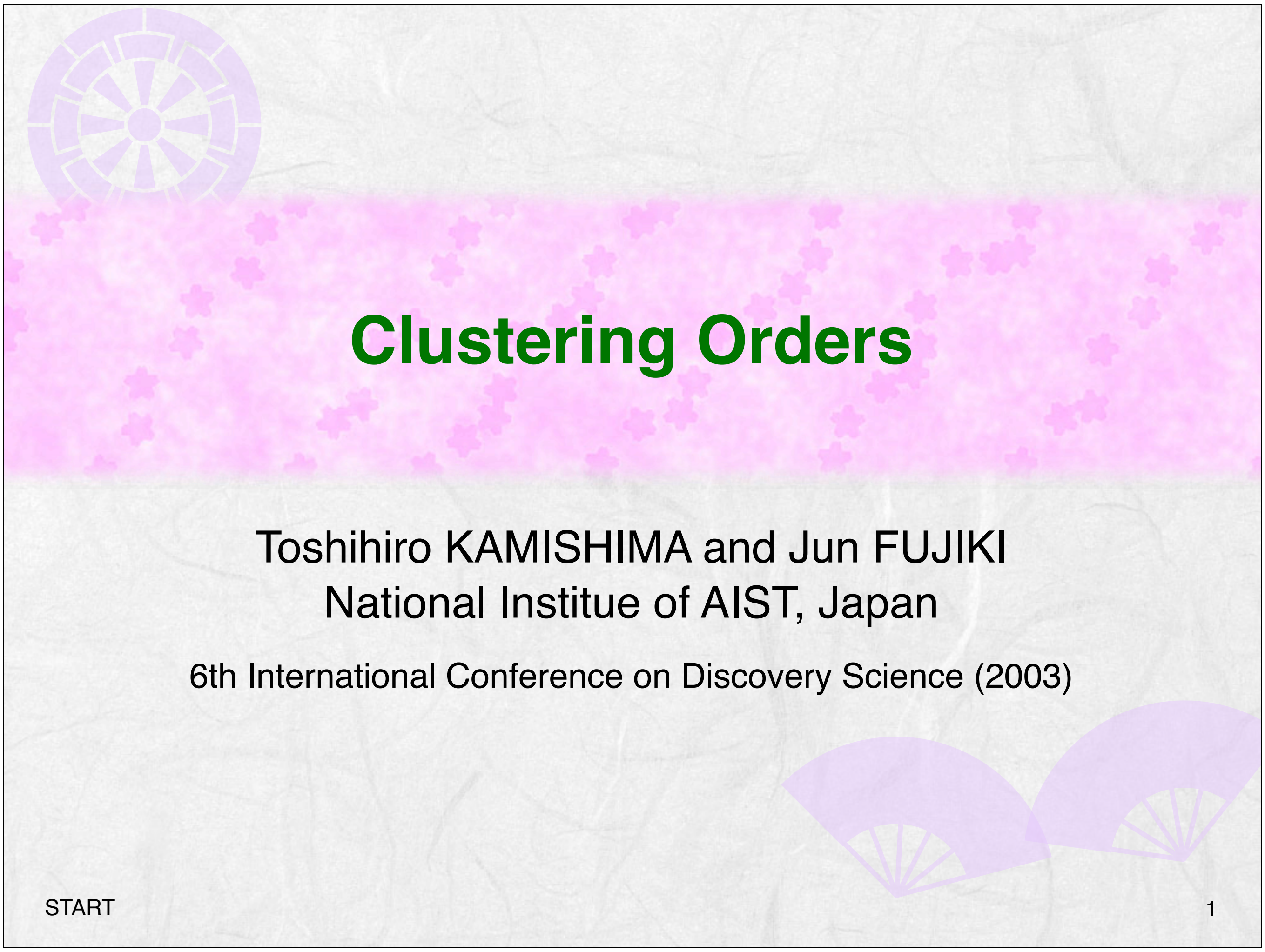
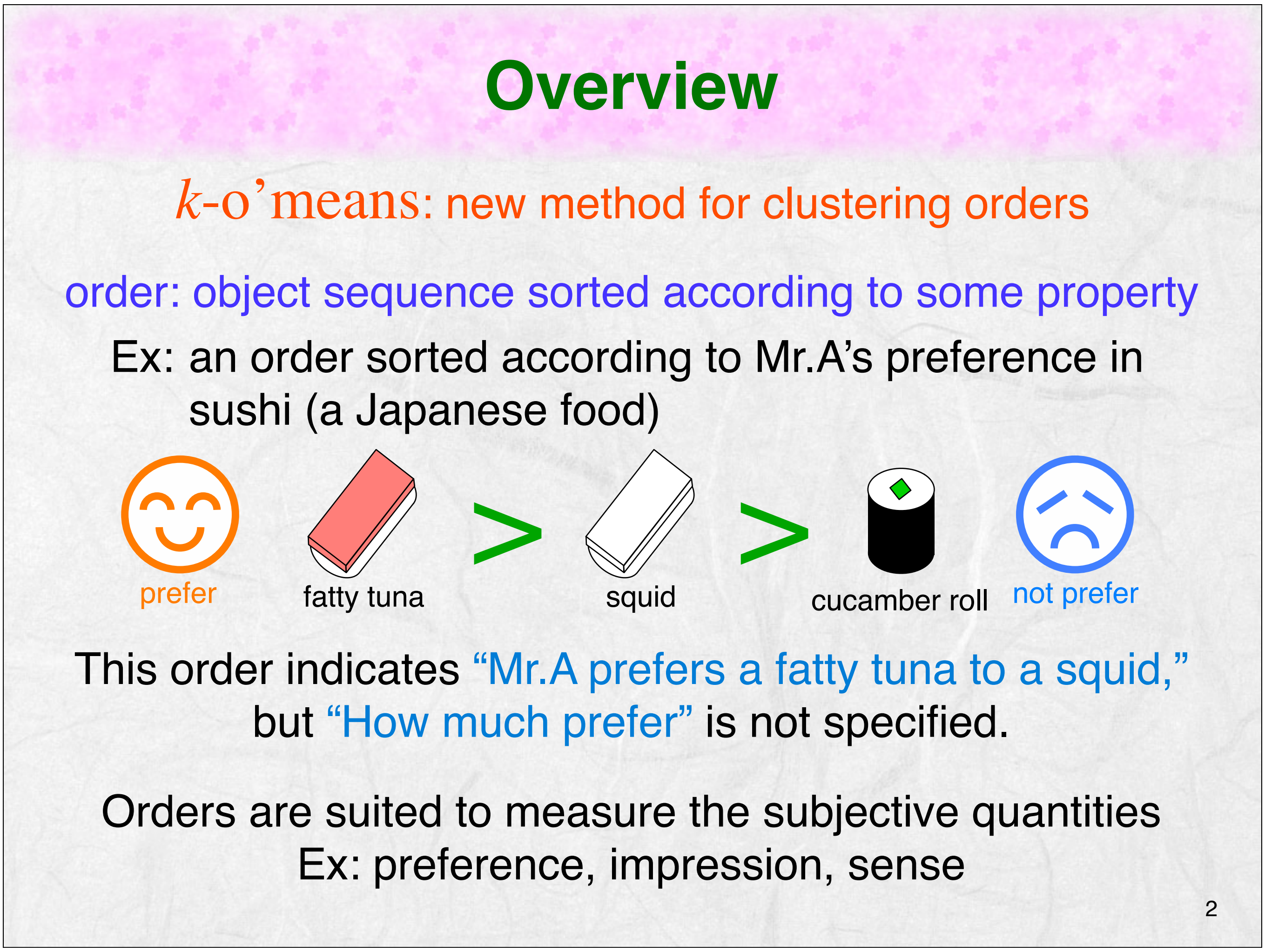


Clustering Orders

Toshihiro KAMISHIMA and Jun FUJIKI
National Institue of AIST, Japan

6th International Conference on Discovery Science (2003)

We would like to talk about the method for clustering orders.



Overview

k-o'means: new method for clustering orders

order: object sequence sorted according to some property

Ex: an order sorted according to Mr.A's preference in sushi (a Japanese food)

prefer
fatty tuna
>
squid
>
cucamber roll
not prefer

This order indicates "Mr.A prefers a fatty tuna to a squid," but "How much prefer" is not specified.

Orders are suited to measure the subjective quantities
Ex: preference, impression, sense

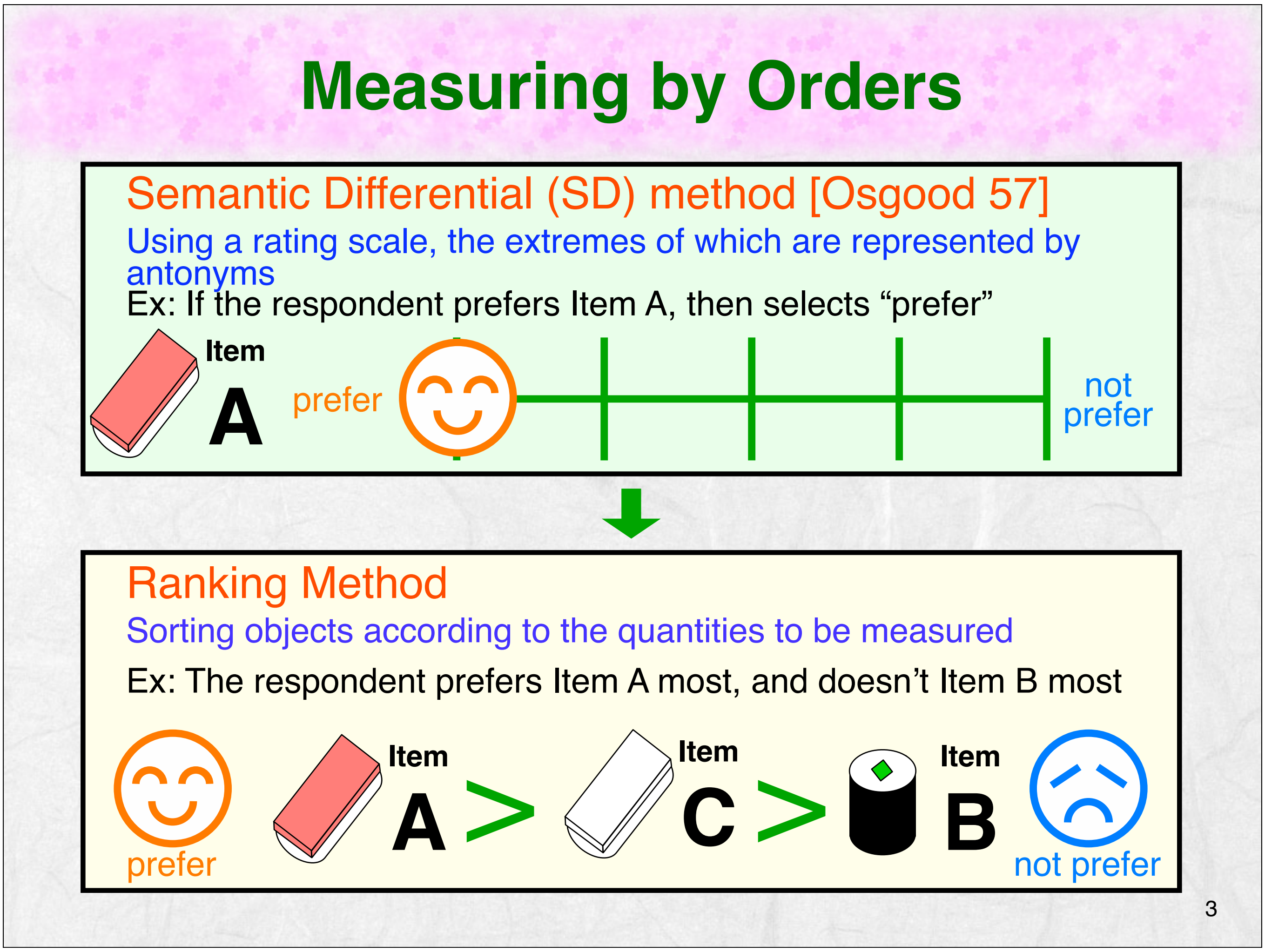
We propose a new method for clustering orders: k-o'means.

