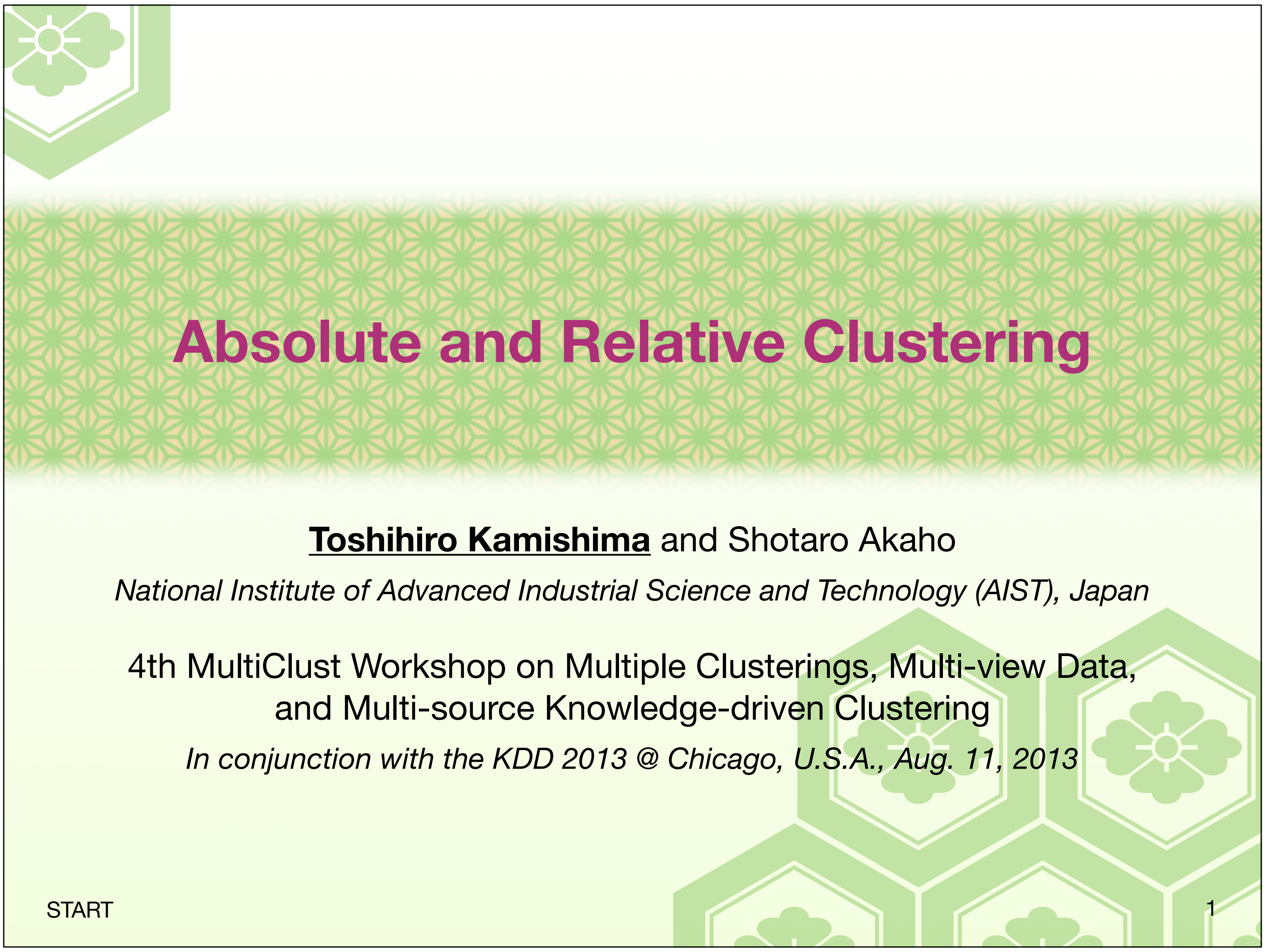


Absolute and Relative Clustering

Toshihiro Kamishima and Shotaro Akaho

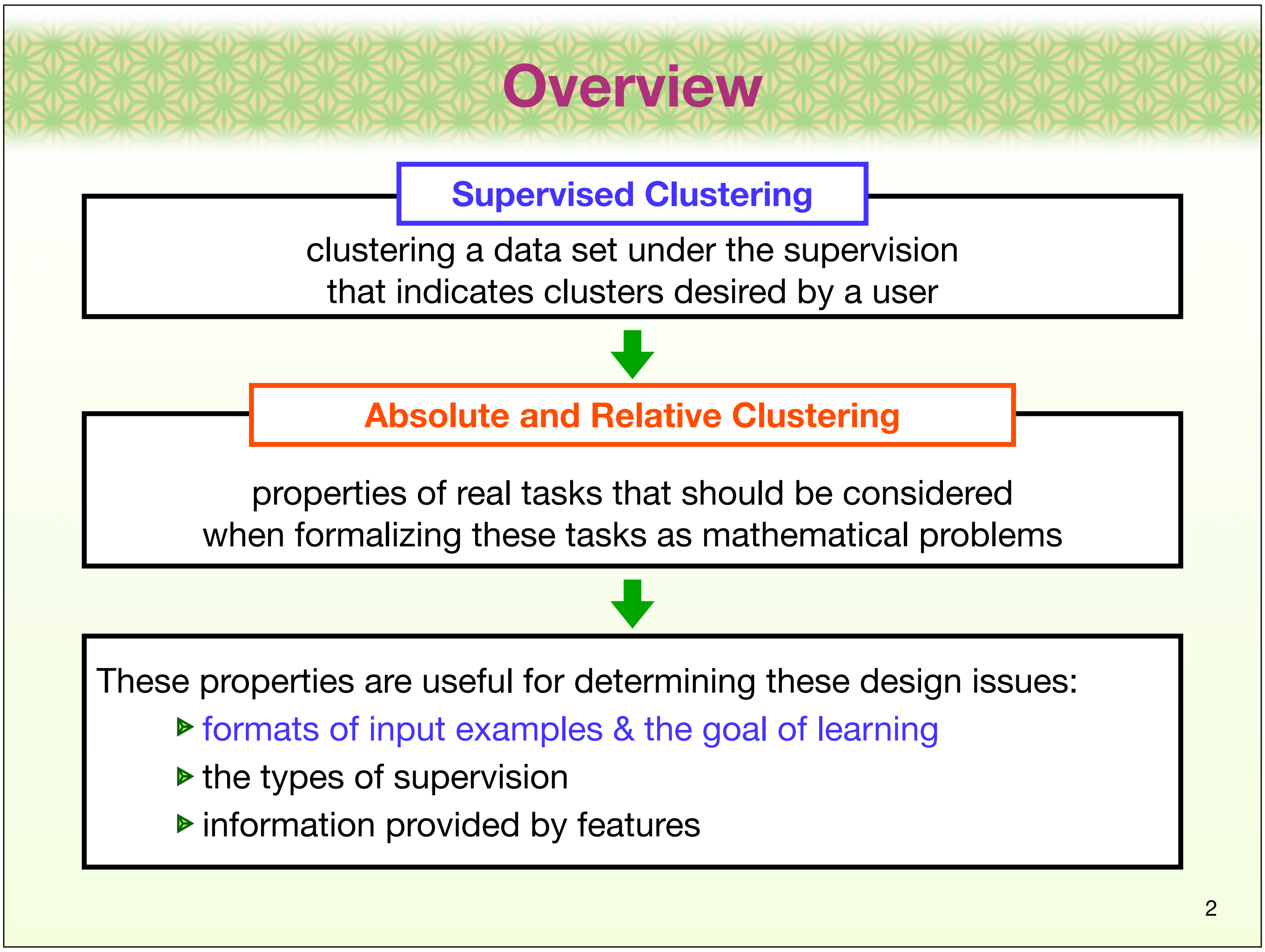
National Institute of Advanced Industrial Science and Technology (AIST), Japan

