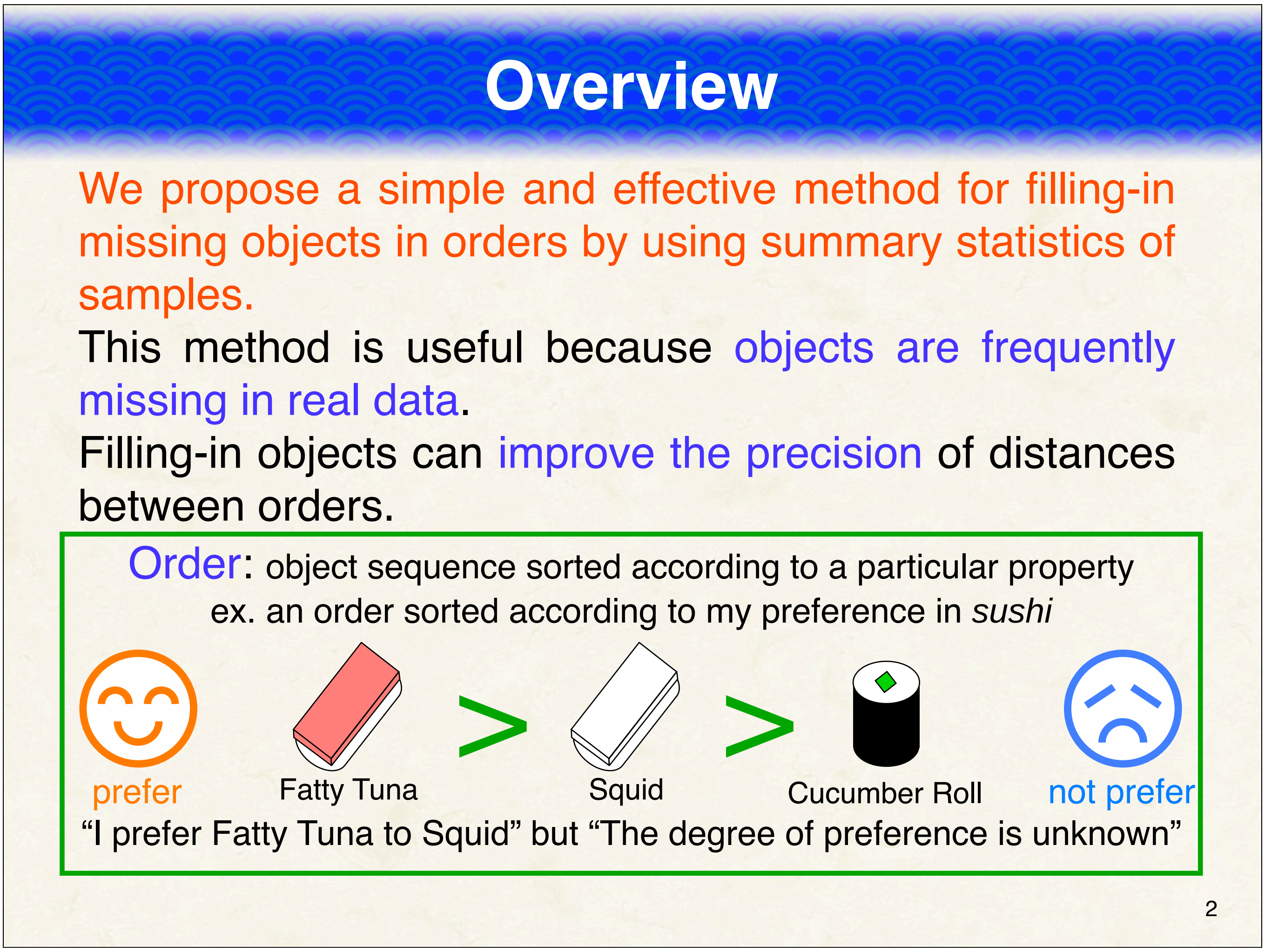
# Filling-in Missing Objects in Orders

Toshihiro Kamishima and Shotaro Akaho
`http://www.kamishima.net/`
National Institute of AIST, Japan

ICDM2004, Brighton, U.K., 1-4/11/2004

We would like to talk about a method for filling-in missing objects in orders.



# Overview

We propose a simple and effective method for filling-in missing objects in orders by using summary statistics of samples.

This method is useful because objects are frequently missing in real data.

Filling-in objects can improve the precision of distances between orders.

Order: object sequence sorted according to a particular property

ex. an order sorted according to my preference in *sushi*

“I prefer Fatty Tuna to Squid” but “The degree of preference is unknown”

We propose a simple and method for filling-in missing objects in orders by using summary statistics of samples.

Such techniques have been developed for numerical or categorical values, but have not for orders.

This method is useful because objects are frequently missing in real data.

