順序中の欠損対象の補完

神嶌 敏弘　赤穂 昭太郎
http://www.kamishima.net/
産業技術総合研究所
人工知能学会 KBS研究会 (2005.2.25-26)

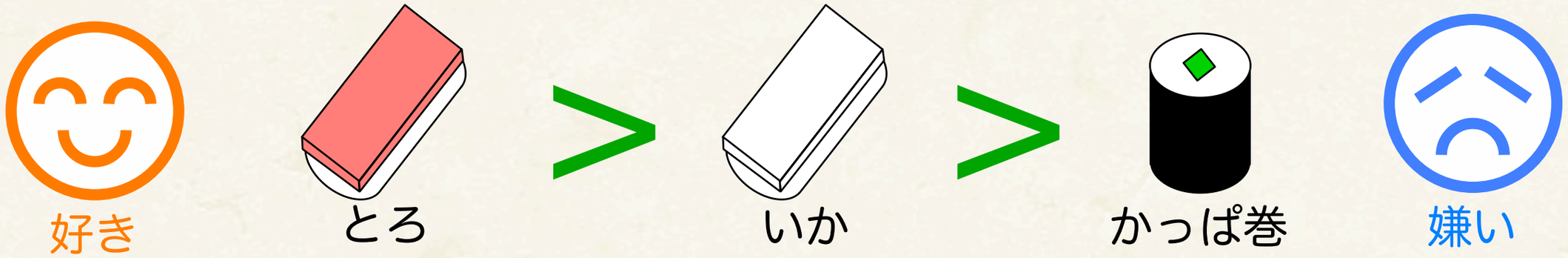
概要

順序中の欠損対象の順位を，順位のサンプルの要約統計量によって補完する簡潔で効果的な方法の提案

- 実データでは対象は頻繁に欠損するので補間手法は有用
- 欠損対象の順位の補完により，順序の類似度をより正確に評価できるようになる

順序: 何らかの基準で対象を整列したもの

例：Aさんが好き[基準]な順に寿司[対象]を整列



“いかよりとろが好き”だが“どれくらい好き”かは不明

何が難しいのか？

サンプルの統計量で欠損値を補完

↓

非常に一般的な手法

例：数値属性の欠損値を平均値で補完
　　カテゴリ属性を最頻値(モード)で補完

順序だとどのような問題があるのか？

- 何を使って補完すればよいのか？

  順序変量で，補完に使えるサンプルの統計量

- どうやって置き換えればよいのか？

  順序変量は相対的な関係のみに意味があるので，単純に置き換えるといった操作では補完できない

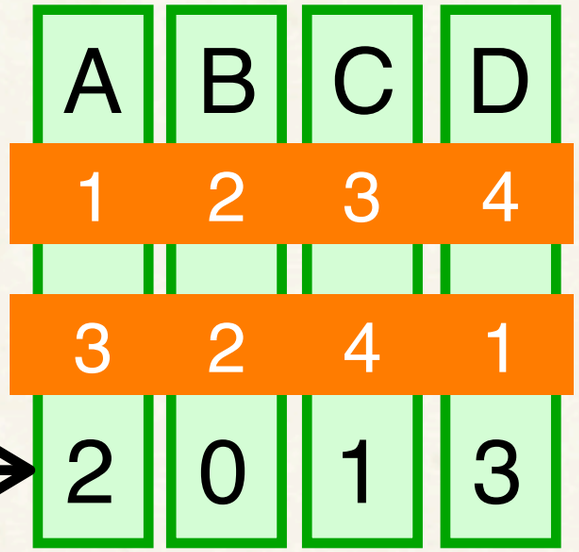