4th MultiClust Workshop on Multiple Clusterings, Multi-view Data, and Multi-source Knowledge-driven Clustering

In conjunction with the KDD 2013 @ Chicago, U.S.A., Aug. 11, 2013

START

I’m Toshihiro Kamishima.
Today, we will not propose a new method.
Our talk is about the property of clustering tasks, absolute clustering and relative clustering.



Overview

Supervised Clustering

clustering a data set under the supervision
that indicates clusters desired by a user

⬇

Absolute and Relative Clustering

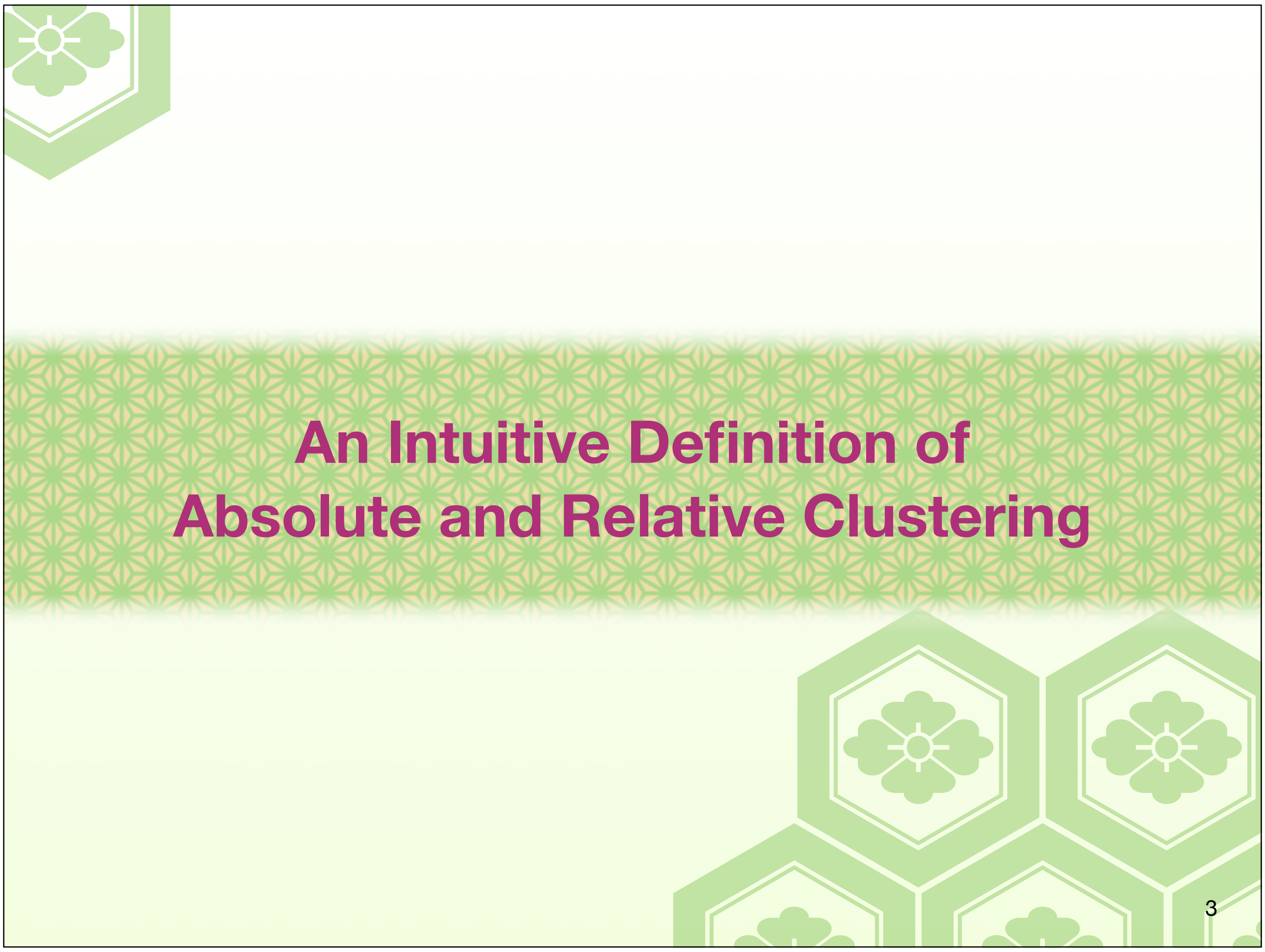
properties of real tasks that should be considered
when formalizing these tasks as mathematical problems

⬇

These properties are useful for determining these design issues:

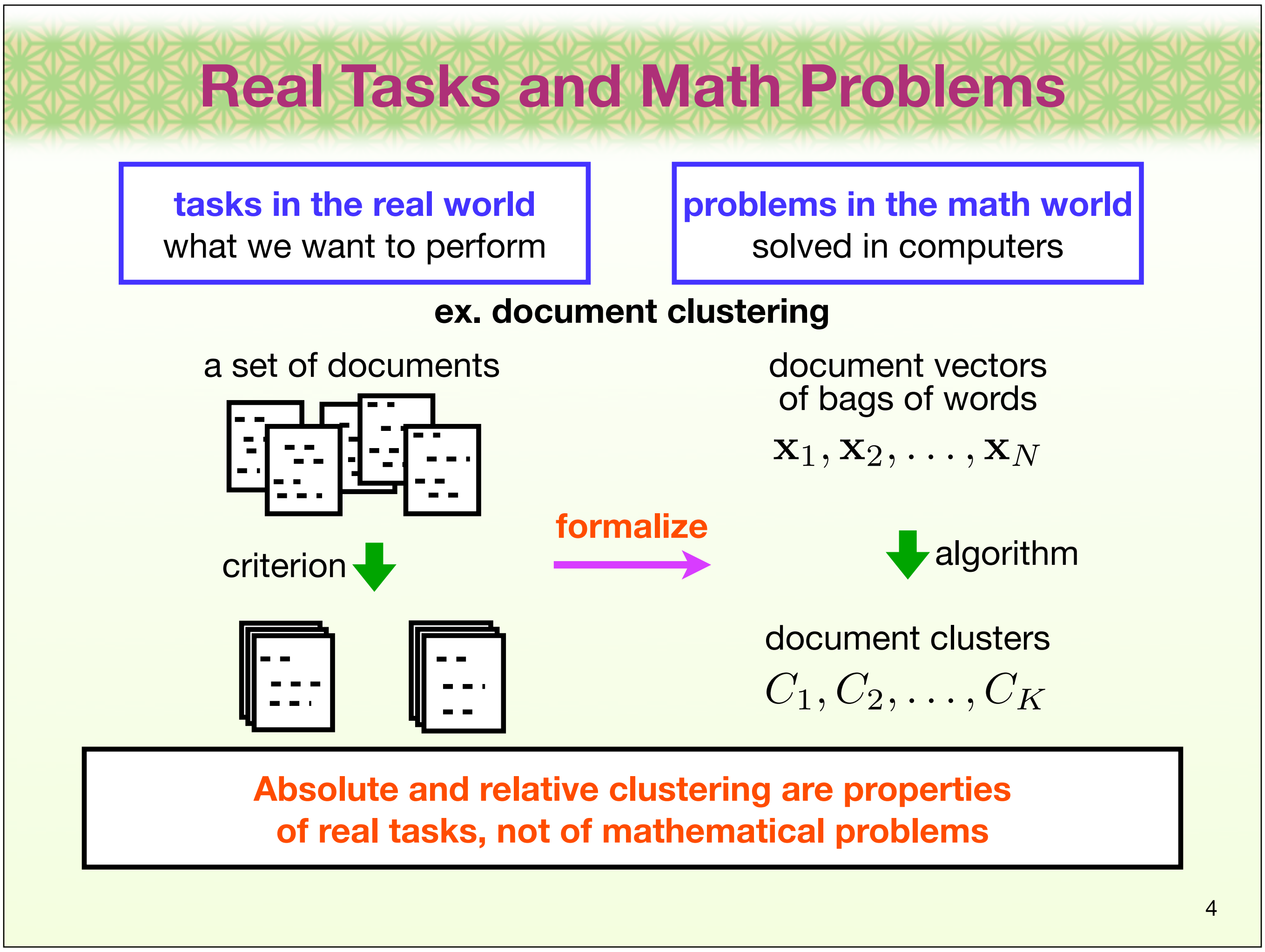
- formats of input examples & the goal of learning
- the types of supervision
- information provided by features

The goal of supervised clustering is to cluster a data set under the supervision that indicates clusters desired by a user.
Absolute and relative clustering are the properties of real tasks that should be considered when formalizing these tasks as mathematical problems.
These properties are useful for determining these design issues.
Today, we talk about this point, formats of input examples & the goal of learning.



An Intuitive Definition of Absolute and Relative Clustering

We begin by an intuitive definition of absolute and relative clustering.



Real Tasks and Math Problems

tasks in the real world
what we want to perform

problems in the math world
solved in computers

ex. document clustering

a set of documents

criterion

formalize

document vectors
of bags of words

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$

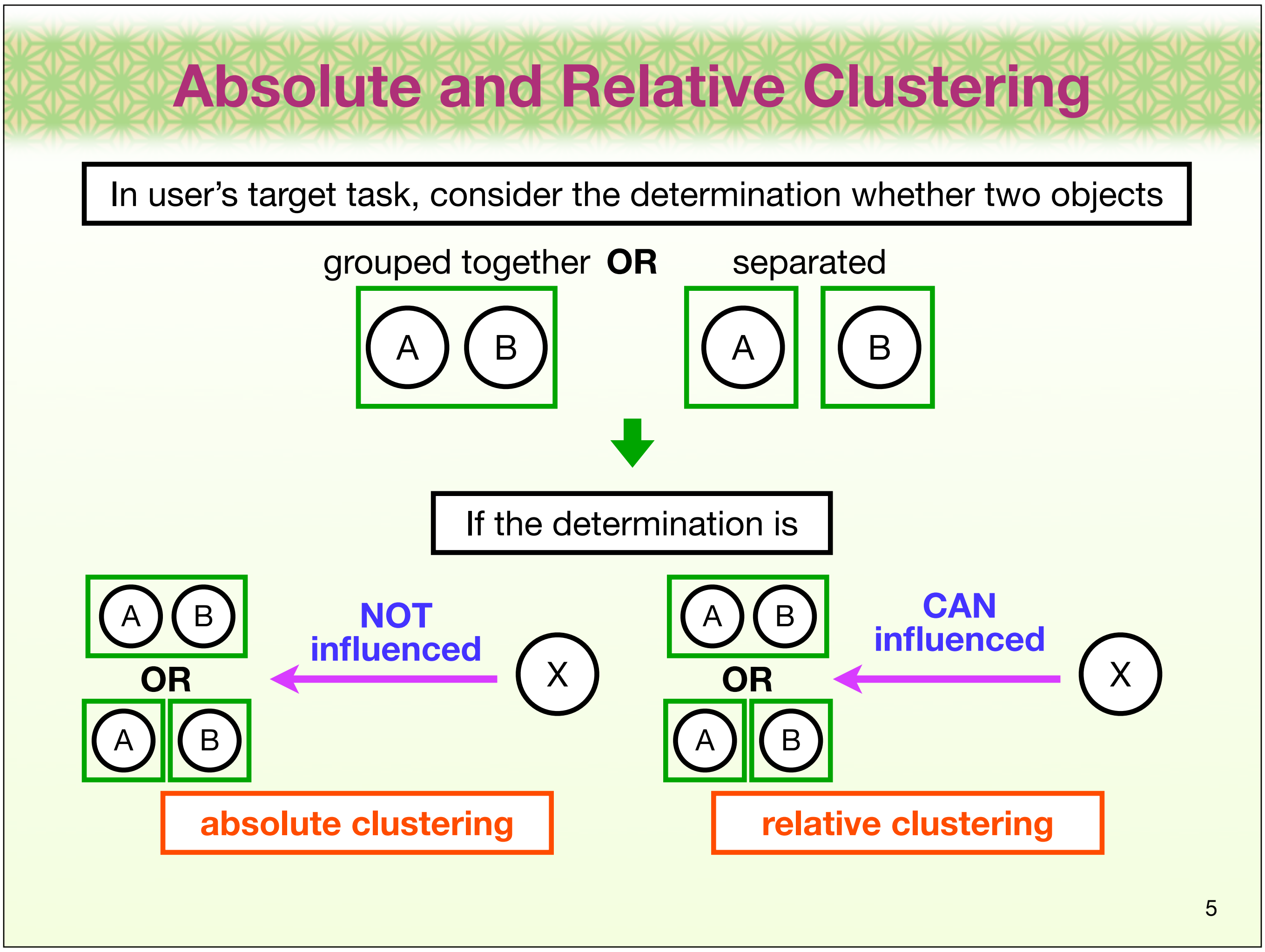
algorithm

document clusters

$C_1, C_2, \ldots, C_K$

Absolute and relative clustering are properties of real tasks, not of mathematical problems

We here differentiate tasks and problems.
Tasks in the real world are what we want to perform, and problems in the mathematical world are solved in computers.
This is an example of document clustering.
A real task is grouping a set of documents based on some appropriate criterion.
This task must be formalized to solve it in computers.
A mathematical problem is to generate clusters from a set of document vectors by applying an algorithm.
Absolute and relative clustering are properties of real tasks, not of mathematical problems.