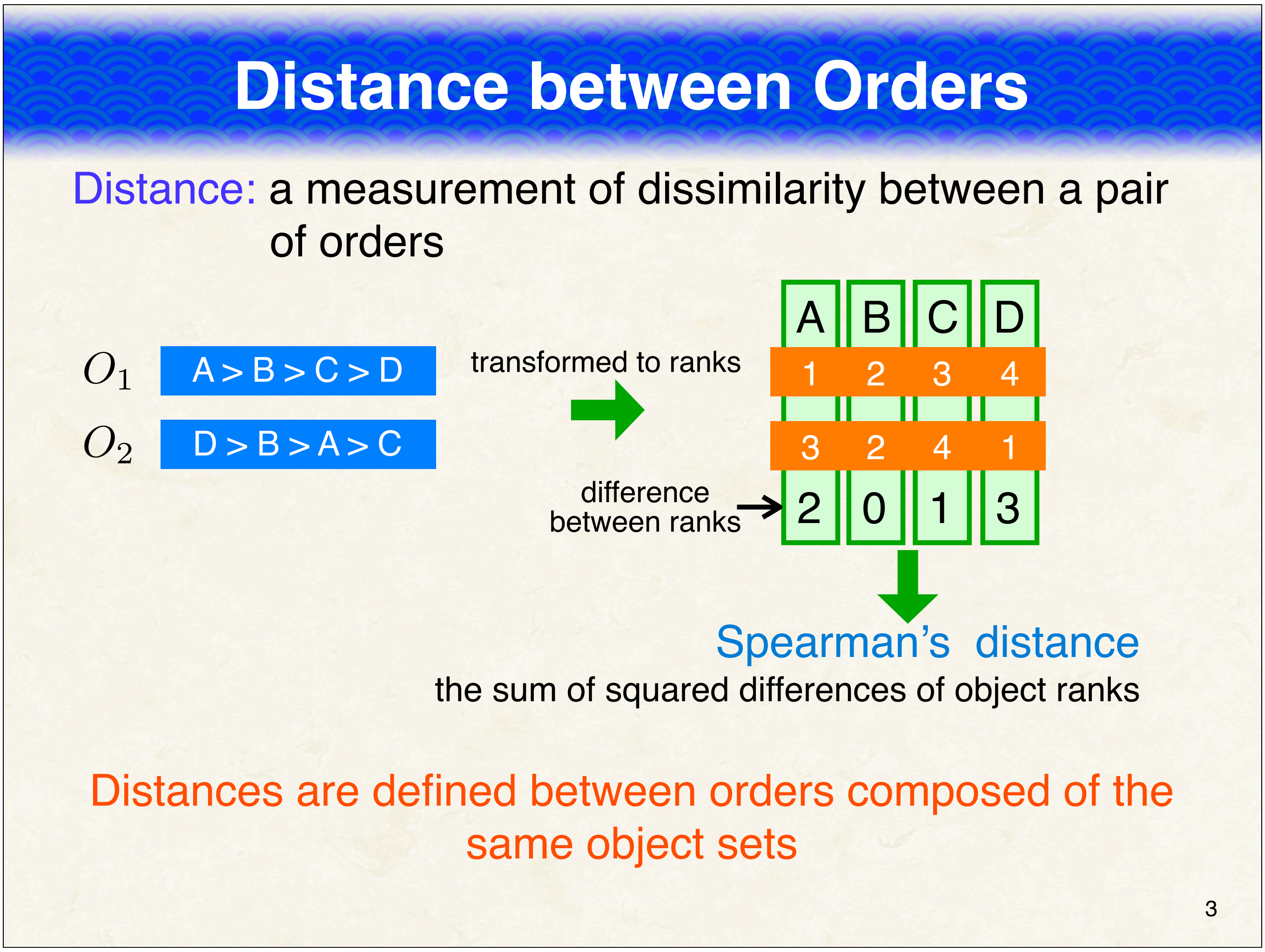
Filling-in objects can improve the precision of distances between orders.

We begin with what is an order.

An order is an object sequence sorted according to a particular property.

For example, an order sorted according to my preference in sushi.

This order indicates that “I prefer a fatty tuna to squid”, but “The degree of preference is unknown.”



# Distance between Orders

Distance: a measurement of dissimilarity between a pair of orders

$O_1$ A > B > C > D

$O_2$ D > B > A > C

transformed to ranks

|  | A | B | C | D |
|---|---|---|---|---|
| $O_1$ | 1 | 2 | 3 | 4 |
| $O_2$ | 3 | 2 | 4 | 1 |
| difference between ranks | 2 | 0 | 1 | 3 |

Spearman's distance

the sum of squared differences of object ranks

Distances are defined between orders composed of the same object sets

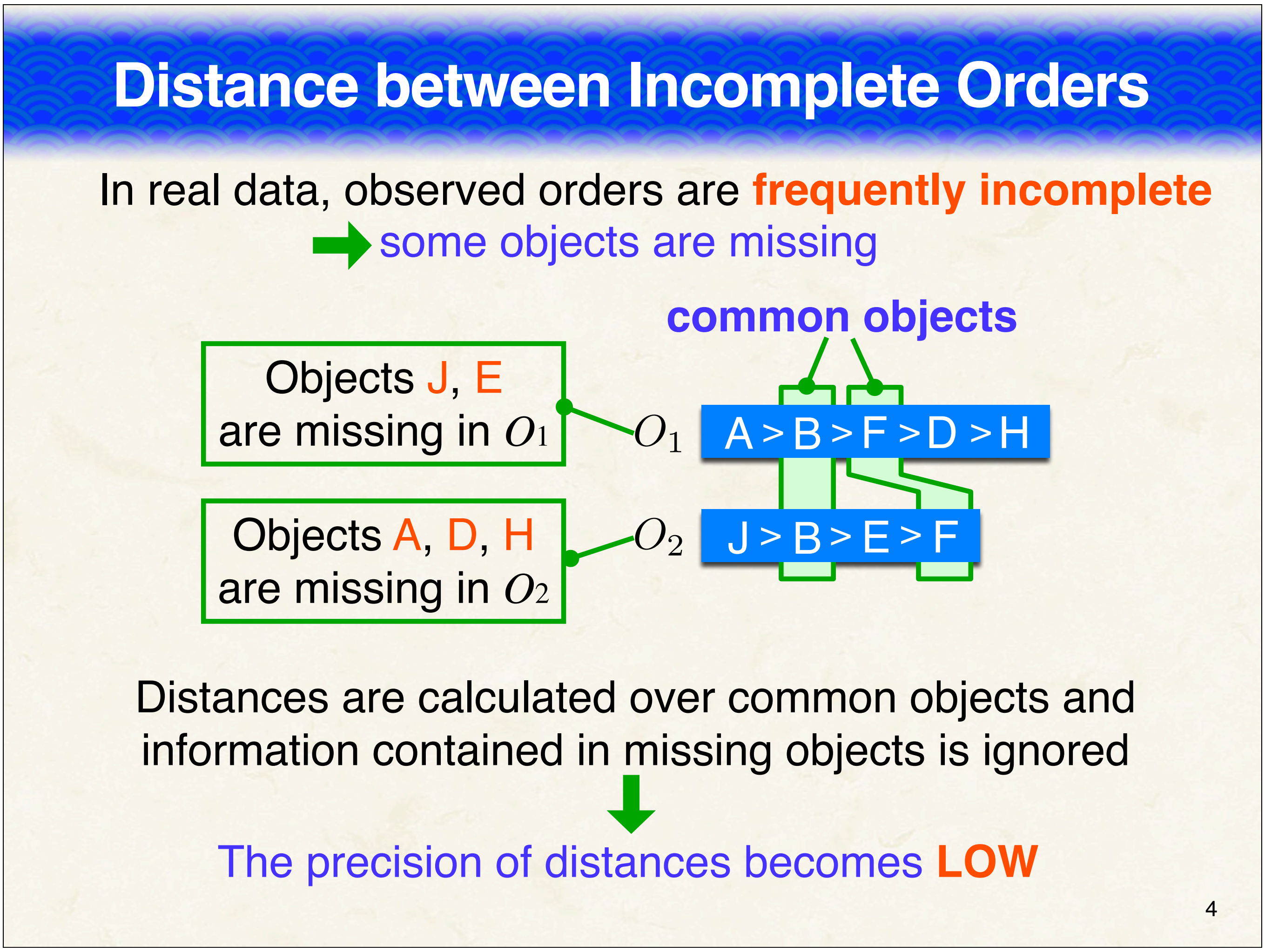
First, we will show what is a filling-in technique and why it is required.