順序間の距離

距離: 二つの順序間の非類似度を測る尺度

O_1 A > B > C > D

O_2 D > B > A > C

順位への変換



Spearmanの距離

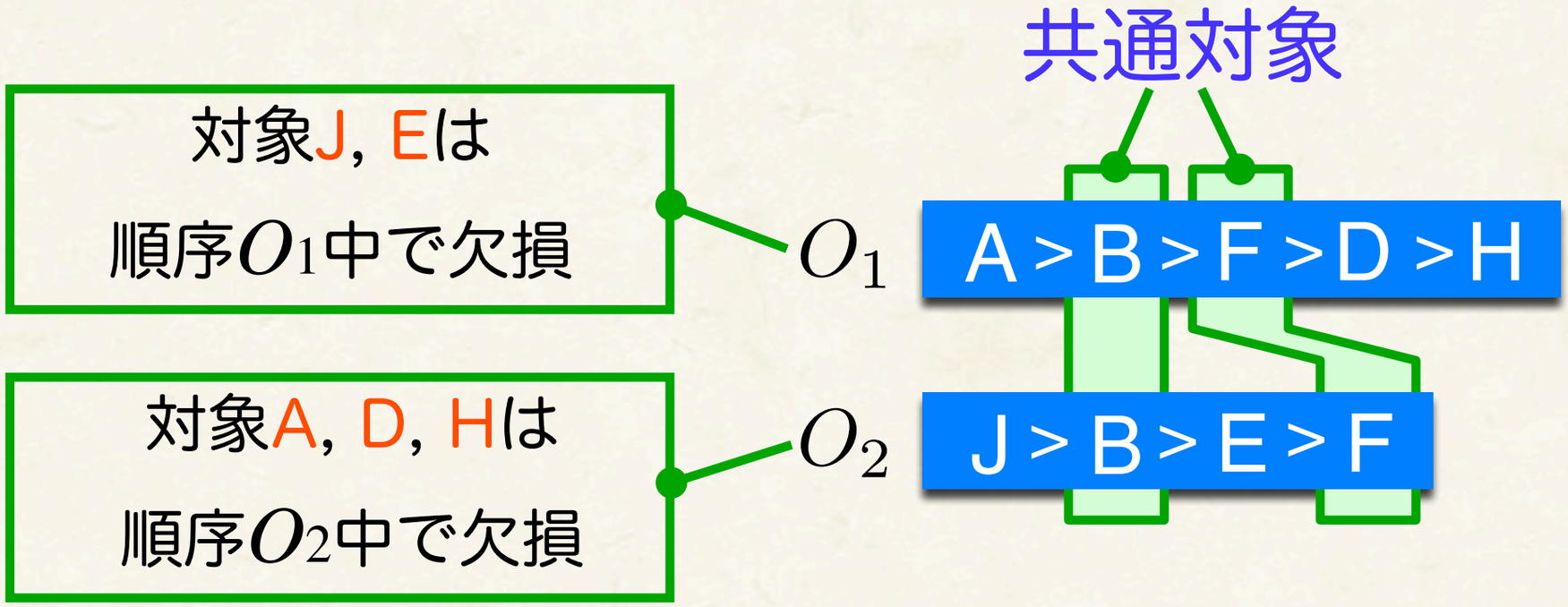
対象間の順位の差の2乗和

距離は同じ対象で構成される順序の間で定義される

不完全順序間の距離の計測法

実データでは不完全順序が頻繁に観測される

➡ 幾つかの対象が欠損



距離は共通対象上で計算されその他の対象は無視

欠損対象に含まれている情報は廃棄される

⬇

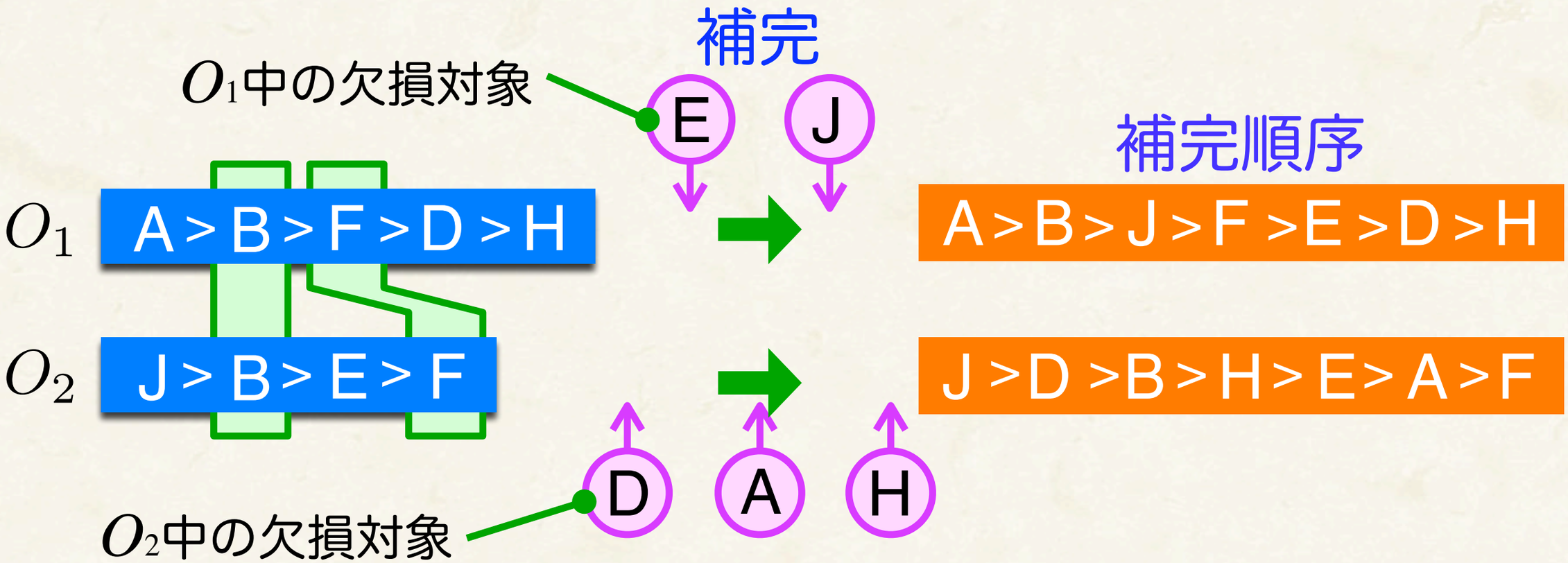
距離の精度は低い

欠損対象の補完

もし欠損対象に含まれている情報が利用できれば

↓

距離の計測精度は高くできる



欠損対象を**何らかのデフォルト値**で補完して

距離を補完順序の間で計測する

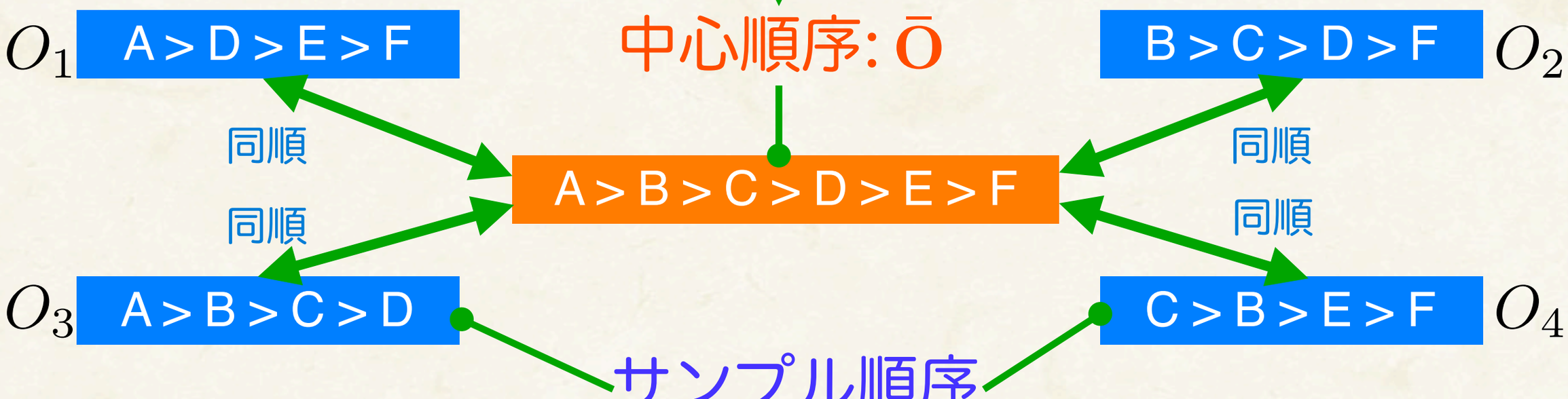
中心順序

順序のデフォルト値にはどんなものが適切か？

デフォルト値: 数値属性 ⟶ サンプルの平均値

カテゴリ属性 ⟶ サンプルの最頻値

順序 ➡ サンプル順序の中心順序