Absolute and Relative Clustering

In user’s target task, consider the determination whether two objects

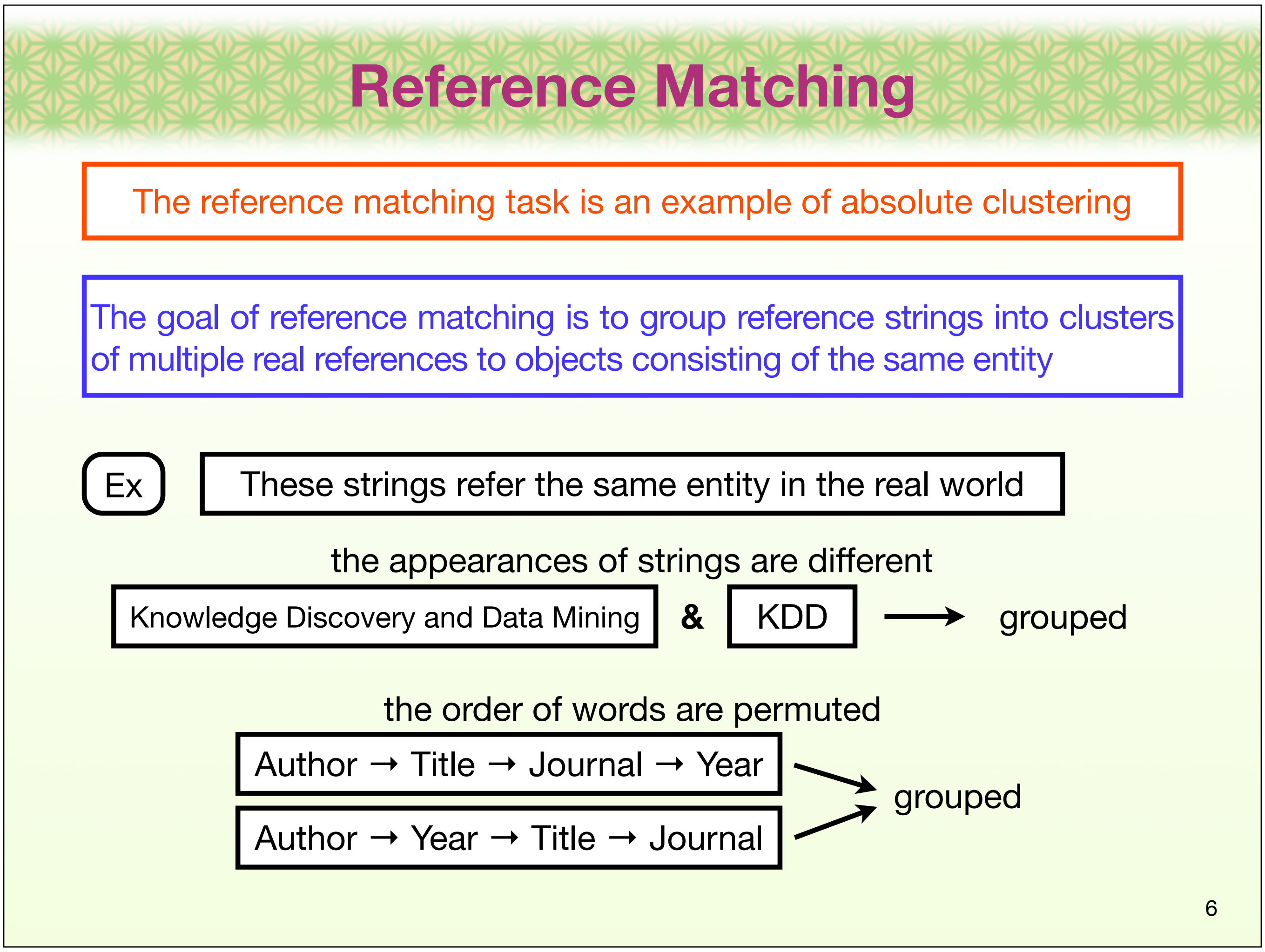
grouped together **OR** separated

If the determination is

absolute clustering

relative clustering

In user’s target task, we consider the determination whether two objects are grouped together or separated.
If the determination is not influenced by any other objects, the task is absolute clustering.
If it can be influenced by some object, the task is relative clustering.
We then show examples.



Reference Matching

The reference matching task is an example of absolute clustering

The goal of reference matching is to group reference strings into clusters of multiple real references to objects consisting of the same entity

