関係を表現
- テンソルの低ランク分解：CP分解とTucker分解
- データの疎性への対処：外部情報の利用

まとめ

ネットワーク構造解析の世界観：

2 × 2 の 4 通りの分類がある

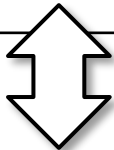
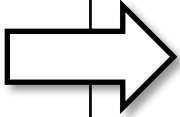
- {内部, 外部}ネットワーク
× {個々のデータ, 内外の関係}についての推論 の4通り

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォ ー カ ス	個々の データ	<ul style="list-style-type: none">• 予測• クラスタリング• 構造ラベリング	<ul style="list-style-type: none">• 予測• クラスタリング• ランキング
	データ 内外の 関係	<ul style="list-style-type: none">• パタン発見• 構造予測	<ul style="list-style-type: none">• リンク予測• 構造変化解析

ネットワーク構造解析のためのモデル： 線形識別モデルと潜在変数モデル

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォールカス 解析の	個々のデータ	<ul style="list-style-type: none">カーネル法パターンマイニング	<ul style="list-style-type: none">ラベル伝播マルコフネットワーク潜在変数モデル
	データ内外の関係	<ul style="list-style-type: none">パターンマイニング構造学習器(HMM、CRF等)	<ul style="list-style-type: none">リンク指標ペアワイズ予測マルコフネットワーク潜在変数モデル

ネットワーク構造解析のためのモデル： 線形識別モデルと潜在変数モデル

		ネットワーク構造の種類	
		内部ネットワーク	外部ネットワーク
フォーカス 解析の	個々のデータ	部分構造に注目した 線形モデルの拡張	ラベル伝播 「隣同志は似ている」
	データ内外の関係		 潜在変数モデル 「付き合い方の似ている同志は似ている」
			 ペアワイズ予測 マルコフネットワーク

連絡先： 鹿島久嗣
kashima@mist.i.u-tokyo.ac.jp