This technique is useful when measuring distance between a pair of orders.

An example of distance between orders is Spearman's distance, that is defined as the sum of squared differences of object ranks.

Though many kinds of distances for orders have been proposed, these are defined between orders composed of the same object sets.

In this example, both of O1 and O2 are composed of the same object set, {A, B, C, D}.



# Distance between Incomplete Orders

In real data, observed orders are **frequently incomplete**

➡ some objects are missing

**common objects**

Objects J, E are missing in $O_1$

$O_1$ A > B > F > D > H

Objects A, D, H are missing in $O_2$

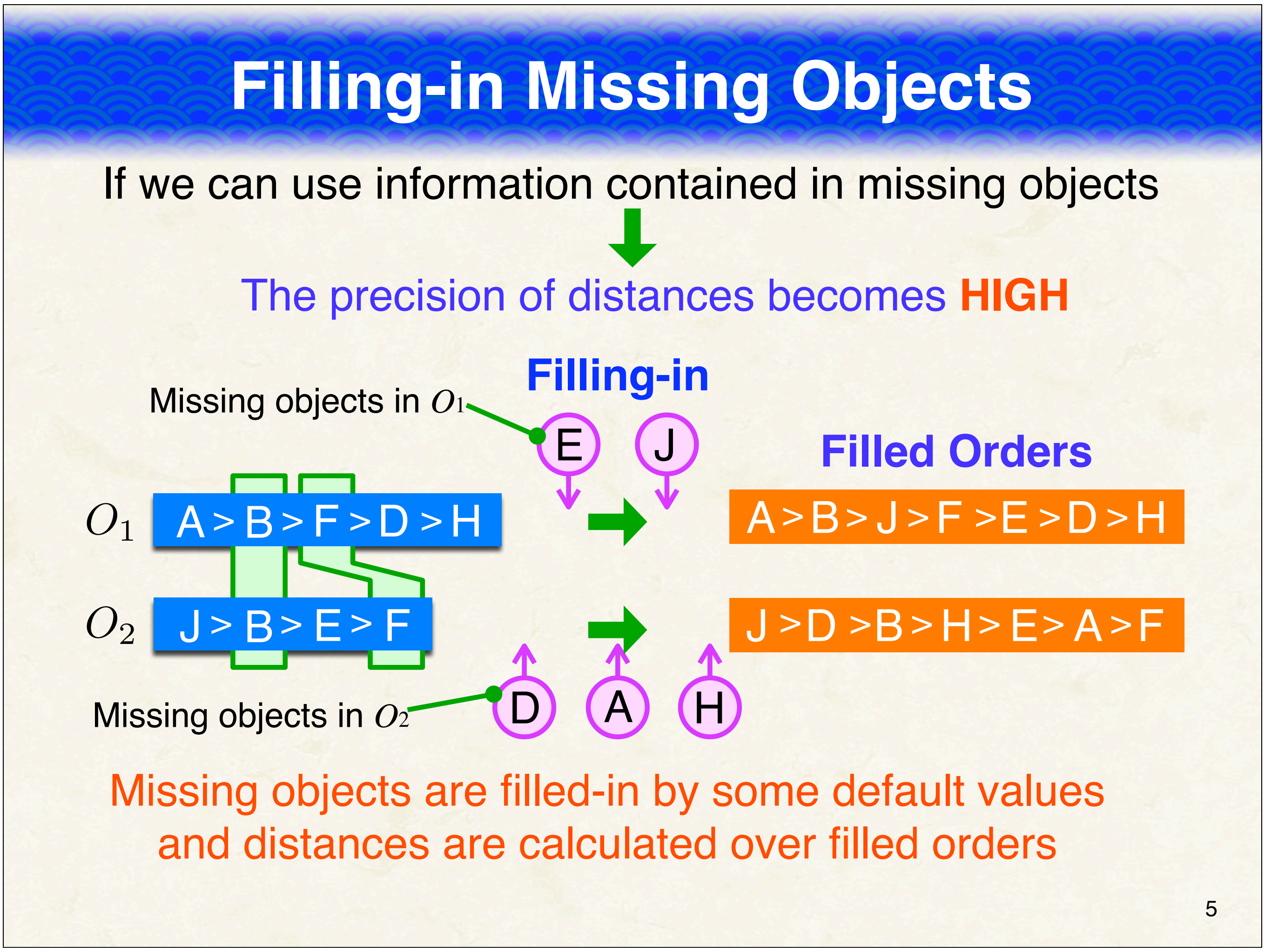
$O_2$ J > B > E > F

Distances are calculated over common objects and information contained in missing objects is ignored