Ex

These strings refer the same entity in the real world

the appearances of strings are different

Knowledge Discovery and Data Mining & KDD → grouped

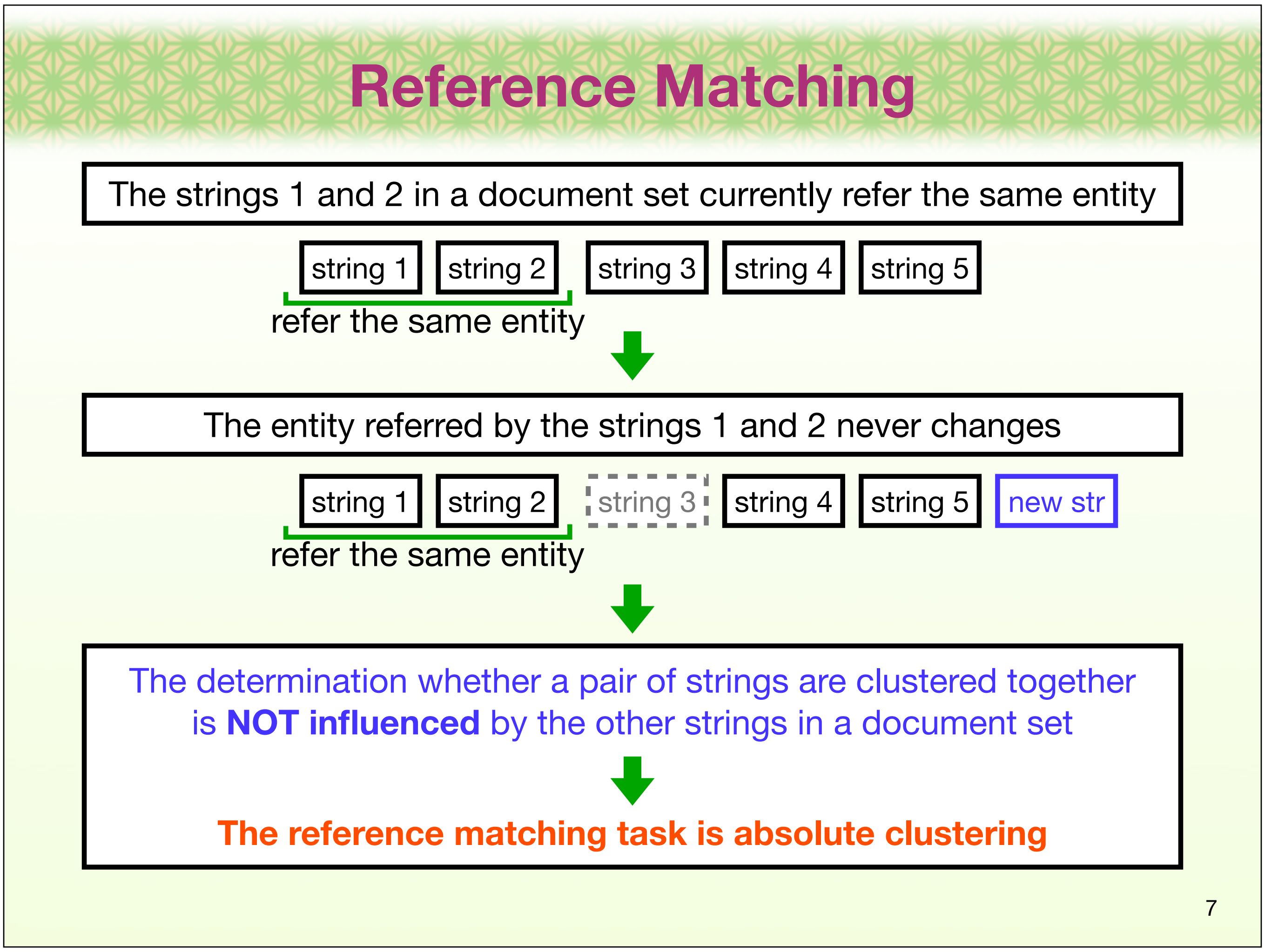
the order of words are permuted

Author → Title → Journal → Year

Author → Year → Title → Journal

→ grouped

The reference matching task is an example of absolute clustering.
The goal of reference matching is to group reference strings into clusters of multiple real references to objects consisting of the same entity.
For example, because these strings refer the same entity in the real world, they are grouped together, even if the appearance of strings are different, or even if the order of words are permuted.



Reference Matching

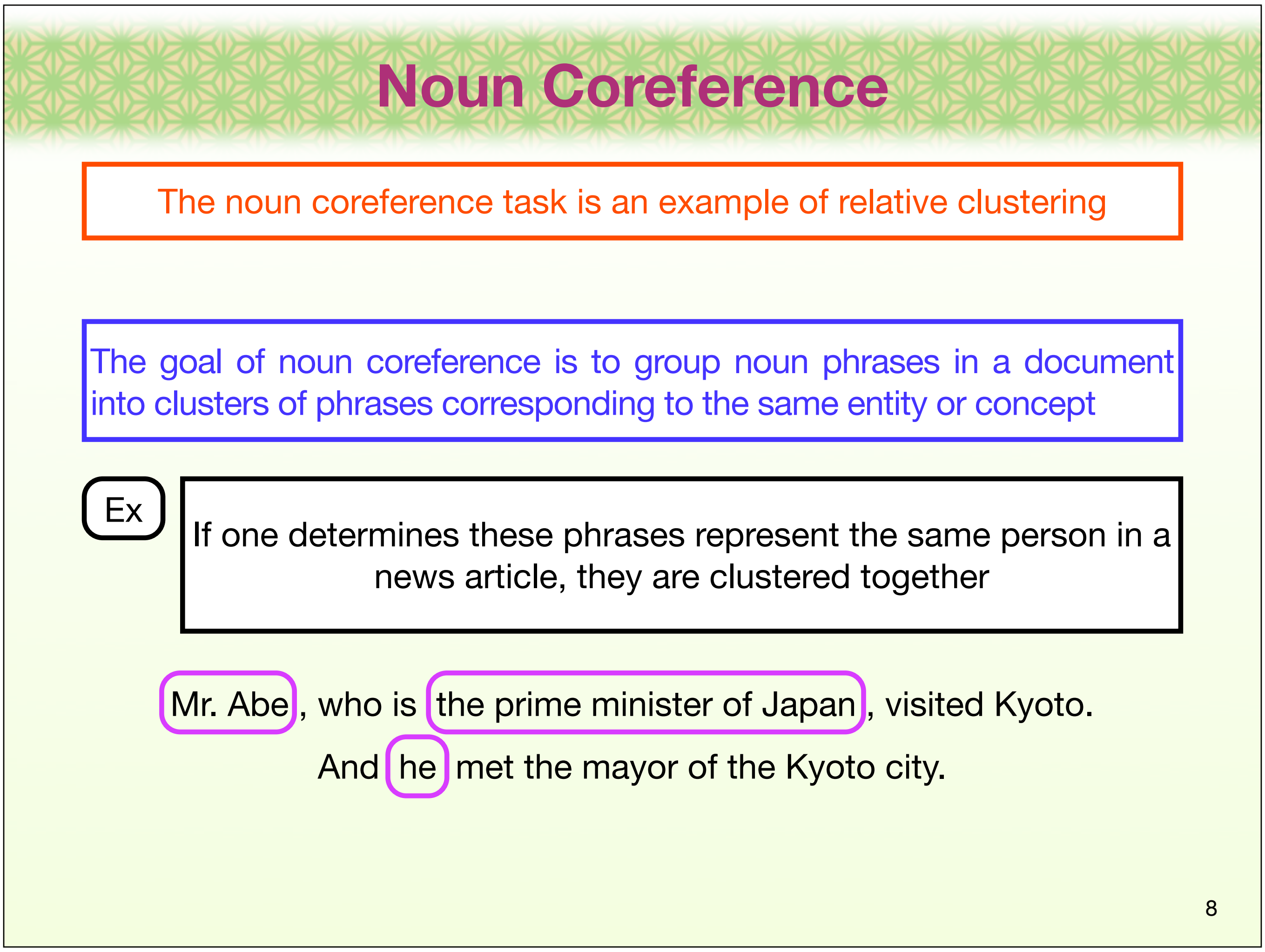
The strings 1 and 2 in a document set currently refer the same entity

The entity referred by the strings 1 and 2 never changes

The determination whether a pair of strings are clustered together is **NOT influenced** by the other strings in a document set

The reference matching task is absolute clustering

The strings 1 and 2 in a document set currently refer the same entity.
If the string 3 is eliminated from the document set, or if a new string is added to the document set, the entity referred by the strings 1 and 2 never changes.
The determination whether a pair of strings are clustered together or not is NOT influenced by the other strings in a document set
Consequently, the reference matching task is absolute clustering.