We define an order as a sequence sorted according to some properties.

This is an example of an order sorted according to Mr.A's preferences in sushi.

This order indicates that "Mr.A prefers a fatty tuna to squid", but "How much prefer" is not represented.

Orders are suited to measure the subjective quantities, for example, preferences, impression, sense.



Measuring by Orders

Semantic Differential (SD) method [Osgood 57]

Using a rating scale, the extremes of which are represented by antonyms

Ex: If the respondent prefers Item A, then selects “prefer”

Ranking Method

Sorting objects according to the quantities to be measured

Ex: The respondent prefers Item A most, and doesn’t Item B most

Traditionally, such subjective quantities are measured by Semantic Differential method.

In this method, the quantities are measured by using a rating scale, the extremes of which are represented by antonyms.

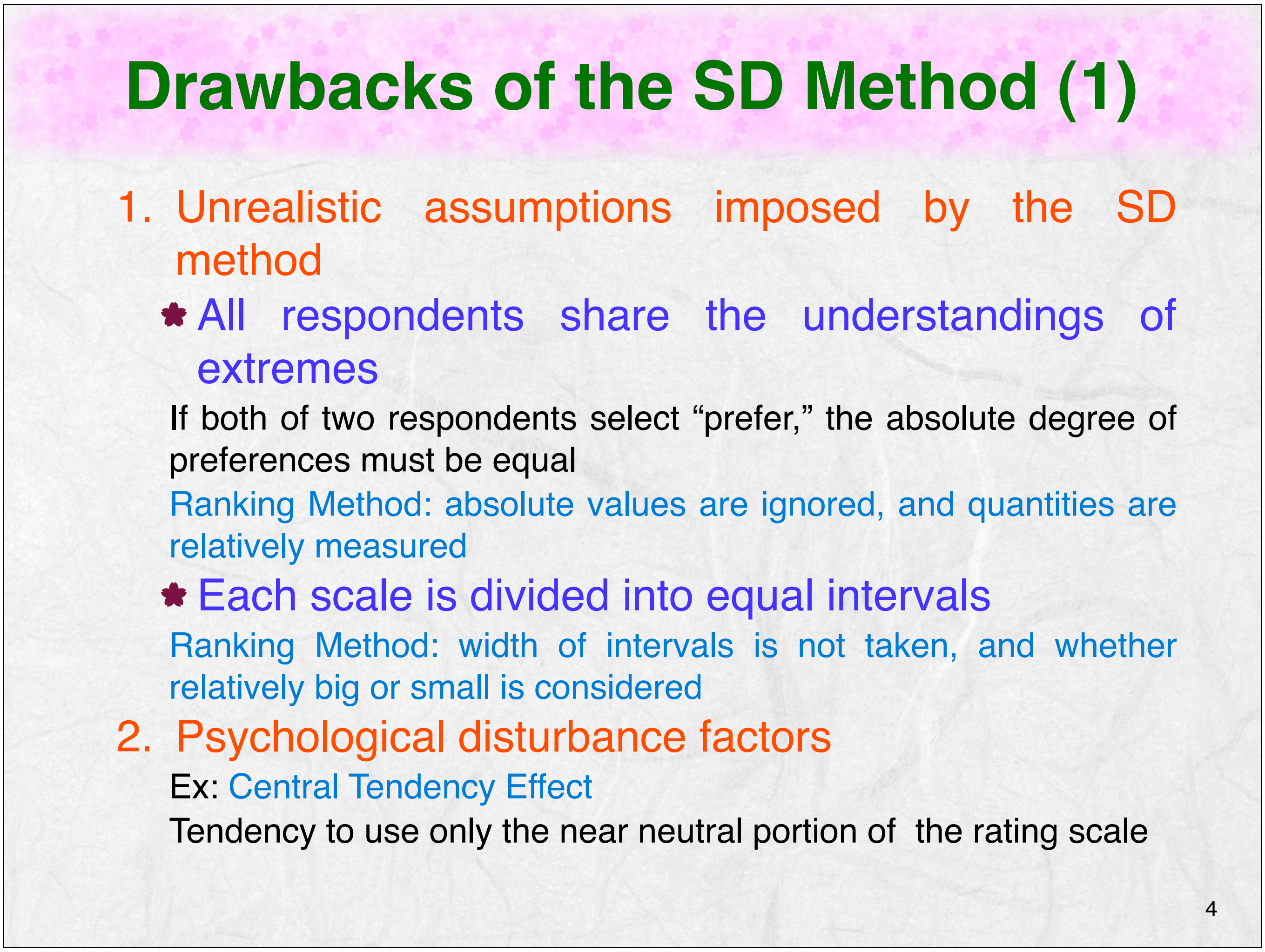
For example, if the respondent prefers Item A, then selects the “prefer” on the scale.

Instead, we use orders. This method is called “ranking method.”

In this method, the objects are sorted according to the quantities to be measured.

In this example, objects are sorted according to the respondent’s preference.

The respondent prefers Item A most, and doesn’t the item B most.



Drawbacks of the SD Method (1)

1. Unrealistic assumptions imposed by the SD method
   - All respondents share the understandings of extremes

     If both of two respondents select “prefer,” the absolute degree of preferences must be equal

     Ranking Method: absolute values are ignored, and quantities are relatively measured
   - Each scale is divided into equal intervals

     Ranking Method: width of intervals is not taken, and whether relatively big or small is considered
2. Psychological disturbance factors

   Ex: Central Tendency Effect

   Tendency to use only the near neutral portion of the rating scale

Why should we use a ranking method?

Because the SD method has the following drawbacks.

First, two unrealistic assumptions are imposed by using the SD method.

The one is that “all respondents share the understandings of extremes.”

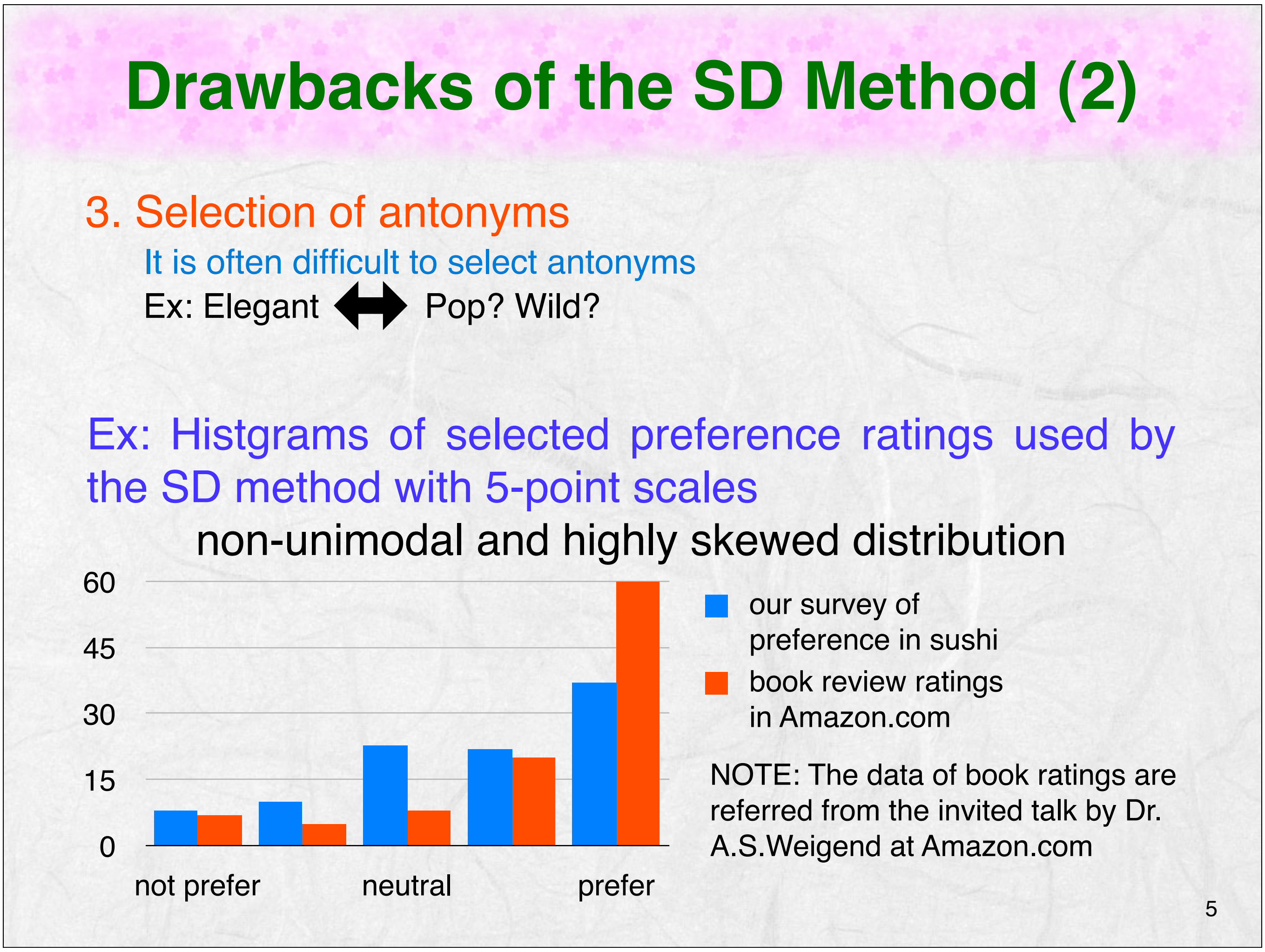
The other is that “each scale is divided into equal intervals.”

On the other hand, in the case of the ranking method, these assumptions are not required, because absolute values are ignored and quantities are relatively measured.

The second drawback is that the SD method suffers from psychological effects.

One example is Central tendency effect, that is a phenomenon respondents tend to use only the near neutral portion of rating scale.

Such effects disturbs the rating values.



Drawbacks of the SD Method (2)

3. Selection of antonyms

It is often difficult to select antonyms

Ex: Elegant ⟷ Pop? Wild?