サンプル順序と平均的に同順

$$\bar{O} = \arg\min \sum_{O_i \in S} \text{Distance}(O, O_i)$$

補完手法

INPUT

サンプルの中心順序

D > B > J > H > A > F > E

欠損対象

D　A　H

補完する順序

J > B > E > F

欠損対象を中心順序と同順になるように整列する

欠損対象の整列

D > H > A

デフォルト順序

デフォルト順序と補完する順序と併合

J > D > B > H > E > A > F

補完順序

デフォルト順序と補完する順序を併合することで補完順序を得る

デフォルト順序の併合

デフォルト順序と補完順序を適切に併合する

仮定

未知の完全順序　J > D > B > H > E > A > F

対象は均一にランダムにサンプリングされる

デフォルト順序/補完する順序　J > B > E > F

デフォルト順序

D > H > A

2.0　4.0　6.0

補完する順序

J > B > E > F

1.6　3.2　4.8　6.4

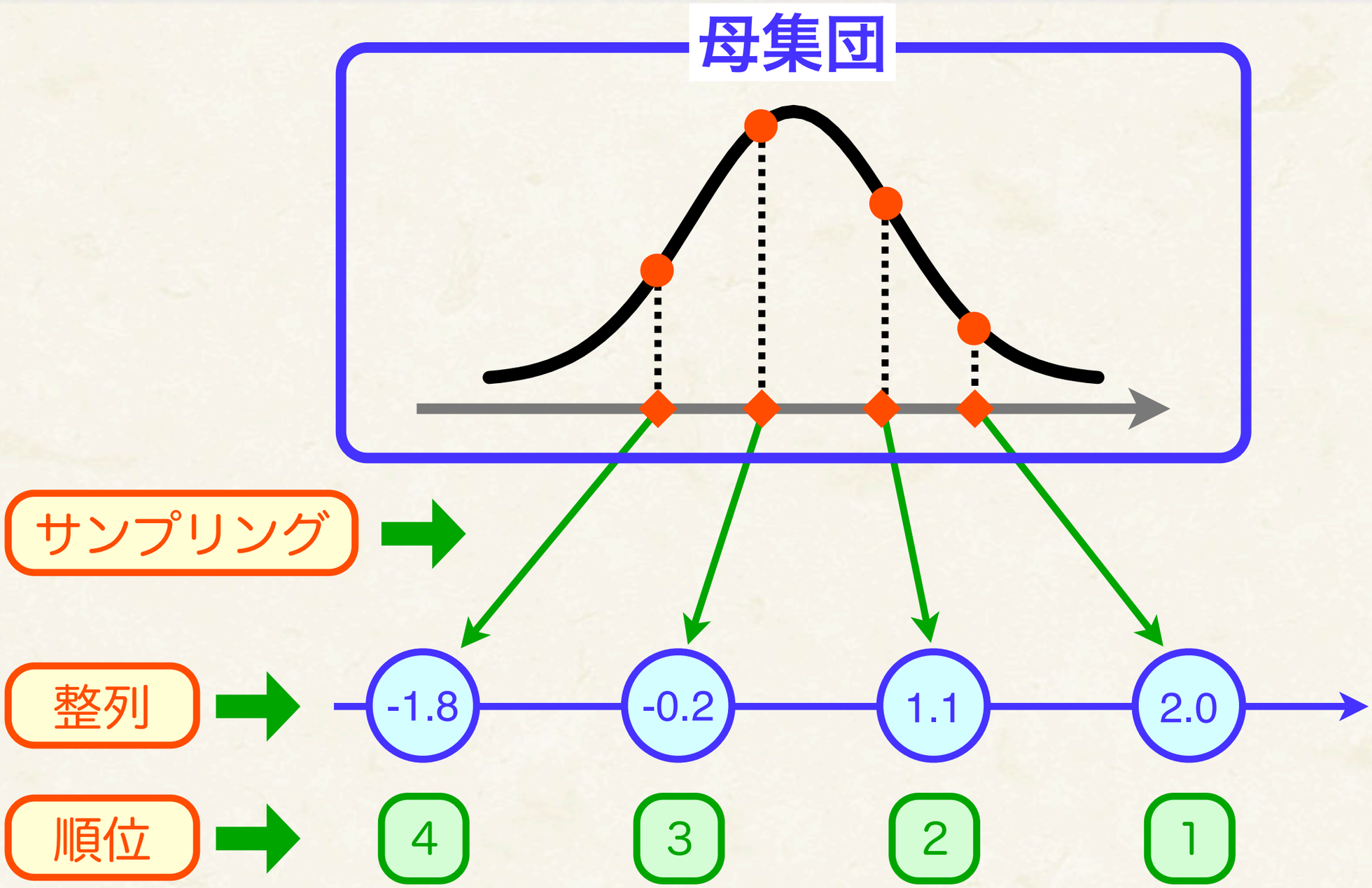
この仮定の下で
未知の完全順序中の