Noun Coreference

The noun coreference task is an example of relative clustering

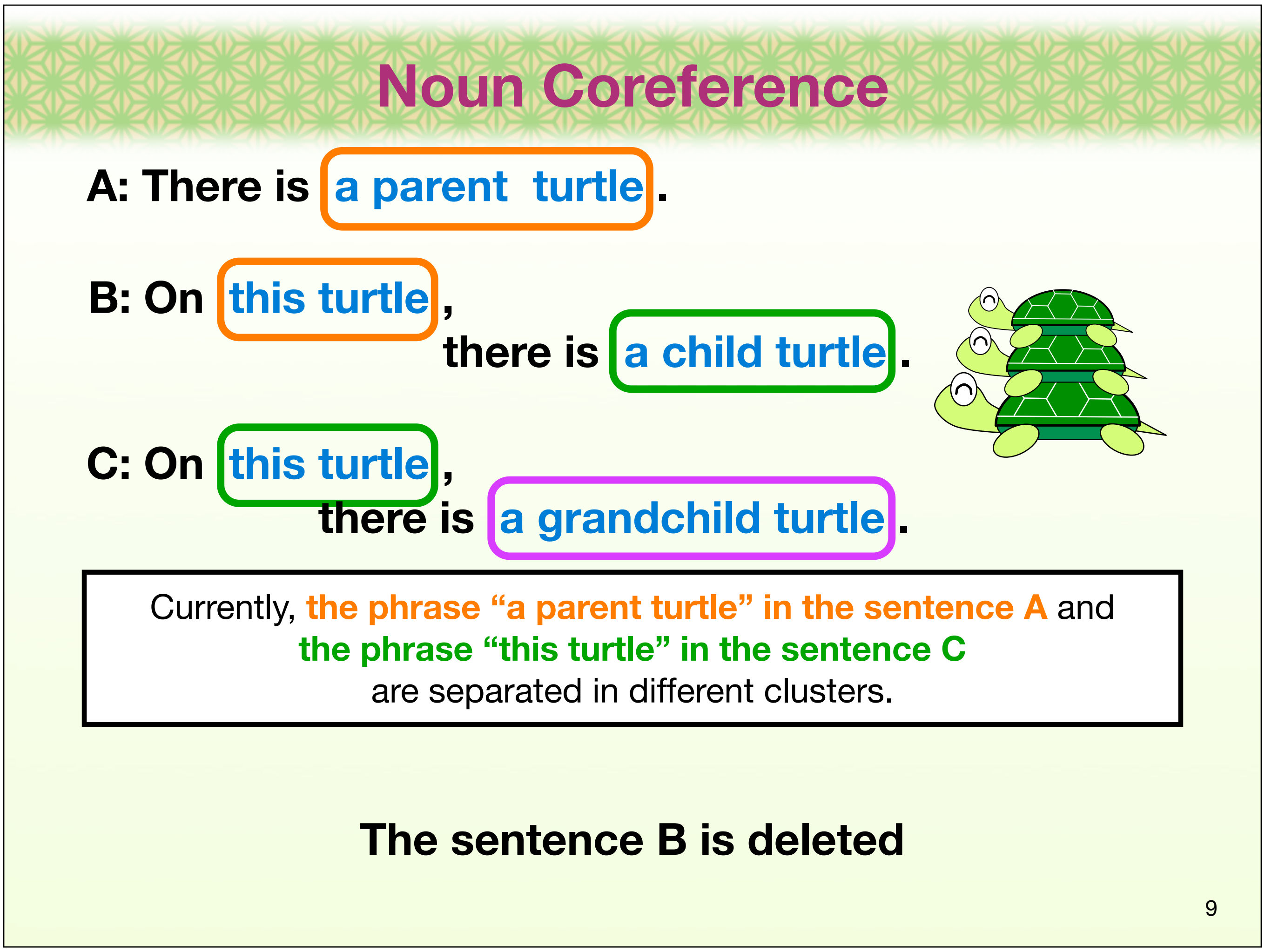
The goal of noun coreference is to group noun phrases in a document into clusters of phrases corresponding to the same entity or concept

Ex

If one determines these phrases represent the same person in a news article, they are clustered together

Mr. Abe, who is the prime minister of Japan, visited Kyoto.
And he met the mayor of the Kyoto city.

The noun coreference task is an example of relative clustering. The goal of this task is to group noun phrases in a document into clusters of phrases corresponding to the same entity or concept. For example, if one determines the these phrases, “Mr Abe,” “the prime minister of Japan,” and “he”, represent the same person in a news article, they are clustered together.



Noun Coreference

A: There is a parent turtle.

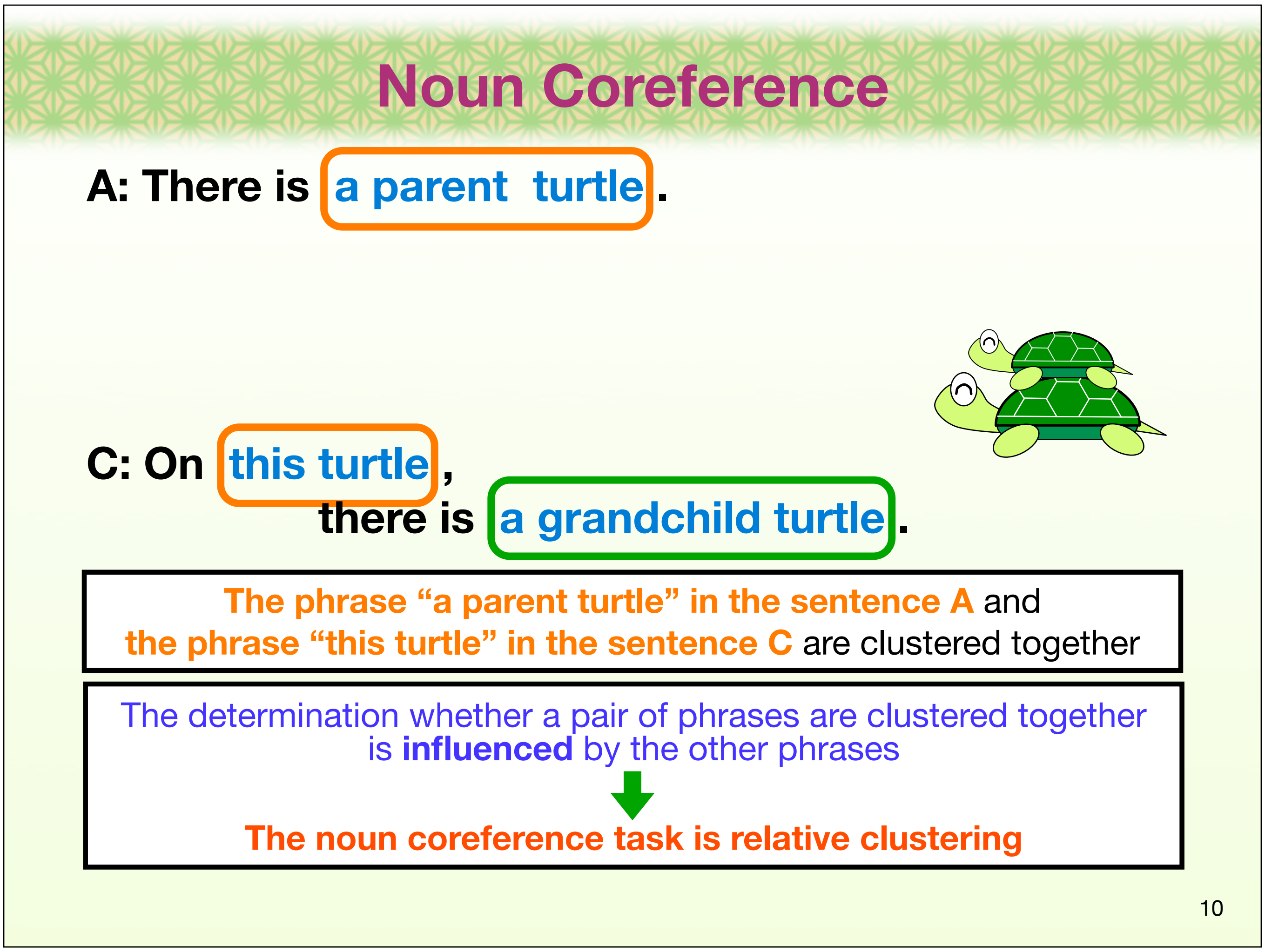
B: On this turtle, there is a child turtle.