Ex: Histgrams of selected preference ratings used by the SD method with 5-point scales

non-unimodal and highly skewed distribution

🟦 our survey of preference in sushi

🟧 book review ratings in Amazon.com

NOTE: The data of book ratings are referred from the invited talk by Dr. A.S.Weigend at Amazon.com

Third, it is often difficult to select antonyms.

For example, what should we select as the antonym of the word “elegant” ? Should we use Pop? or Wild?

To clarify these drawbacks, we show data collected by the SD method.

These are histgrams of selected preference ratings. One is our data, preferences in sushi, the other is the book review ratings in Amazon.com.

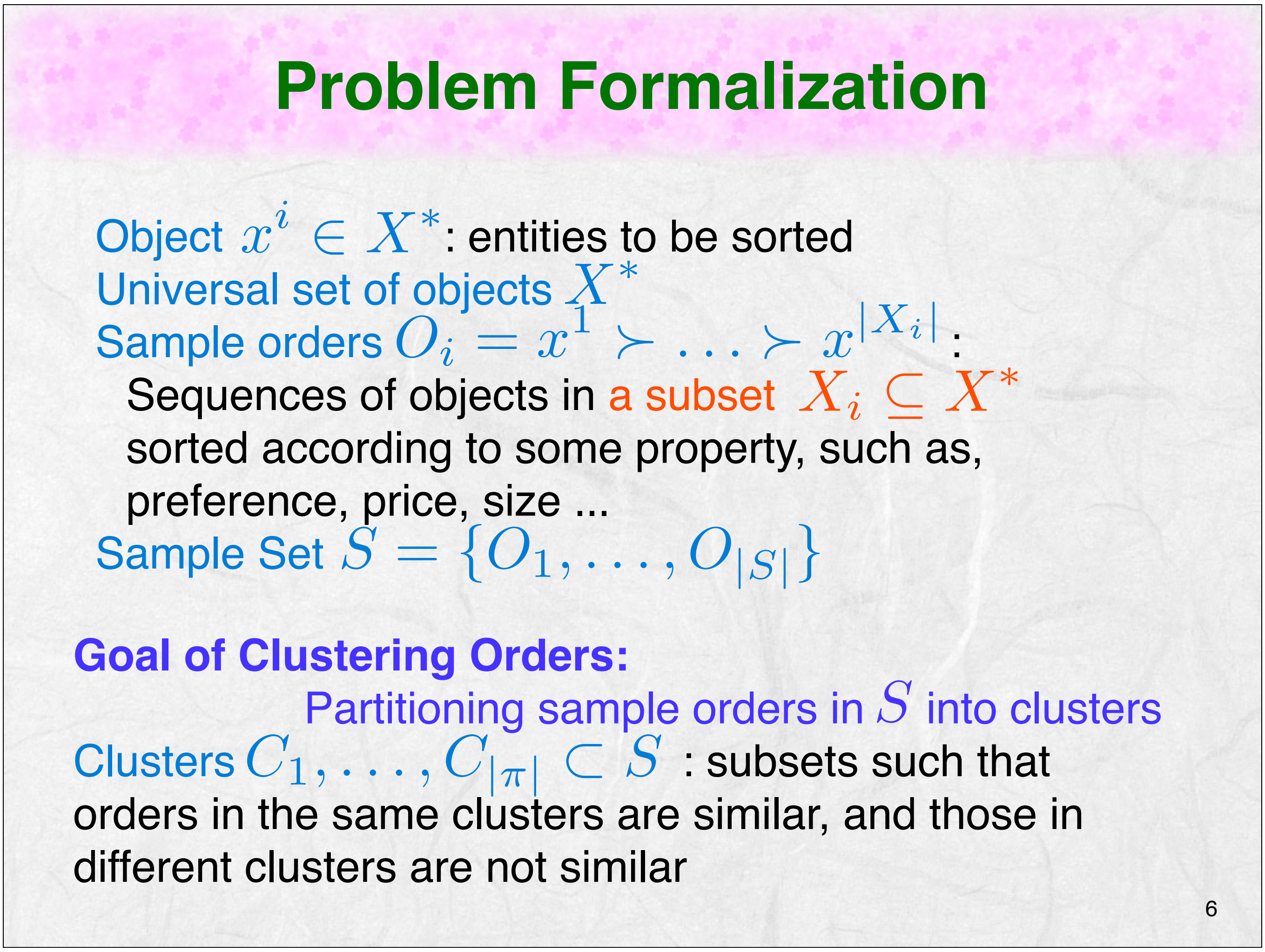
According to assumptions imposed by the SD method, ideally, these distributions should be symmetric and unimodal. However, distributions of actual data are non-unimodal and highly skewed.

We have shown several many drawbacks of the SD method, but the method is widely used.

Because many analysis techniques for rating scores have been developed.

Conversely, a few analysis techniques are available for orders, thus the ranking method has not been used.

Therefore, we develop a new method for clustering orders. We then show this method.



Problem Formalization

Object $x^i \in X^*$: entities to be sorted

Universal set of objects X^*

Sample orders $O_i = x^1 \succ \ldots \succ x^{|X_i|}$:

Sequences of objects in a subset $X_i \subseteq X^*$ sorted according to some property, such as, preference, price, size ...

Sample Set $S = \{O_1, \ldots, O_{|S|}\}$

Goal of Clustering Orders:

Partitioning sample orders in S into clusters

Clusters $C_1, \ldots, C_{|\pi|} \subset S$: subsets such that orders in the same clusters are similar, and those in different clusters are not similar

Sample orders are sorted sequences of objects according to some property, for example, preferences, prices, sizes, and so on.

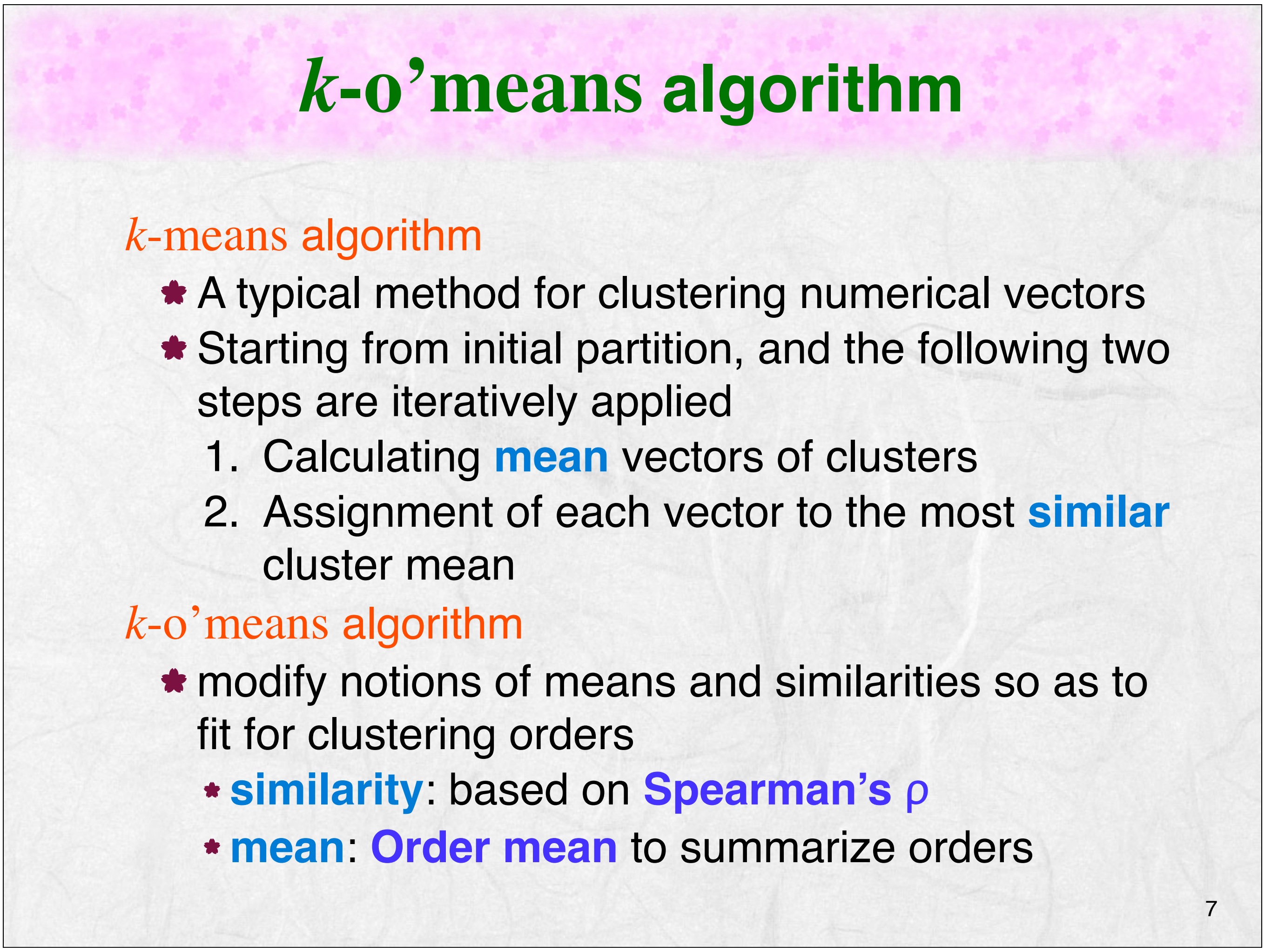
The goal of clustering orders is to partition given a set of sample orders into clusters,

such that he orders in the same cluster are similar, and those in different clusters are not similar.

Here, we want to emphasize that not all the objects are sorted.

For example, even though there are one hundred possible objects, respondents may sort only ten objects among them.

This makes it difficult to cluster orders.



k-o’means algorithm

k-means algorithm

- A typical method for clustering numerical vectors
- Starting from initial partition, and the following two steps are iteratively applied
  1. Calculating **mean** vectors of clusters
  2. Assignment of each vector to the most **similar** cluster mean

k-o’means algorithm

- modify notions of means and similarities so as to fit for clustering orders
  - **similarity**: based on **Spearman’s** ρ
  - **mean**: **Order mean** to summarize orders

We then show the algorithm for clustering orders, k-o’means.

This algorithm is the modified version of a k-means algorithm, that is a typical method for clustering numerical vectors.

In the k-means, these two steps are iteratively applied.