順位の期待値を計算

補完順序　J > D > B > H > E > A > F

1.6　2.0　3.2　4.0　4.8　6.0　6.4

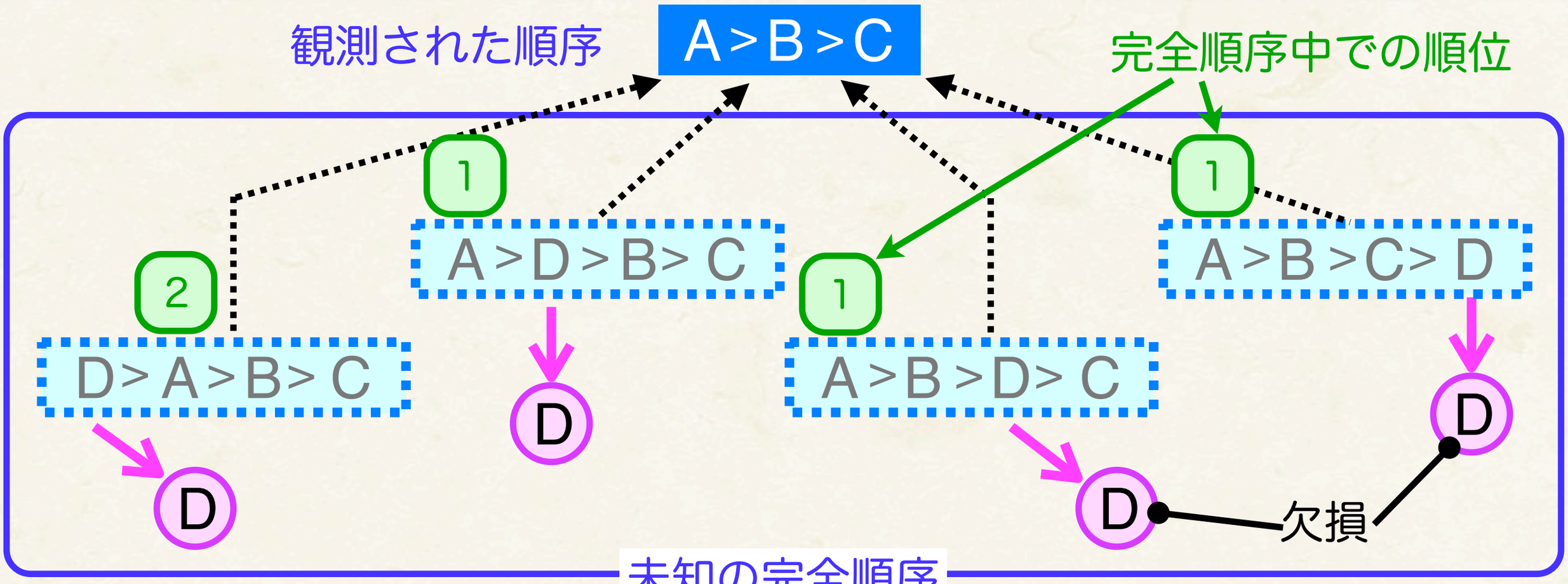
これらの順位の期待値に従って整列することで
二つの順序を併合する

順序統計



サンプル中で第n位の対象の分布は？

期待順位



元の順序は等確率(=1/4)でこれらのうちのいずれか

$$\text{期待順位} = \frac{\langle\text{完全順序の長さ}\rangle + 1}{\langle\text{観測順序の長さ}\rangle + 1} \times \langle\text{観測順序中の順位}\rangle$$

$$\text{対象Aの期待順位} = 1 \times 2 \times \frac{1}{4} + 3 \times 1 \times \frac{1}{4} = \frac{4+1}{3+1} \times 1 = \frac{5}{4}$$

実験

- 順序の補完手法を，利用者の応答順序に基づいた「なんとなく協調フィルタリング」に適用し，その有効性を検証する
- このフィルタリング手法では，利用者間の嗜好順序のSpearmanの距離を用いて測る