⬇

The precision of distances becomes **LOW**

In cases of real data, observed orders are frequently incomplete.
That is to say, some objects are missing.
For example, object J and E are contained in the order O2, but not in the O1.
In this case, we say that object J and E are missing in O1.
Consider to calculate the distance between incomplete orders.
The distances are defined between orders that composed of the same object sets, so distances are of necessity calculated over common objects and potentially useful information contained in missing objects is ignored.
Therefore, the precision of distances becomes low.



# Filling-in Missing Objects

If we can use information contained in missing objects

The precision of distances becomes **HIGH**

Filling-in

Missing objects in $O_1$

E J

Filled Orders

$O_1$ A > B > F > D > H

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

$O_2$ J > B > E > F

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

Missing objects in $O_2$

D A H

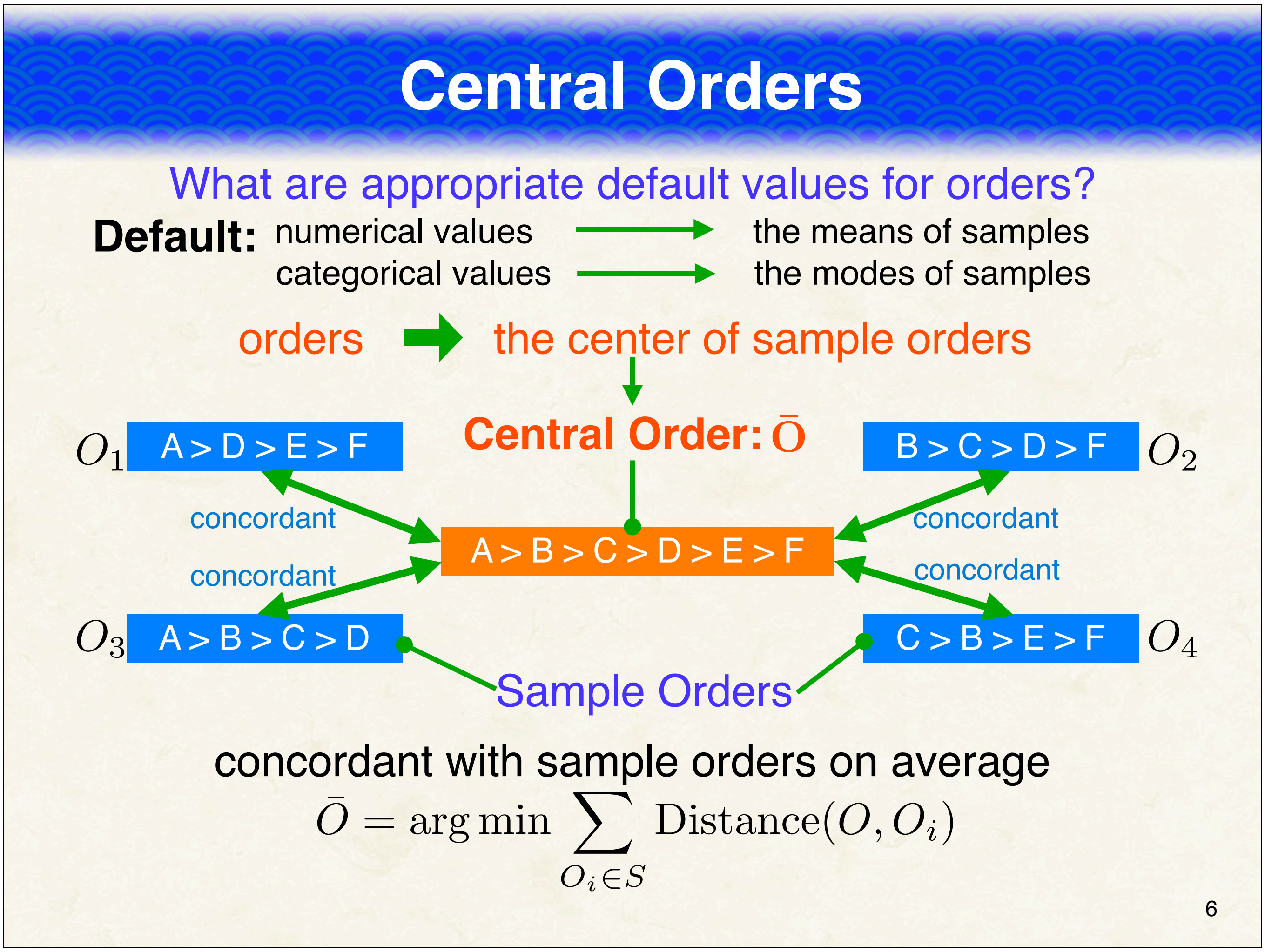
Missing objects are filled-in by some default values and distances are calculated over filled orders

If we could use information contained in missing objects, the precision of distances would become high.

To use the ignored information, missing objects are filled-in by some default values and distances are calculated over filled orders.

This is the reason why a filling-in technique is required.

Next, we show our filling-in technique of missing objects in orders.



# Central Orders

What are appropriate default values for orders?

**Default:** numerical values → the means of samples
categorical values → the modes of samples

orders → the center of sample orders

Central Order: $\bar{O}$

$O_1$ A > D > E > F

$O_2$ B > C > D > F

A > B > C > D > E > F

$O_3$ A > B > C > D

$O_4$ C > B > E > F

concordant

concordant

concordant

concordant

Sample Orders

concordant with sample orders on average

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

Before showing our filling-in technique, we discuss what are appropriate default values for orders.