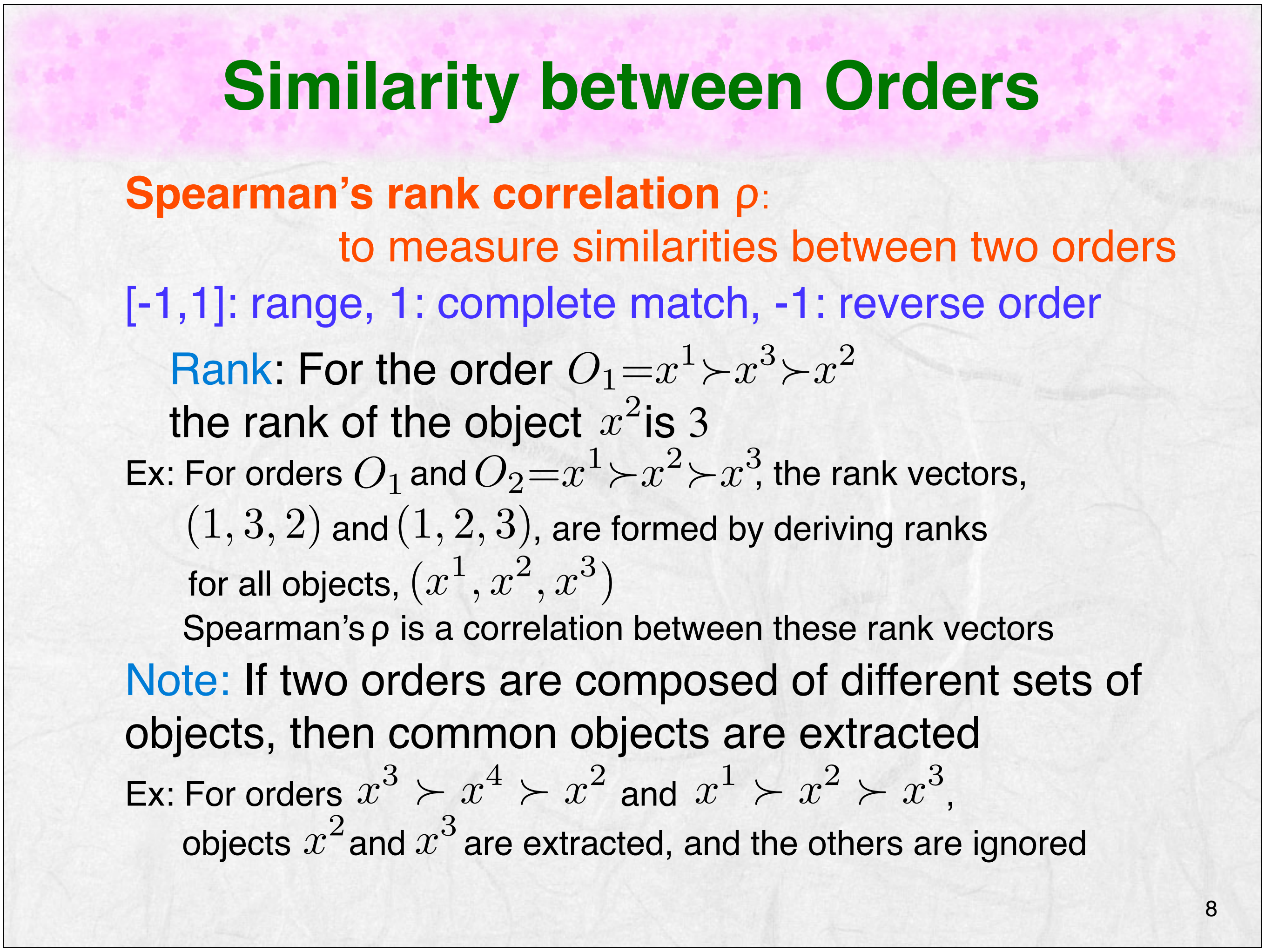
Calculating the means of clusters, and assignment each vector to the most similar cluster mean.

To modify this k-means so as to fit for clustering orders, the notions of means and similarities for orders are required.

As the similarities, we adpoted the Spearman’s rho.

And, as the means, we developed the notions of order means.

We then show these notions.



Similarity between Orders

Spearman's rank correlation ρ:

to measure similarities between two orders

[-1,1]: range, 1: complete match, -1: reverse order

Rank: For the order $O_1 = x^1 \succ x^3 \succ x^2$
the rank of the object x^2 is 3

Ex: For orders O_1 and $O_2 = x^1 \succ x^2 \succ x^3$, the rank vectors, $(1, 3, 2)$ and $(1, 2, 3)$, are formed by deriving ranks for all objects, (x^1, x^2, x^3)

Spearman's ρ is a correlation between these rank vectors

Note: If two orders are composed of different sets of objects, then common objects are extracted

Ex: For orders $x^3 \succ x^4 \succ x^2$ and $x^1 \succ x^2 \succ x^3$, objects x^2 and x^3 are extracted, and the others are ignored

The Spearman's rho is a well-known statistics to measure the similarities between two orders.

This is defined as correlation coefficients between ranks of objects.

The rank is the number to represent the position of the object in the specified order.

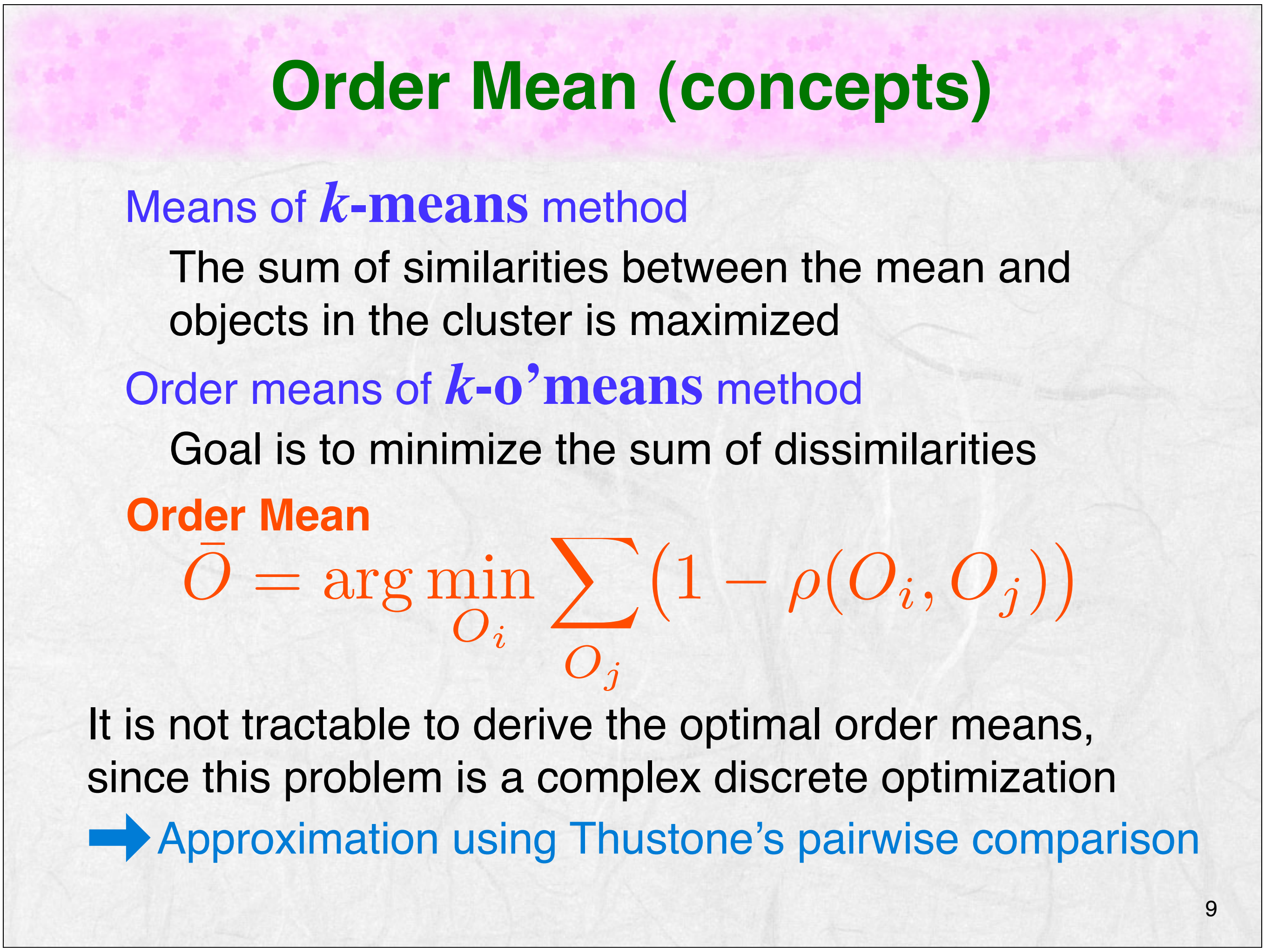
For example, in the order O1, the rank of the object x2 is 3.

Rank vectors are formed by deriving ranks for all objects.

The rho is defined as correlation between these rank vectors.

It should be notice that, if two orders are composed of different sets of objects, then only common objects are extracted.

For example, two orders, 3 4 2 and 1 2 3 are given, objects x2 and x3 are extracted, and the other objects are ignored.



Order Mean (concepts)

Means of ***k*-means** method

The sum of similarities between the mean and objects in the cluster is maximized

Order means of ***k*-o'means** method

Goal is to minimize the sum of dissimilarities

Order Mean

$$\bar{O} = \arg\min_{O_i} \sum_{O_j} (1 - \rho(O_i, O_j))$$

It is not tractable to derive the optimal order means, since this problem is a complex discrete optimization

➡ Approximation using Thustone's pairwise comparison

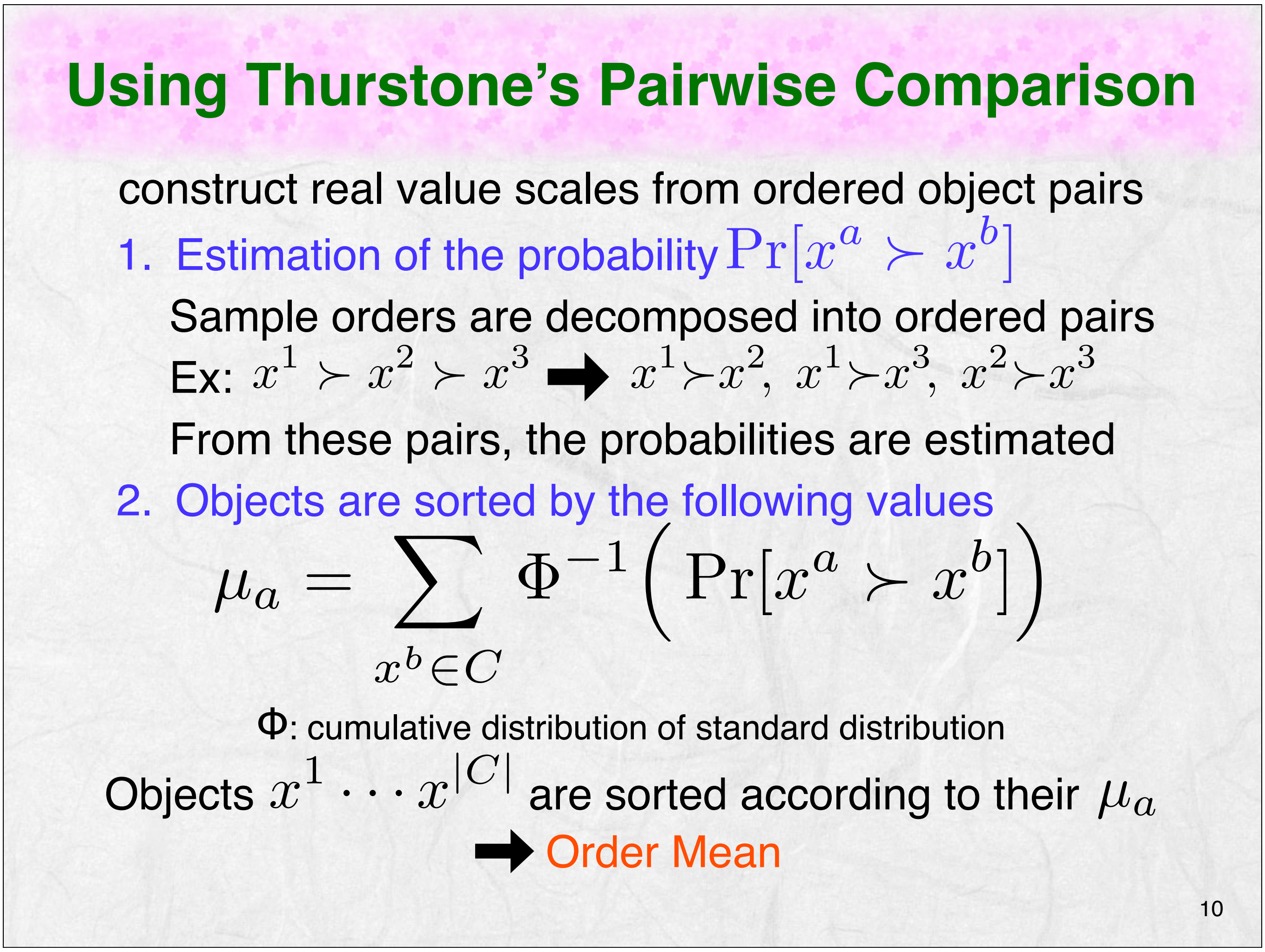
We turn to the notion of the order mean.