C: On this turtle, there is a grandchild turtle.

Currently, the phrase “a parent turtle” in the sentence A and the phrase “this turtle” in the sentence C are separated in different clusters.

The sentence B is deleted

Consider the case of these three sentences.
Five nouns are grouped into three clusters.
Currently, the phrase “a parent turtle” in the sentence A and the phrase “this turtle” in the sentence C are separated in different clusters.
The sentence B is deleted.



Noun Coreference

A: There is **a parent turtle**.

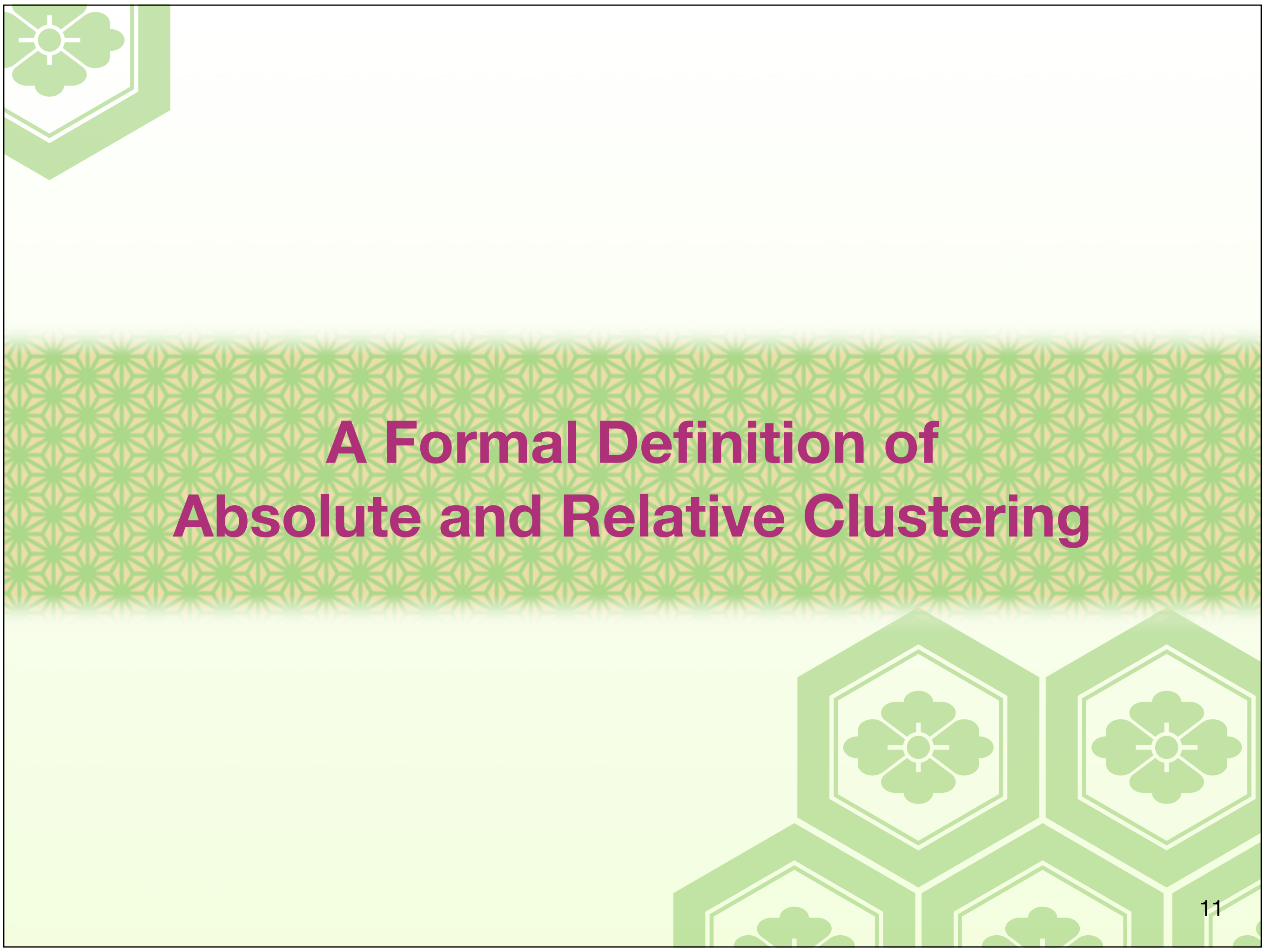
C: On **this turtle**,
there is **a grandchild turtle**.

The phrase “a parent turtle” in the sentence A and
the phrase “this turtle” in the sentence C are clustered together

The determination whether a pair of phrases are clustered together
is **influenced** by the other phrases