In the case of numerical values, missing values are filled-in by the means of samples.

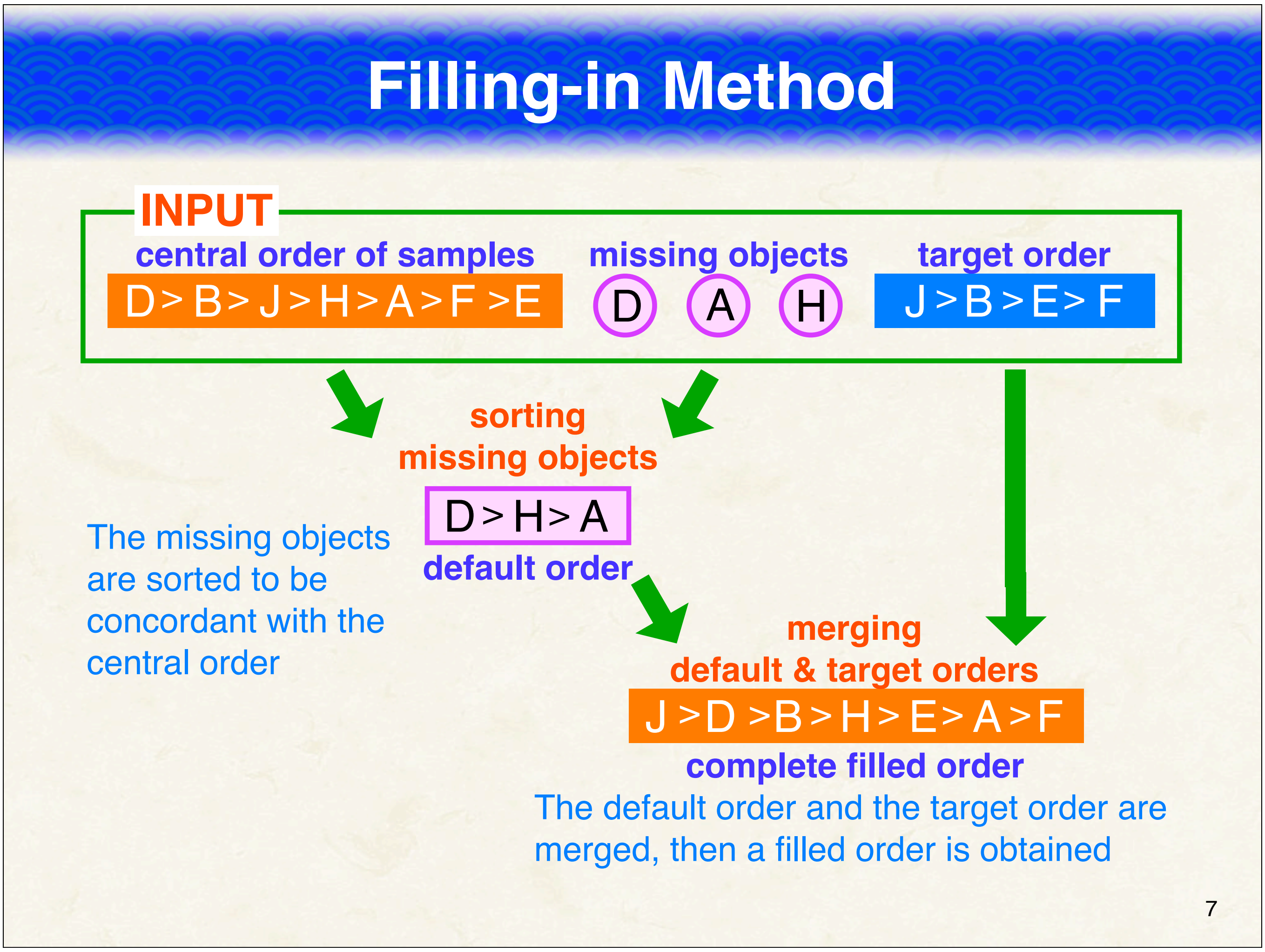
In the case of categorical values, missing values are filled-in by the modes of samples.

By analogy, it is appropriate to fill-in missing objects in orders by the center of sample orders.

This is an example of a central order.

Intuitively, a central order is concordant with sample orders on average.

This definition is analogous to the center of numerical vectors.



# Filling-in Method

INPUT

central order of samples

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

missing objects

D A H

target order

J > B > E > F

sorting missing objects

D > H > A

default order

The missing objects are sorted to be concordant with the central order

merging default & target orders

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

complete filled order

The default order and the target order are merged, then a filled order is obtained