嗜好順序が短く，対象が高頻度で欠損している場合

↓

類似度の計測は不正確

→ 不適切な対象が推薦される

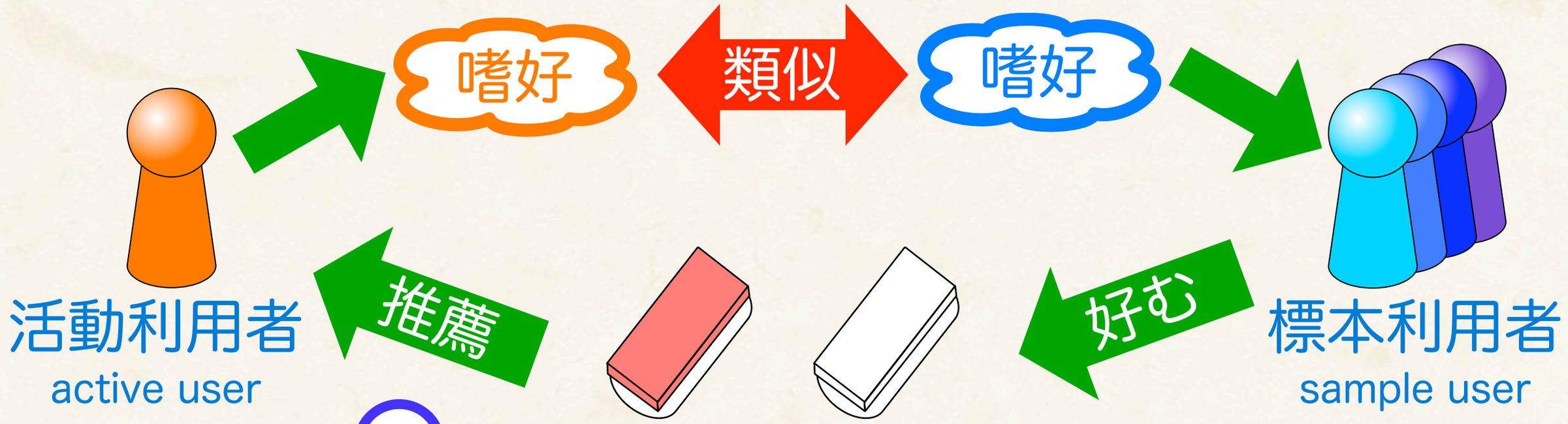
欠損対象を提案手法で補間することで
より適切な対象を推薦できるようにする

協調フィルタリング

「口コミ情報」を用いて利用者が好む対象を見つける方法

① 活動利用者の嗜好をシステムに提示

② 利用者DBから，活動利用者と類似した嗜好を持つ標本利用者を検索



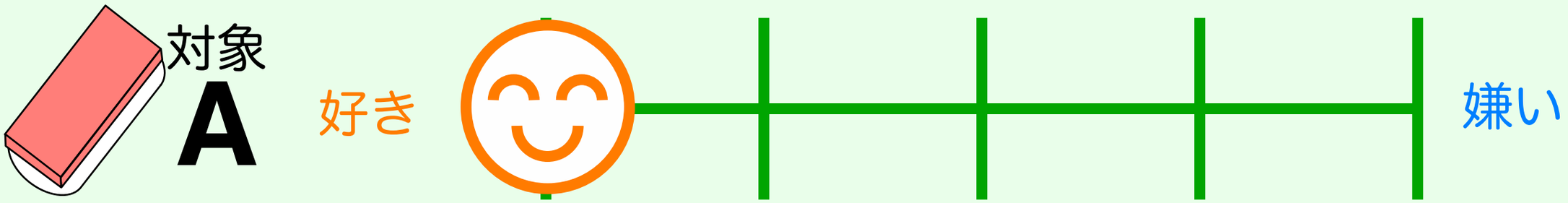
③ 類似した嗜好を持つ標本利用者が好む対象を，活動利用者に推薦

SD法と順位法

Semantic Differential 法 (SD法)