Let me remind you that means of the k-means are defined so as to maximize the sum of similarities between the mean and objects in the cluster.

By analogy, order means are defined so as to minimize the sum of dissimilarities.

Unfortunately, it is not tractable to derive the optimal order means, since this problem is a complex discrete optimization.

Therefore, we adopted approximations using the Thurstone's pairwise comparison.



Using Thurstone’s Pairwise Comparison

construct real value scales from ordered object pairs

1. Estimation of the probability $\Pr[x^a \succ x^b]$

   Sample orders are decomposed into ordered pairs

   Ex: $x^1 \succ x^2 \succ x^3$ ➡ $x^1 \succ x^2,\ x^1 \succ x^3,\ x^2 \succ x^3$

   From these pairs, the probabilities are estimated

2. Objects are sorted by the following values

$$\mu_a = \sum_{x^b \in C} \Phi^{-1}\Big(\Pr[x^a \succ x^b]\Big)$$

Φ: cumulative distribution of standard distribution

Objects $x^1 \cdots x^{|C|}$ are sorted according to their μ_a

➡ Order Mean

Thurstone’s pairwise comparison is a method to construct a real value scale from ordered object pairs.

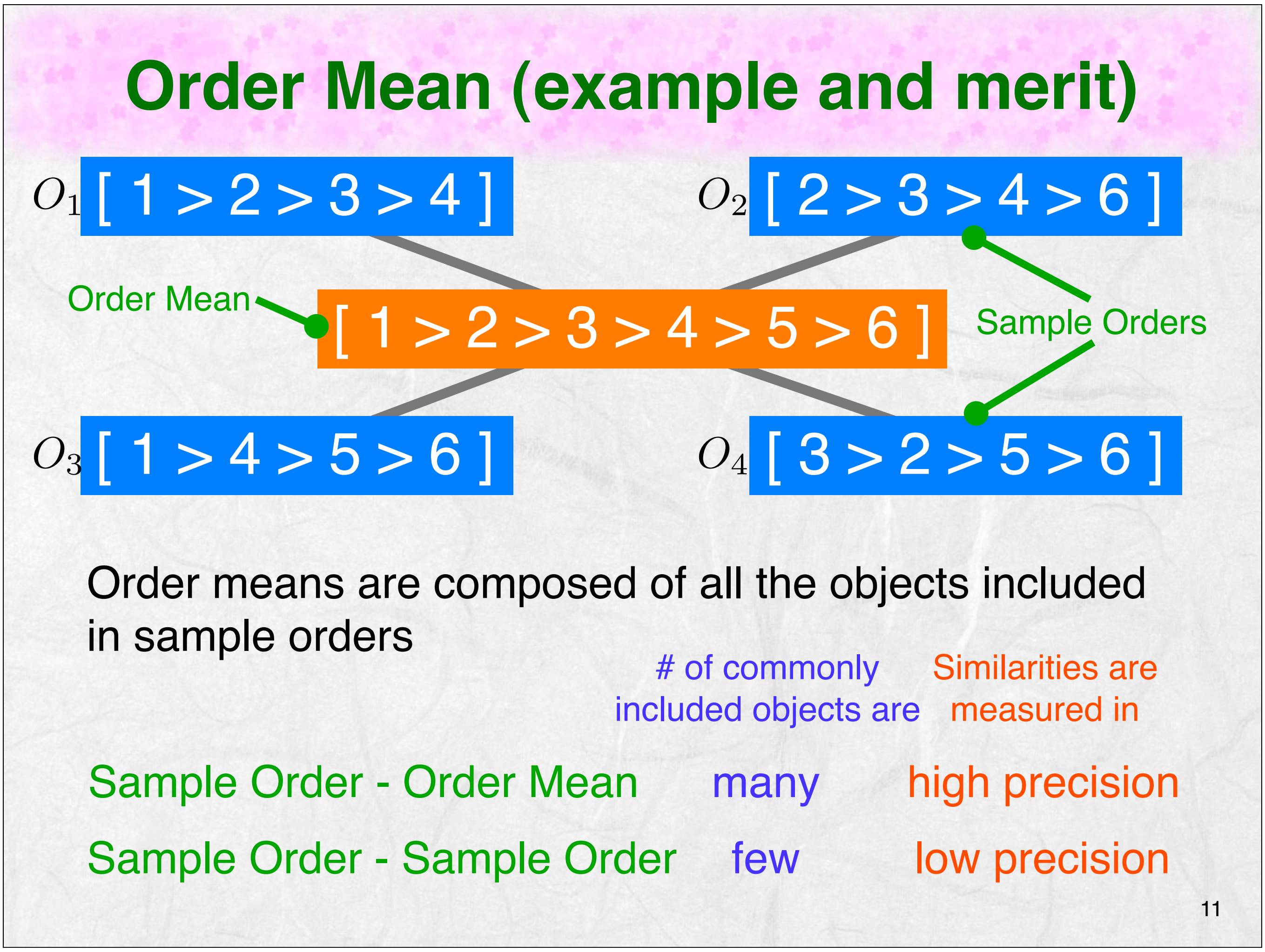
First, the probability the object x_a precedes x_b is estimated.

Sample orders are decomposed into ordered object pairs.

Then, from the frequency of these pairs, the probabilities are estimated.

Second, from these probabilities, these mu_a are derived for each of possible objects.

By sorting objects according to their mu_a, the order mean is derived.



Order Mean (example and merit)

O_1 [1 > 2 > 3 > 4]

O_2 [2 > 3 > 4 > 6]

Order Mean

[1 > 2 > 3 > 4 > 5 > 6]

Sample Orders

O_3 [1 > 4 > 5 > 6]

O_4 [3 > 2 > 5 > 6]

Order means are composed of all the objects included in sample orders

	# of commonly included objects are	Similarities are measured in
Sample Order - Order Mean	many	high precision
Sample Order - Sample Order	few	low precision

We give an example of an order mean.

4 sample orders are given, and this order mean is derived.

Here, it should be notice that order means are composed of all the objects included in sample orders.

Therefore, when measuring similarities between a sample order and an order mean, there are many common objects between them, and the similarities are precisely measured.

On the other hand, when measuring similarities between two sample orders, there may be only a few common objects.

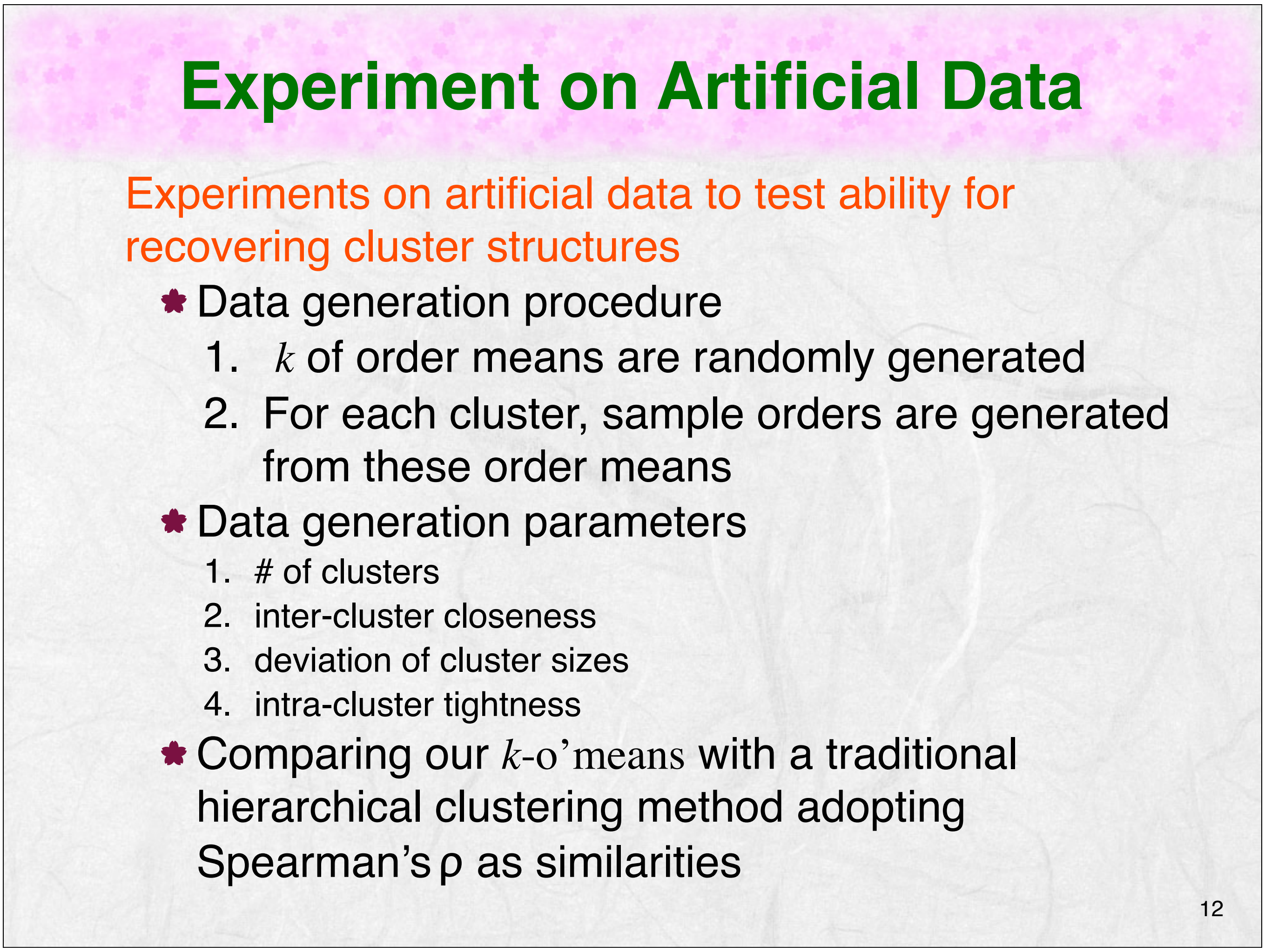
In such cases, similarities cannot be measured precisely any more.