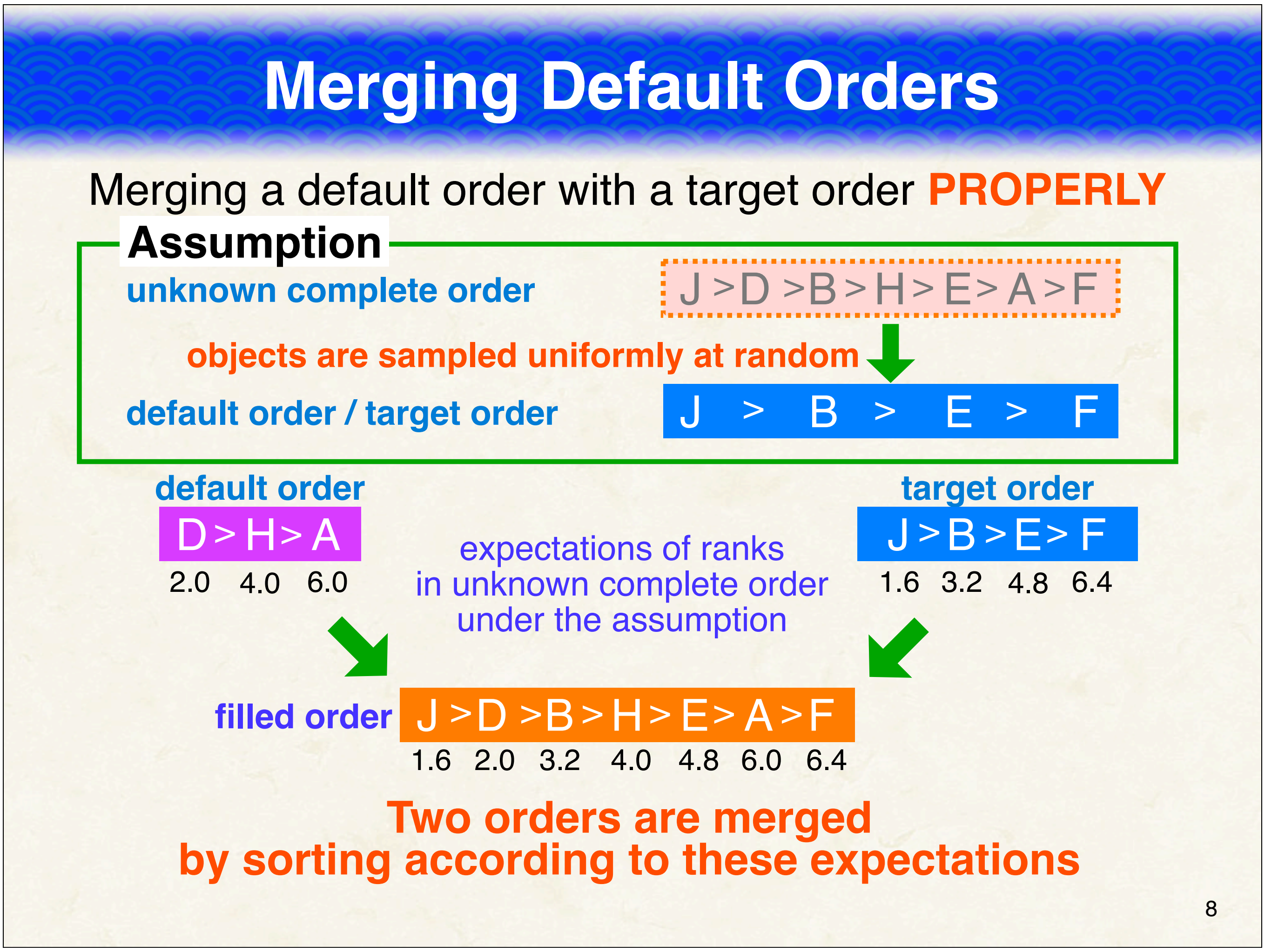
The method to fill-in missing objects by a central order is as follows.

The target order to fill-in, the missing objects of the target order, and the central order of samples are given.

First, the missing objects are sorted to be concordant with the central order.

We call the resultant order a default order.

Finally, the default order and the target order are merged, then the filled order is obtained.



# Merging Default Orders

Merging a default order with a target order **PROPERLY**

**Assumption**

**unknown complete order** J > D > B > H > E > A > F

**objects are sampled uniformly at random**

**default order / target order** J > B > E > F

**default order**

D > H > A

2.0 4.0 6.0

expectations of ranks
in unknown complete order
under the assumption

**target order**

J > B > E > F

1.6 3.2 4.8 6.4

**filled order** J > D > B > H > E > A > F

1.6 2.0 3.2 4.0 4.8 6.0 6.4

**Two orders are merged**
**by sorting according to these expectations**

All that we have to do is merging a default order with a target order properly.

First, we introduce an assumption, to clarify the condition that our merging method is proper.

A default order or a target order is generated by sampling objects uniformly at random from an unknown complete order.

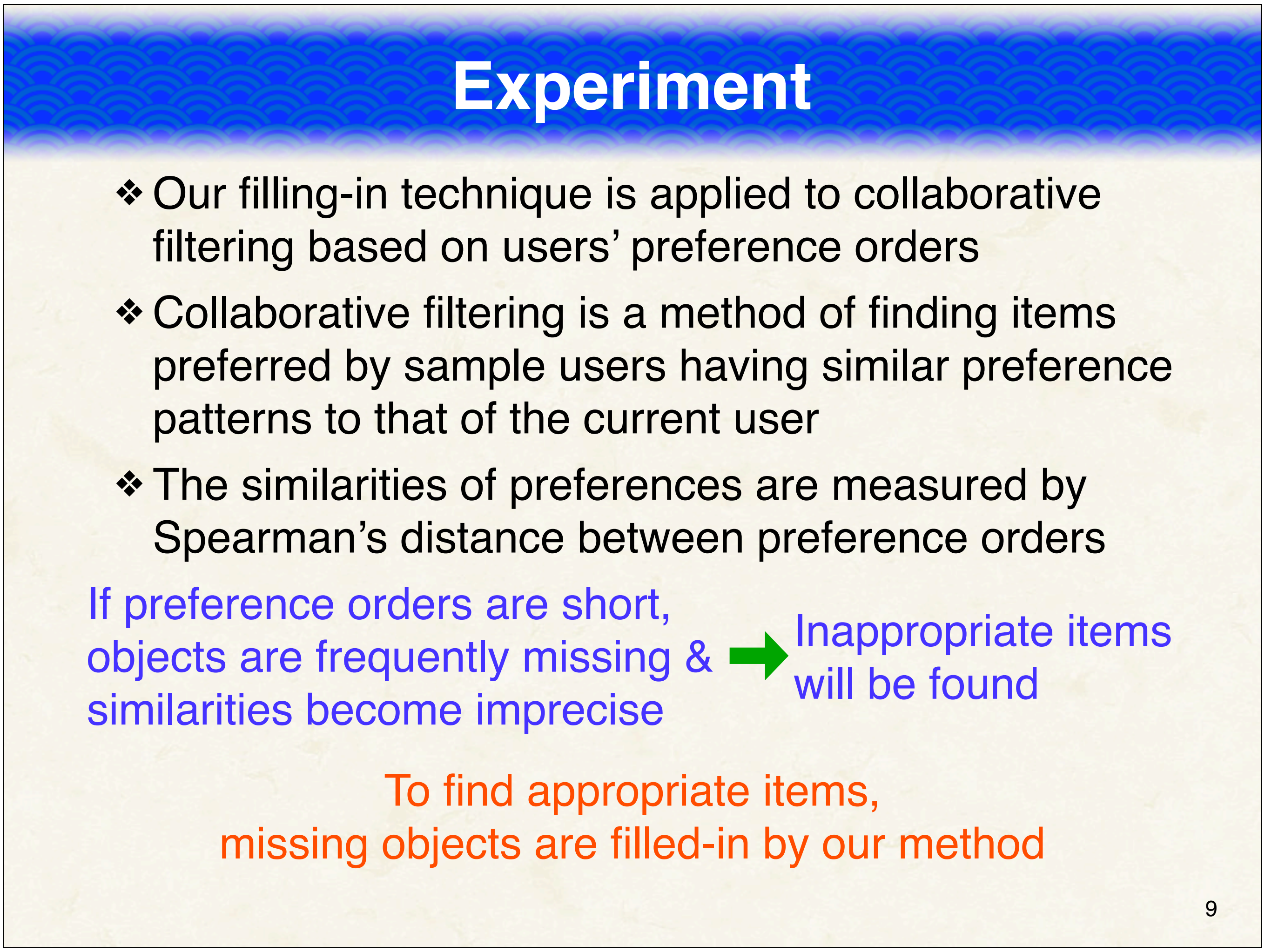
Under this assumption, for each object in a default order and a target order, an expectation of ranks in unknown complete order can be calculated.