両端を対義語で表した尺度を用いる

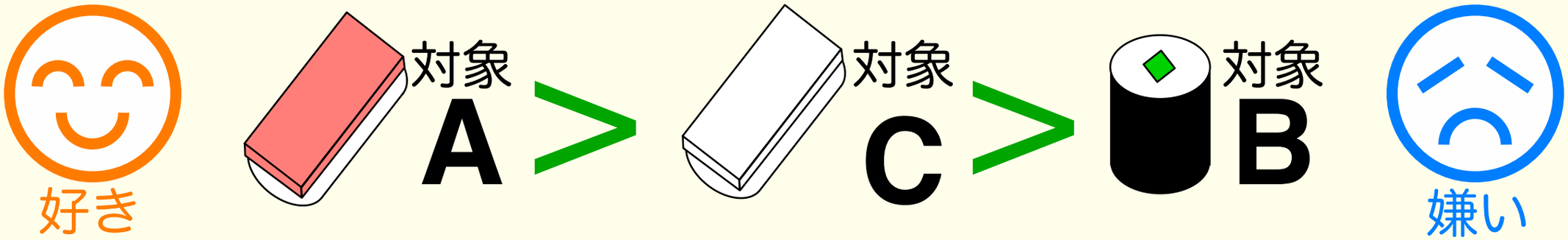
例：被験者が 対象A を好きならば，尺度の「好き」を選ぶ



順位法

計測する度合いの強さの順に対象を整列

例：被験者は 対象A が最も好きで，対象B が最も嫌い



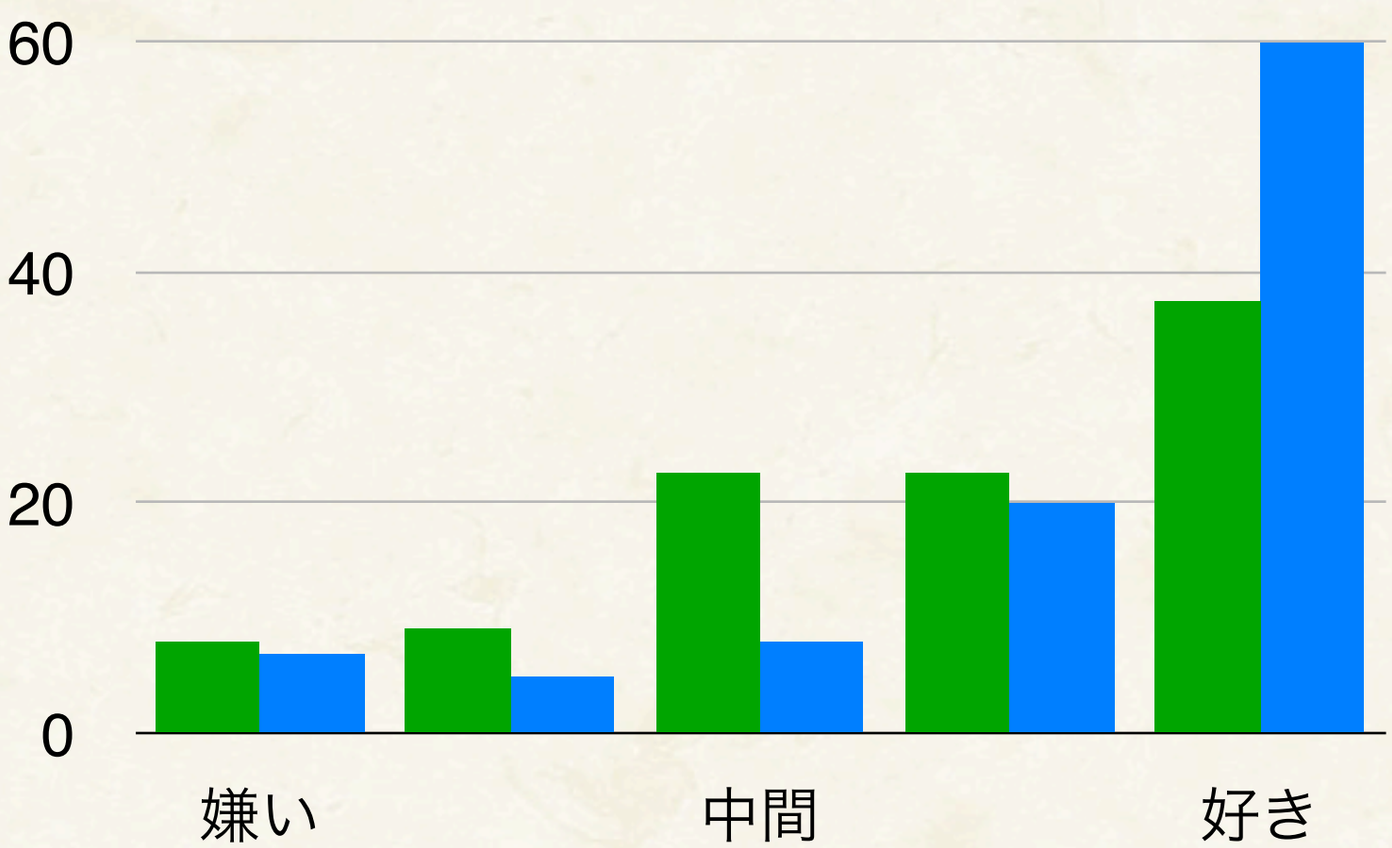
なぜ順序を使うのか？

SD法の仮定 ➡ 理想的なスコアの分布は対称で単峰

⬇

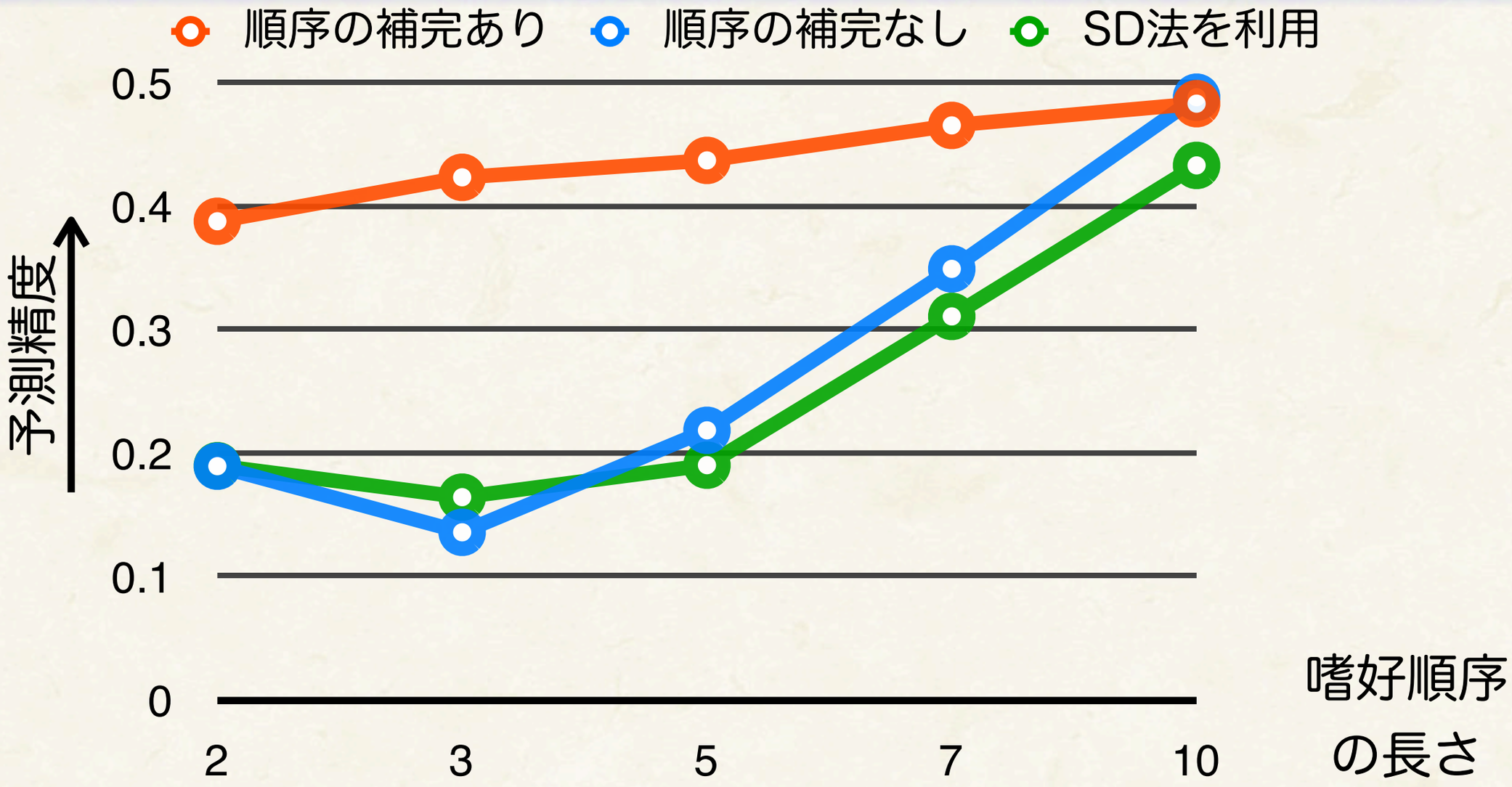
現実の分布はかなり歪んだ分布

5段階のSD法で，被験者が選択した値の分布の例