This is the most important merit of introducing order means.

We have shown our k-o'means algorithm. We next apply this algorithm to two types of data.

One is artificial data to clarify the advantages of our method.

The other is real data, that are questionnaire survey about preference in sushi.



Experiment on Artificial Data

Experiments on artificial data to test ability for recovering cluster structures

- Data generation procedure
  1. *k* of order means are randomly generated
  2. For each cluster, sample orders are generated from these order means
- Data generation parameters
  1. # of clusters
  2. inter-cluster closeness
  3. deviation of cluster sizes
  4. intra-cluster tightness
- Comparing our *k*-o'means with a traditional hierarchical clustering method adopting Spearman's ρ as similarities

To test our k-o'means algorithm, it is applied to artificial data.

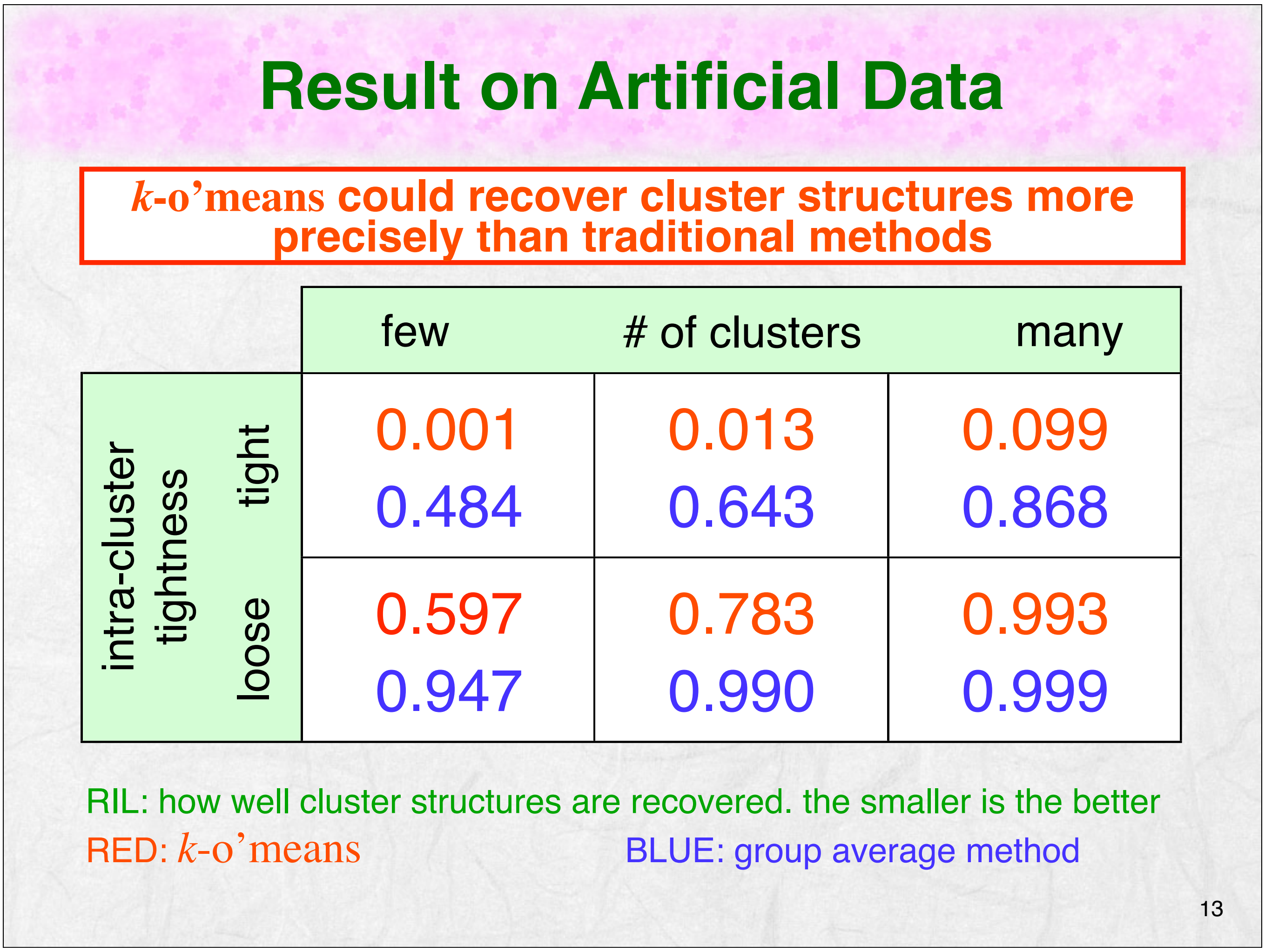
To generate data, first, k of order means are randomly generated.

Then, for each cluster, sample orders are generated from these order means.

We tested whether our k-o'means can recover these imposed cluster structures or not.

For comparison, a group average method is also applied.

This is a traditional hierarchical clustering method, and can be used for clustering orders by adopting Spearman's rho as similarities.



Result on Artificial Data

***k*-o'means could recover cluster structures more precisely than traditional methods**

		few	# of clusters	many
intra-cluster tightness	tight	0.001 0.484	0.013 0.643	0.099 0.868
	loose	0.597 0.947	0.783 0.990	0.993 0.999

RIL: how well cluster structures are recovered. the smaller is the better

RED: *k*-o'means BLUE: group average method

We show RIL, that presents how well imposed cluster structures are recovered.

The smaller is the better.

Red entries show results by our k-o'means, and Blue entries show results by a group average method.

Clearly, our k-o'means is superior.

As shown before, our k-o'means adopted notions of order means.

So, similarities between an order mean and a sample order can be measured precisely.

However, in the case of a group average method, since similarities are measured between a pair of sample orders, its precision becomes low.

Therefore, our k-o'means could recover cluster structures more precisely than traditional methods.

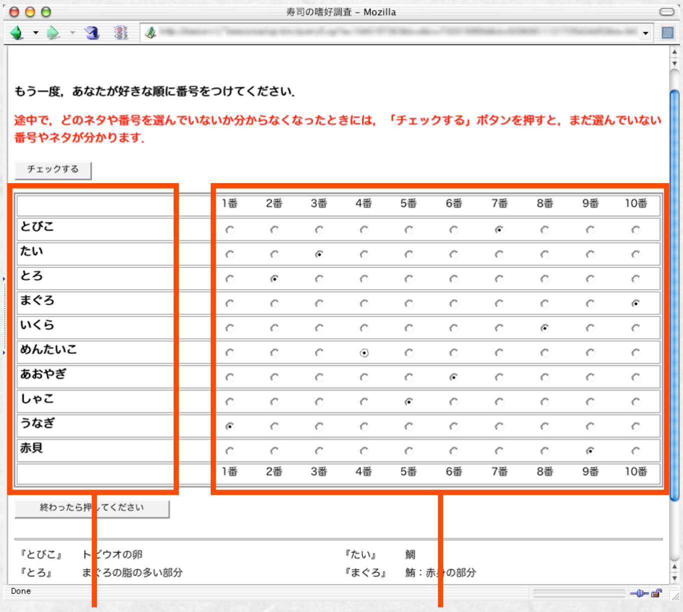
Experiment on Survey Data

Questionnaire survey data about preference in sushi

- 100 kinds of sushi
- 10 randomly selected sushi are shown to each respondent
- Each respondent select ranks in terms of his/her preference in sushi
- # of respondents = 1025

Find clusters of respondents having similar preference, and analyze characteristics of these clusters

WWW form to input orders



寿司の嗜好調査 – Mozilla

もう一度，あなたが好きな順に番号をつけてください．

途中で，どのネタや番号を選んでいないか分からなくなったときには，「チェックする」ボタンを押すと，まだ選んでいない番号やネタが分かります．

チェックする

	1番	2番	3番	4番	5番	6番	7番	8番	9番	10番
とびこ	○	○	○	○	○	○	●	○	○	○
たい	○	○	●	○	○	○	○	○	○	○
とろ	○	●	○	○	○	○	○	○	○	○
まぐろ	○	○	○	○	○	○	○	○	○	●
いくら	○	○	○	○	○	○	○	●	○	○
めんたいこ	○	○	○	●	○	○	○	○	○	○
あおやぎ	○	○	○	○	○	●	○	○	○	○
しゃこ	○	○	○	○	●	○	○	○	○	○
うなぎ	●	○	○	○	○	○	○	○	○	○
赤貝	○	○	○	○	○	○	○	○	●	○
	1番	2番	3番	4番	5番	6番	7番	8番	9番	10番