This can be done based on a theory of an order statistics.

These two orders are merged by sorting according to these expectations.

Now, we can fill-in missing objects in orders.

Next, we experimentally show the effectiveness of this method.



# Experiment

- Our filling-in technique is applied to collaborative filtering based on users' preference orders
- Collaborative filtering is a method of finding items preferred by sample users having similar preference patterns to that of the current user
- The similarities of preferences are measured by Spearman's distance between preference orders

If preference orders are short, objects are frequently missing & similarities become imprecise → Inappropriate items will be found

To find appropriate items, missing objects are filled-in by our method

Our filling-in technique is applied to collaborative filtering based on users' preference orders.

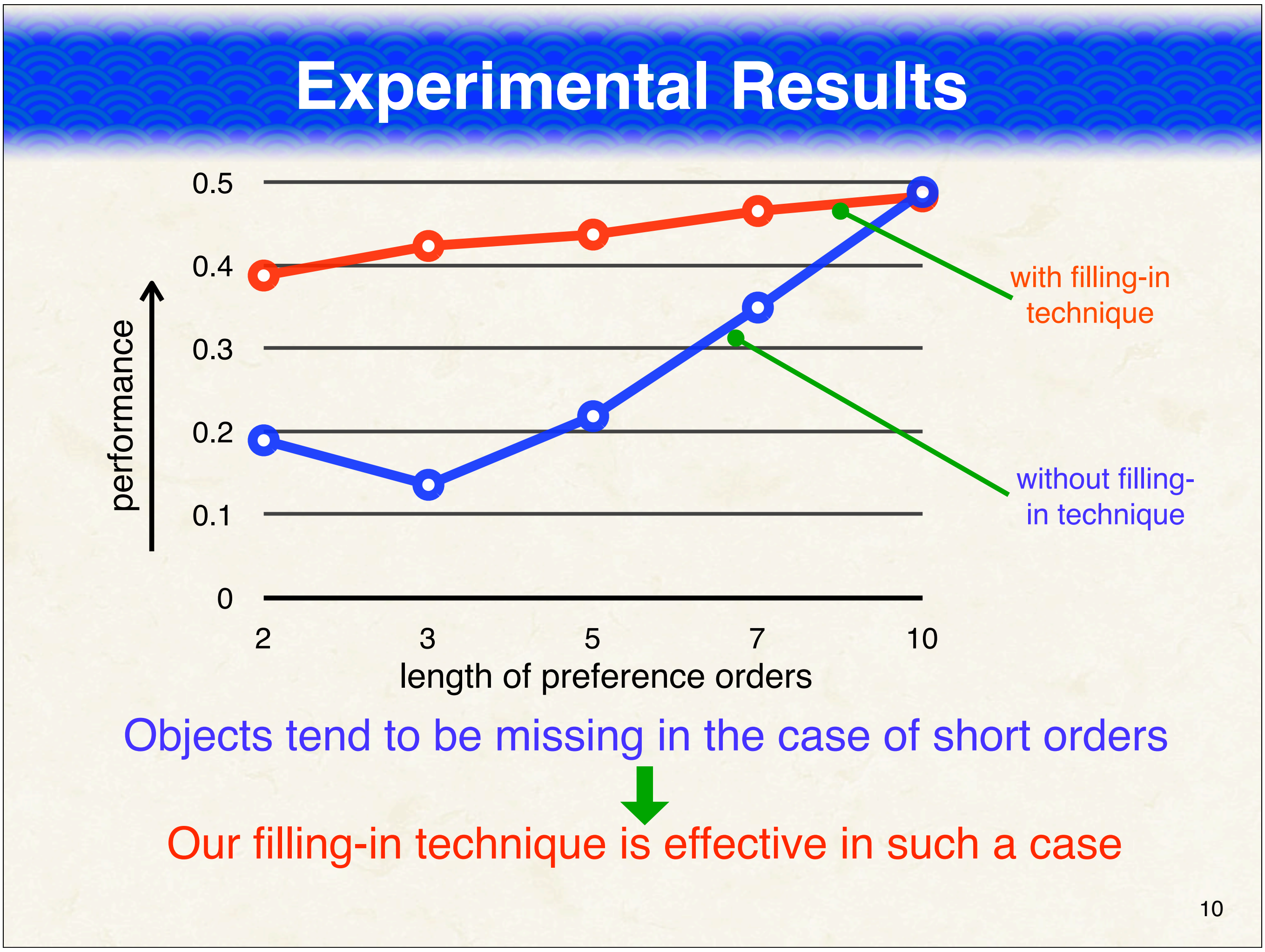
Collaborative filtering is a method to find items preferred by sample users having similar preference patterns to that of the current user.