※ Amazonの書評のデータは
KDD2003のA.S.Weigendの
招待講演より引用

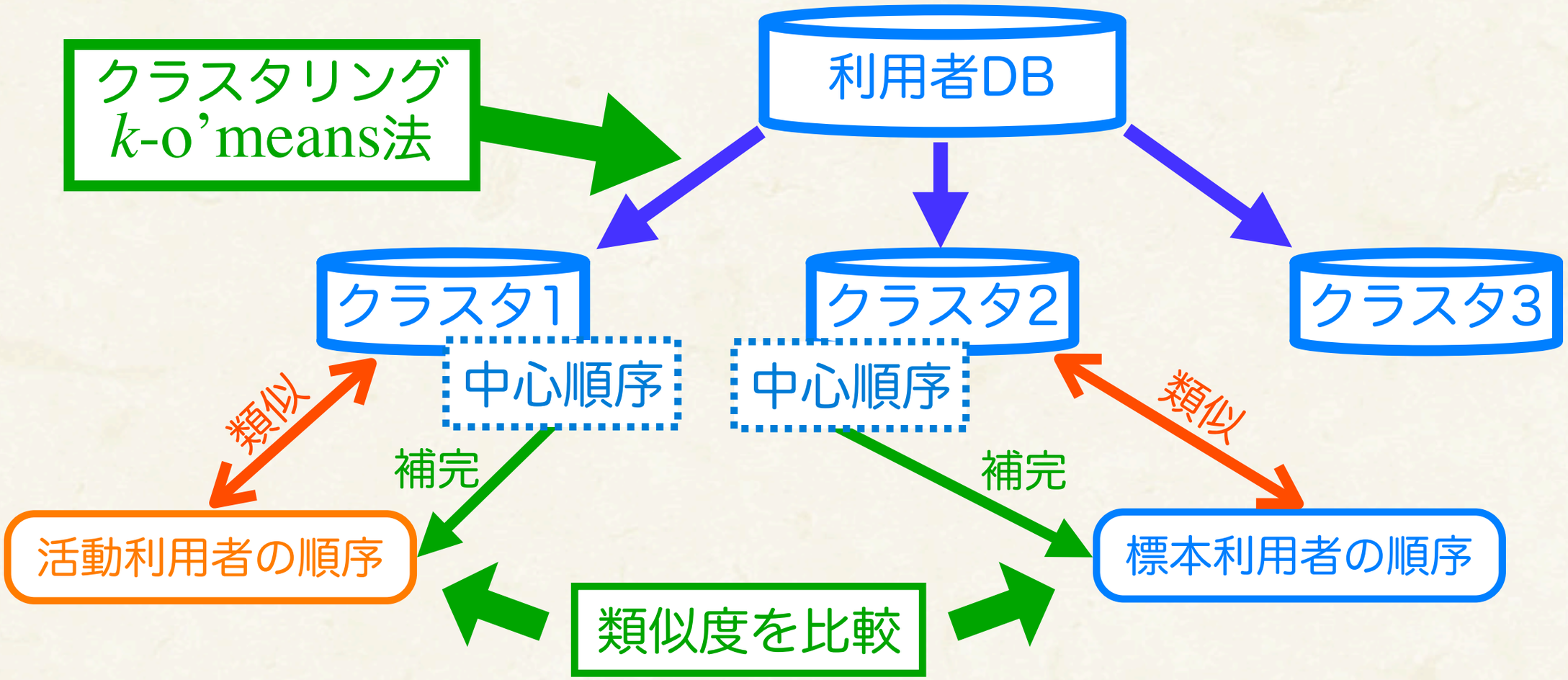
実験結果



応答順序が短い場合，対象はより頻繁に欠損する

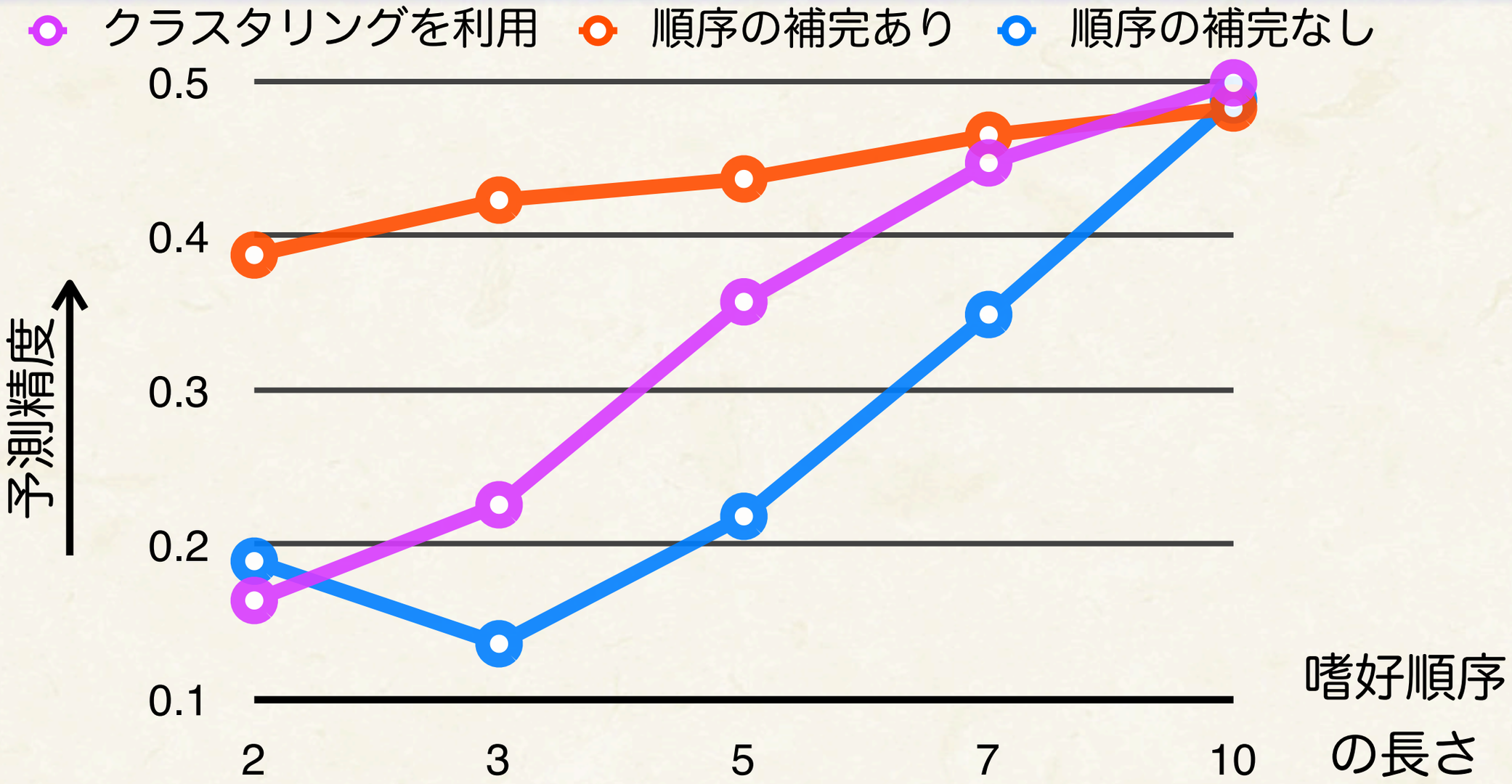
順序の補完手法はこの場合に効果的

クラスタリングとの併用



より精度の高い順序で補完するため，DB全体の中心順序ではなく，補完対象の順序が含まれるクラスタの中心順序で補完する

実験結果(クラスタリング併用)



応答順序が短い場合クラスタへの割当てを正確に判定できない

クラスタの中心順序で補完するのは順序が長い場合のみ有効

まとめ

順序中の欠損対象の順位を，サンプルの中心順序によって補完する方法を提案した

- 嗜好順序に基づく協調フィルタリング手法にてきようすることで，実験的にその効果を示した
- この手法は計算量的に効率的

$$O(\max(|X'|, |\tilde{X}| \log |\tilde{X}|))$$

$|X'|$：補完順序の長さ　　$|\tilde{X}|$：デフォルト順序の長さ

クラスタリングなどの他の手法への適用などを予定

追加情報: **http://www.kamishima.net/**