終わったら押してください

『とびこ』	トビウオの卵	『たい』	鯛
『とろ』	まぐろの脂の多い部分	『まぐろ』	鮪：赤身の部分

Done

sushi name

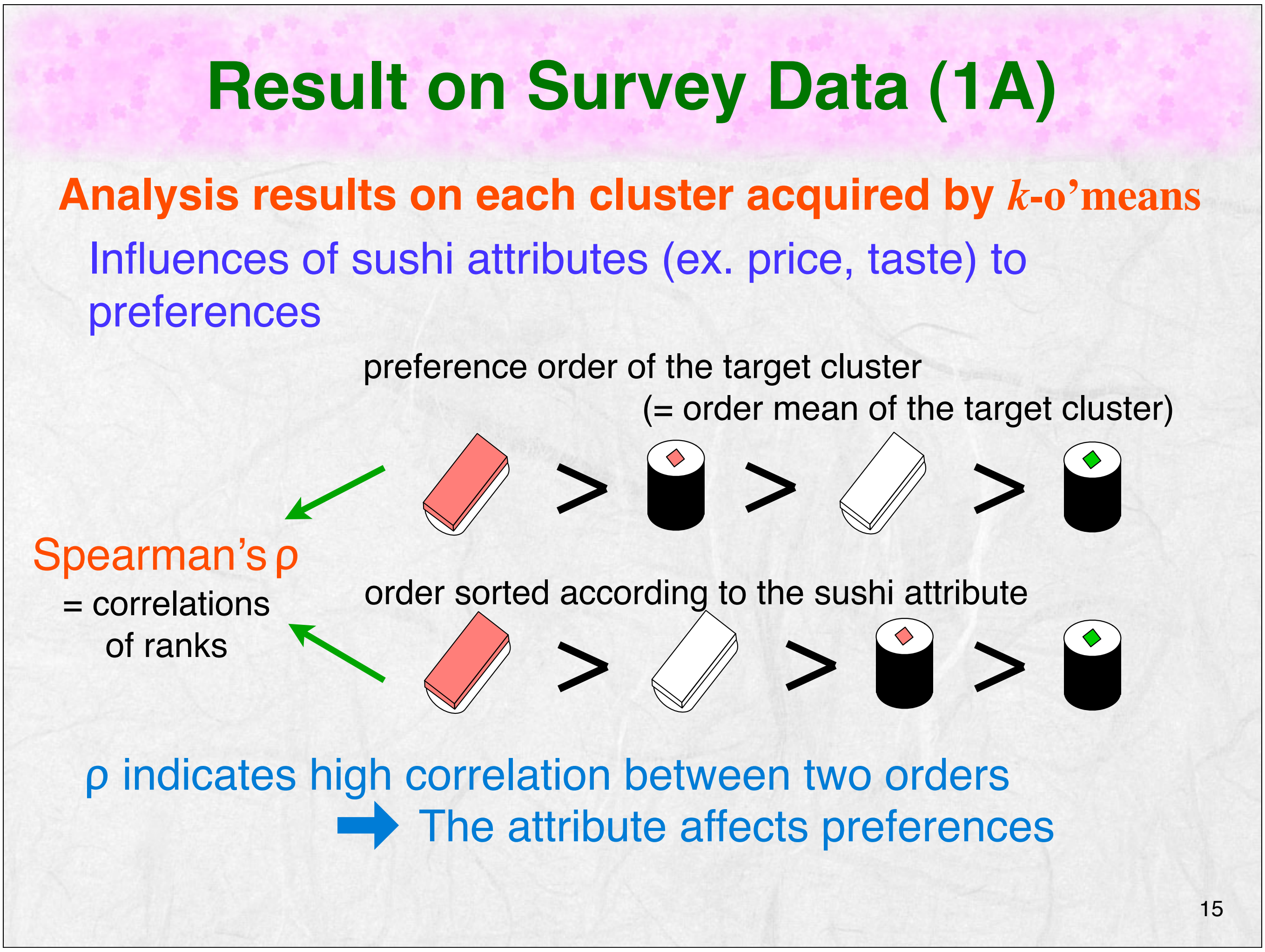
select ranks

We then apply our k-o’means to real data, that is questionnaire survey data about preference in sushi.

10 randomly selected sushi are shown to each respondent.

Each respondent select ranks in terms of preference in sushi.

We found clusters of respondents having similar preference, and analyzed characteristics of these clusters.



Result on Survey Data (1A)

Analysis results on each cluster acquired by *k*-o'means

Influences of sushi attributes (ex. price, taste) to preferences

preference order of the target cluster
(= order mean of the target cluster)

Spearman's ρ
= correlations of ranks

order sorted according to the sushi attribute

ρ indicates high correlation between two orders
➡ The attribute affects preferences

First, we analyzed influences of sushi attributes to preferences.

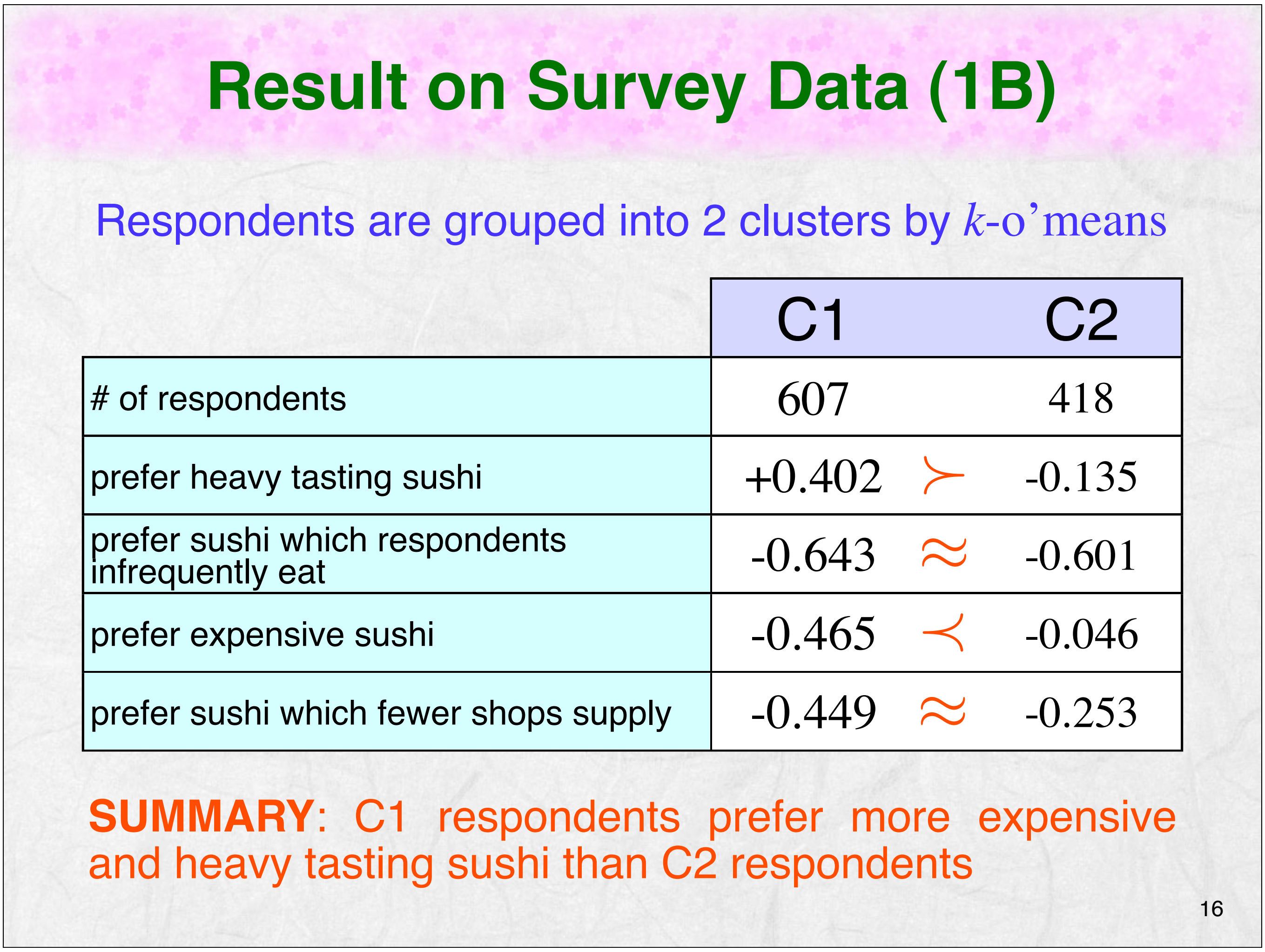
We mean attributes, for example, prices or taste of sushi.

For this aim, we compared the two orders.

The one is the preference order of the target cluster, that is, the order mean of the cluster.

The other is the order sorted according to the sushi attribute.

If these two orders are correlated, it can be concluded that the attribute affects preferences.



Result on Survey Data (1B)

Respondents are grouped into 2 clusters by *k*-o'means

	C1		C2
# of respondents	607		418
prefer heavy tasting sushi	+0.402	≻	-0.135
prefer sushi which respondents infrequently eat	-0.643	≈	-0.601
prefer expensive sushi	-0.465	≺	-0.046
prefer sushi which fewer shops supply	-0.449	≈	-0.253

SUMMARY: C1 respondents prefer more expensive and heavy tasting sushi than C2 respondents

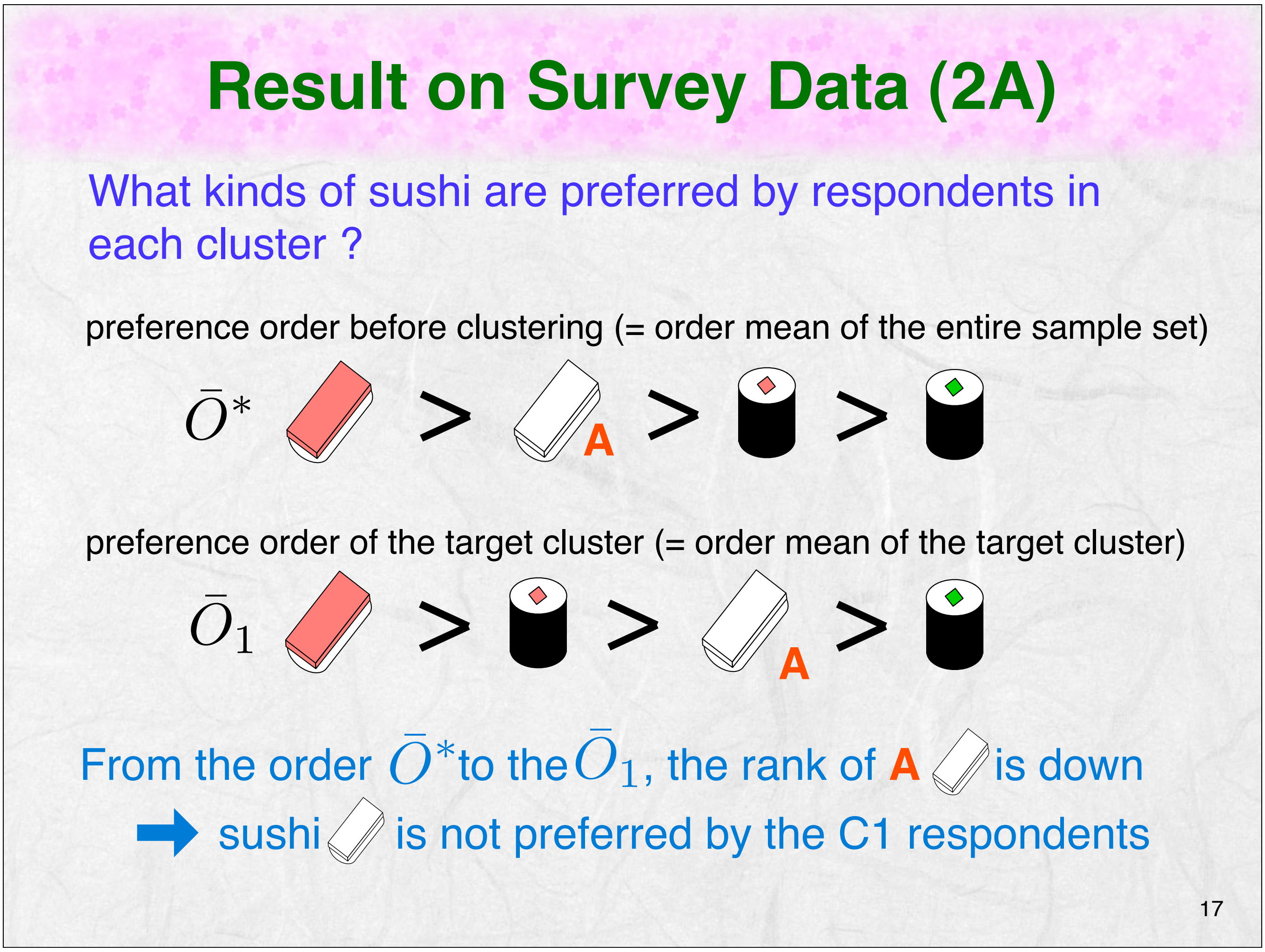
Respondents are grouped into 2 clusters by our k-o'means.