In this experiment, the similarities of preference are measured by Spearman's distance between preference orders.

If preference orders are short, objects are frequently missing and similarities become imprecise.

As a result, inappropriate items will be found.

To find appropriate items, missing objects are filled-in by our method.



# Experimental Results

Objects tend to be missing in the case of short orders

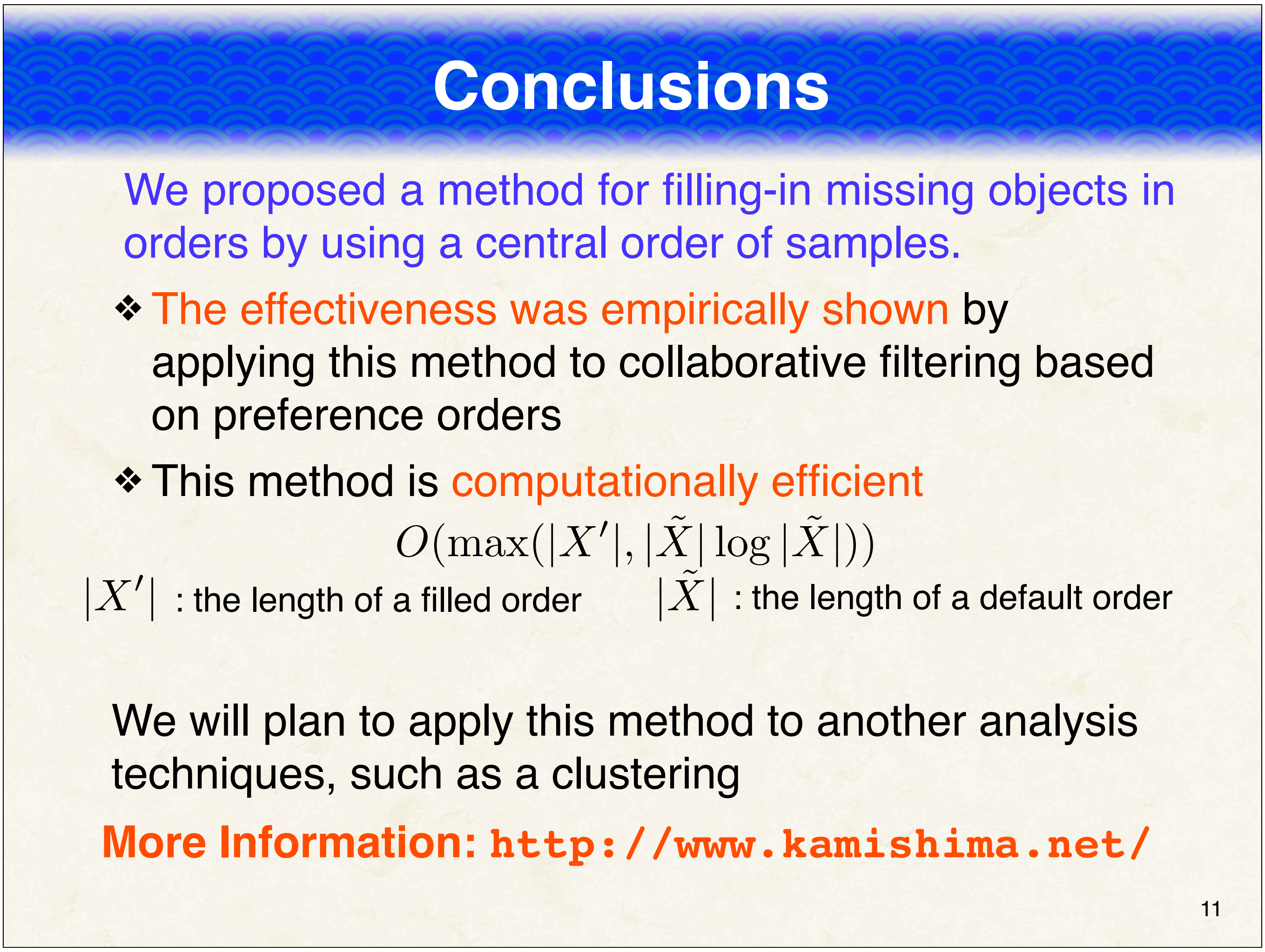
Our filling-in technique is effective in such a case

The blue line shows the performance of original collaborating filtering method.

The performance was drastically dropped in accordance with the decrease of the length of preference orders.

The red line shows the performance of the method adopting our filling-in technique.

The performance was improved especially if the preference orders were short.



# Conclusions

We proposed a method for filling-in missing objects in orders by using a central order of samples.

- ❖ The effectiveness was empirically shown by applying this method to collaborative filtering based on preference orders
- ❖ This method is computationally efficient

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

$|X'|$ : the length of a filled order    $|\tilde{X}|$ : the length of a default order

We will plan to apply this method to another analysis techniques, such as a clustering

**More Information: `http://www.kamishima.net/`**

We would like to conclude our talk.

We proposed a method for filling-in missing objects in orders by using a central order of samples, and its effectiveness was empirically shown.

Additionally, this method is computationally efficient.

That’s all we have to say. Thank you.