These are correlation between preferences and attributes.

In terms of heaviness, or oiliness, of sushi, the correlation of the cluster C1 is much higher than that of C2.

So, C1 respondents prefer heavy tasting sushi.

In summary, C1 respondents prefer more expensive and heavy tasting sushi than C2.



Result on Survey Data (2A)

What kinds of sushi are preferred by respondents in each cluster ?

preference order before clustering (= order mean of the entire sample set)

$\bar{O}^{*}$ > A > >

preference order of the target cluster (= order mean of the target cluster)

$\bar{O}_{1}$ > > A >

From the order $\bar{O}^{*}$ to the $\bar{O}_{1}$, the rank of A is down

➡ sushi is not preferred by the C1 respondents

Next, I explored what kinds of sushi are preferred by respondents in each cluster.

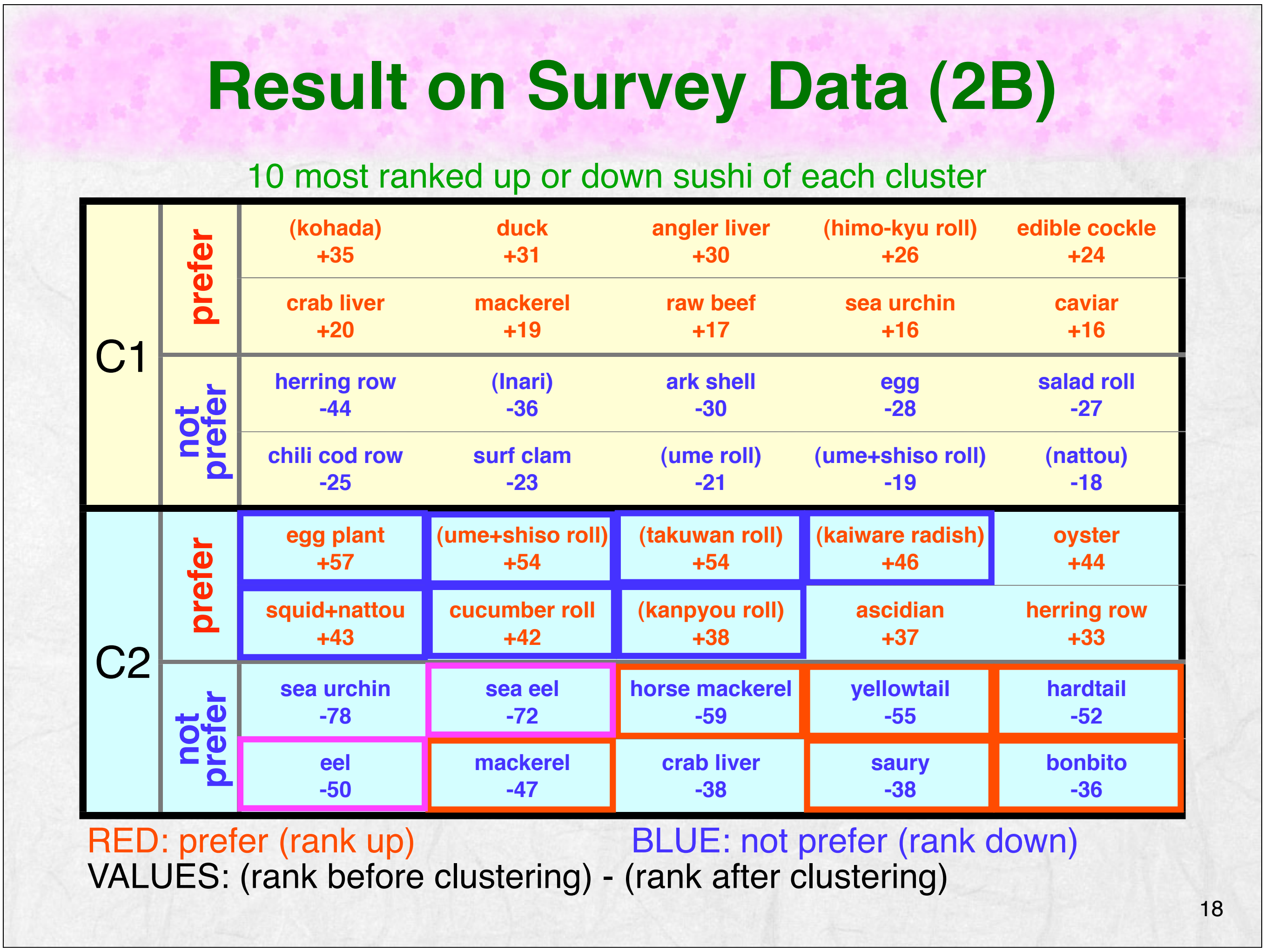
To this aim, for each kind of sushi, we checked its rank in the following two orders.

The one is the preference order before clustering, that is the order mean of the entire sample set.

We think this order represents the neutral preferences.

The other is the preference order after clustering, that is the order mean of the target cluster.

For example, if the rank of some sushi A is down, it can be concluded that the sushi A is not preferred.



Result on Survey Data (2B)

10 most ranked up or down sushi of each cluster

C1	prefer	(kohada) +35	duck +31	angler liver +30	(himo-kyu roll) +26	edible cockle +24
		crab liver +20	mackerel +19	raw beef +17	sea urchin +16	caviar +16
	not prefer	herring row -44	(Inari) -36	ark shell -30	egg -28	salad roll -27
		chili cod row -25	surf clam -23	(ume roll) -21	(ume+shiso roll) -19	(nattou) -18
C2	prefer	egg plant +57	(ume+shiso roll) +54	(takuwan roll) +54	(kaiware radish) +46	oyster +44
		squid+nattou +43	cucumber roll +42	(kanpyou roll) +38	ascidian +37	herring row +33
	not prefer	sea urchin -78	sea eel -72	horse mackerel -59	yellowtail -55	hardtail -52
		eel -50	mackerel -47	crab liver -38	saury -38	bonbito -36

RED: prefer (rank up) BLUE: not prefer (rank down)

VALUES: (rank before clustering) - (rank after clustering)

We picked up 10 most ranked up and down sushi of each cluster.

Red entries show preferred, that is, ranked up sushi.

Blue entries show not preferred, that is, ranked down sushi.

Since the rank changes are larger, we observe C2 cluster.

[PUSH] These (red square) are so-called “blue fish,” rather oily and smelly.

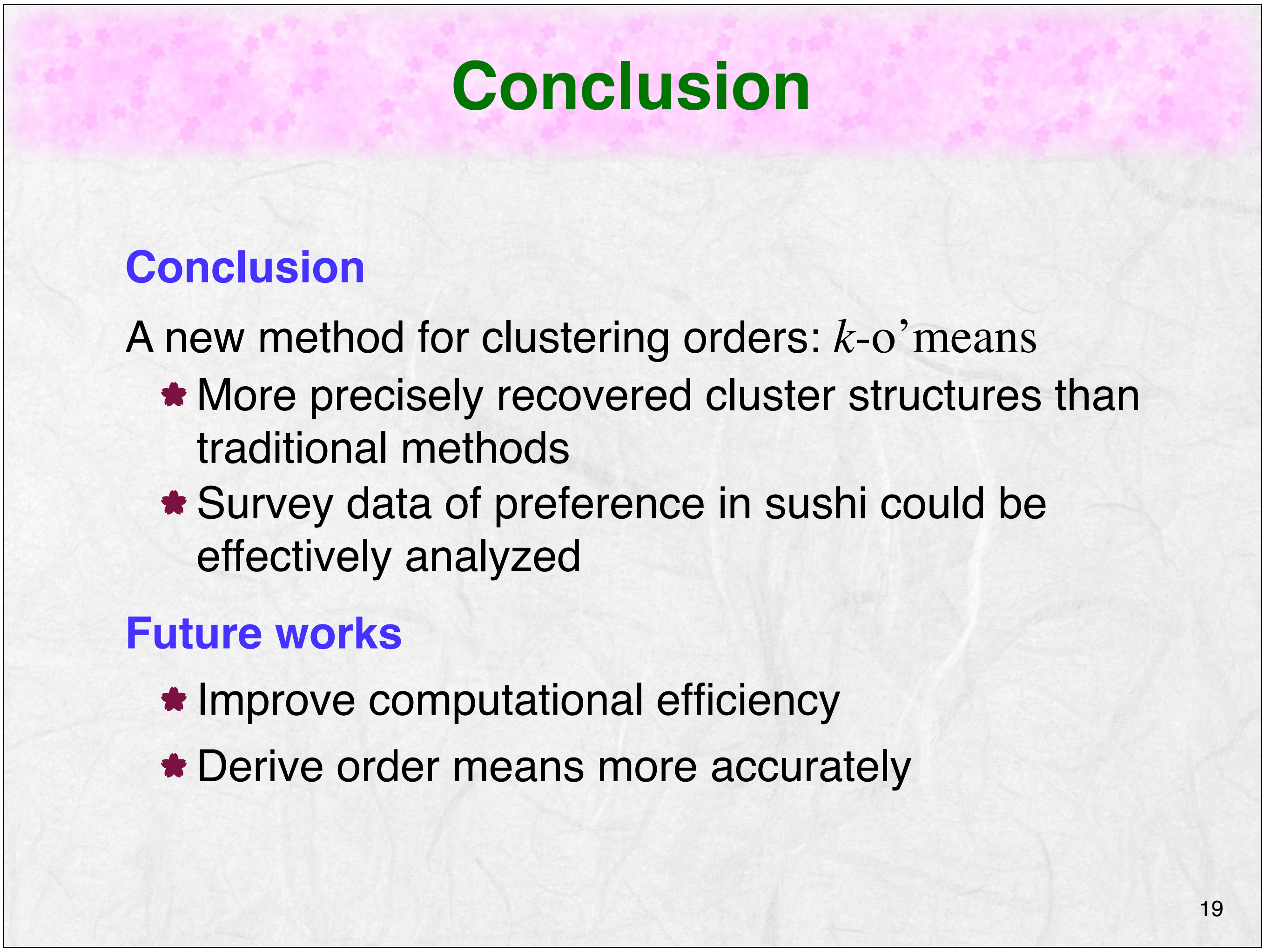
[PUSH] And these (magenta square) eels are very oily.

This result agrees with the previous result that C2 respondents don’t prefer oily sushi.

[PUSH] These (blue square) are very economic sushi.

These are ranked up, but are still in the neutral portions of the preference order.

Therefore, it should say that C2 respondents doesn’t dislike these sushi.



Conclusion

Conclusion

A new method for clustering orders: k-o'means

- More precisely recovered cluster structures than traditional methods
- Survey data of preference in sushi could be effectively analyzed

Future works

- Improve computational efficiency
- Derive order means more accurately

We would like to conclude our talk.

We proposed a method for clustering orders: k-o'means.

This method could more precisely recover cluster structures than traditional methods

Survey data of preference in sushi can be effectively analyzed.

We plan to improve computational efficiency, and develop a method that can derive order means more accirately.

That's all we have to say. Thank you.