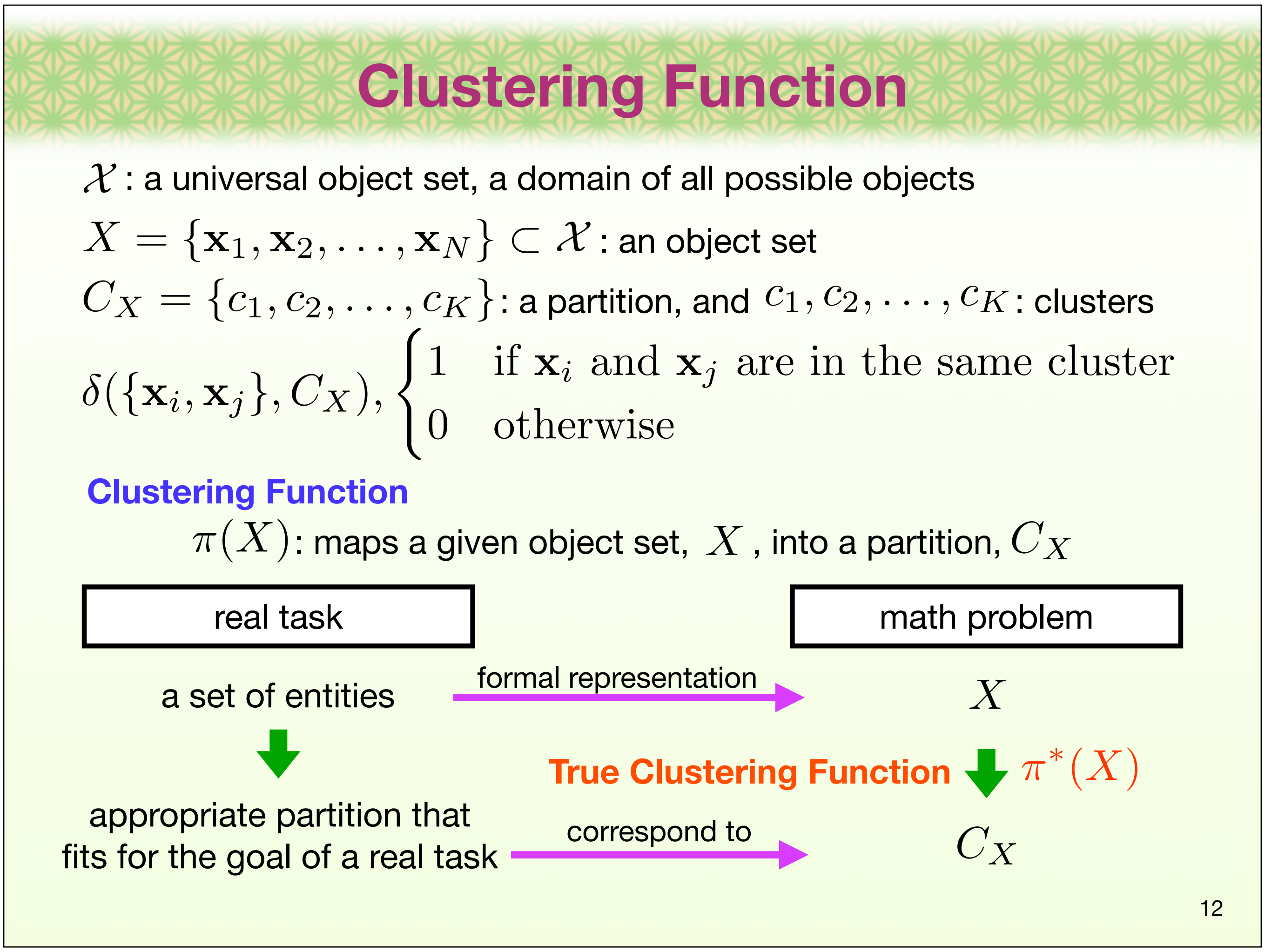
The noun coreference task is relative clustering

Now, three nouns are grouped into two clusters.
The phrase “a parent turtle” in the sentence A and the phrase “this turtle” in the sentence C are clustered together.
The determination whether a pair of phrases are clustered together is influenced by the other phrases.
Consequently, the noun coreference task is relative clustering.



A Formal Definition of Absolute and Relative Clustering

We next give a formal definition.



Clustering Function

$\mathcal{X}$: a universal object set, a domain of all possible objects

$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\} \subset \mathcal{X}$: an object set

$C_X = \{c_1, c_2, \ldots, c_K\}$: a partition, and $c_1, c_2, \ldots, c_K$: clusters

$$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, C_X), \begin{cases} 1 & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

Clustering Function

$\pi(X)$: maps a given object set, X , into a partition, C_X

real task

math problem

a set of entities

formal representation

X

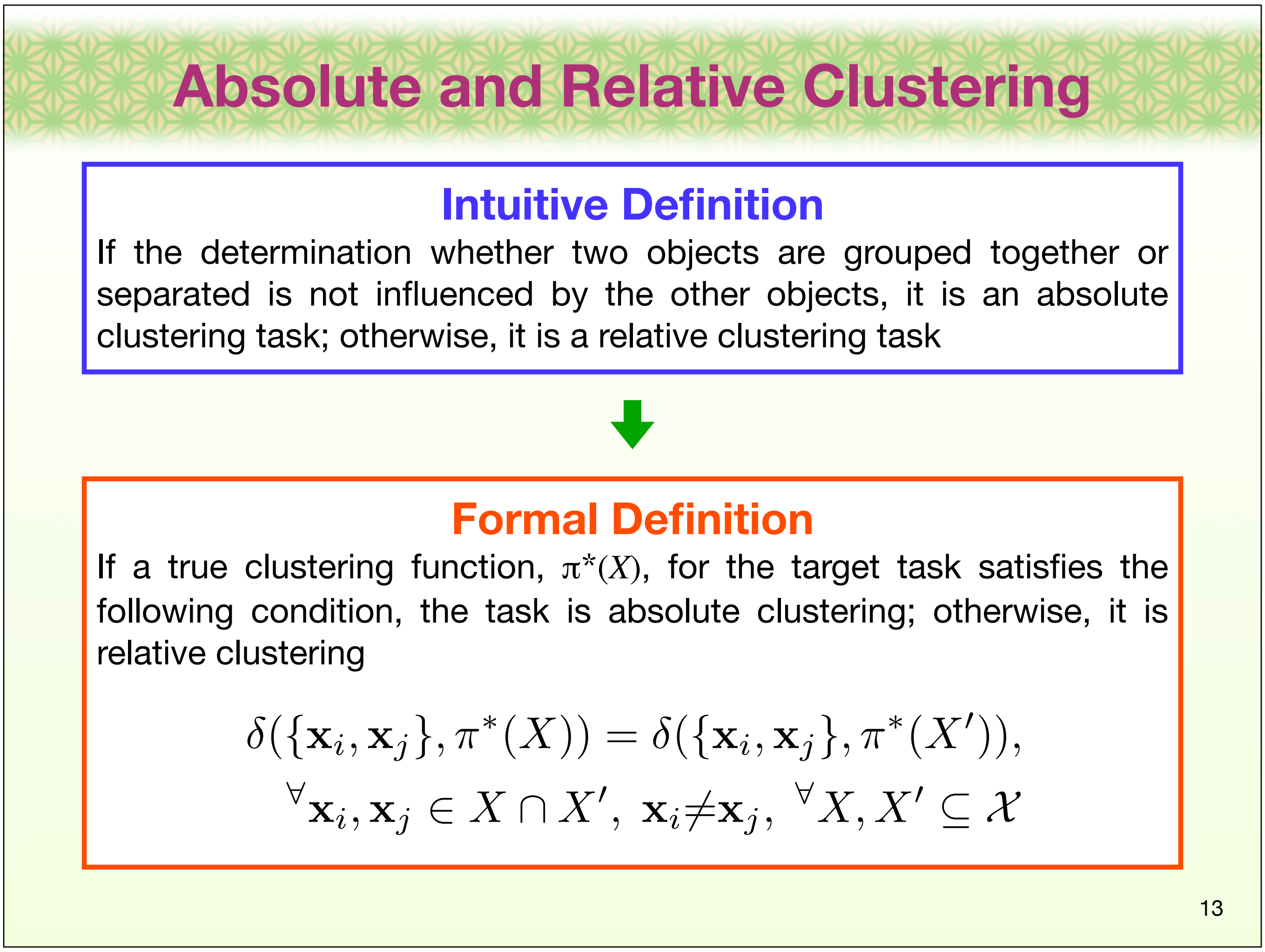
True Clustering Function $\pi^*(X)$

appropriate partition that fits for the goal of a real task

correspond to

C_X

Basic notations are as follows:
delta is an indicator function to represent whether two objects are in the same cluster or not.
A clustering function maps a given object set into a partition.
A true clustering function used for deriving an appropriate partition that fits for the goal of a real task.



Absolute and Relative Clustering

Intuitive Definition

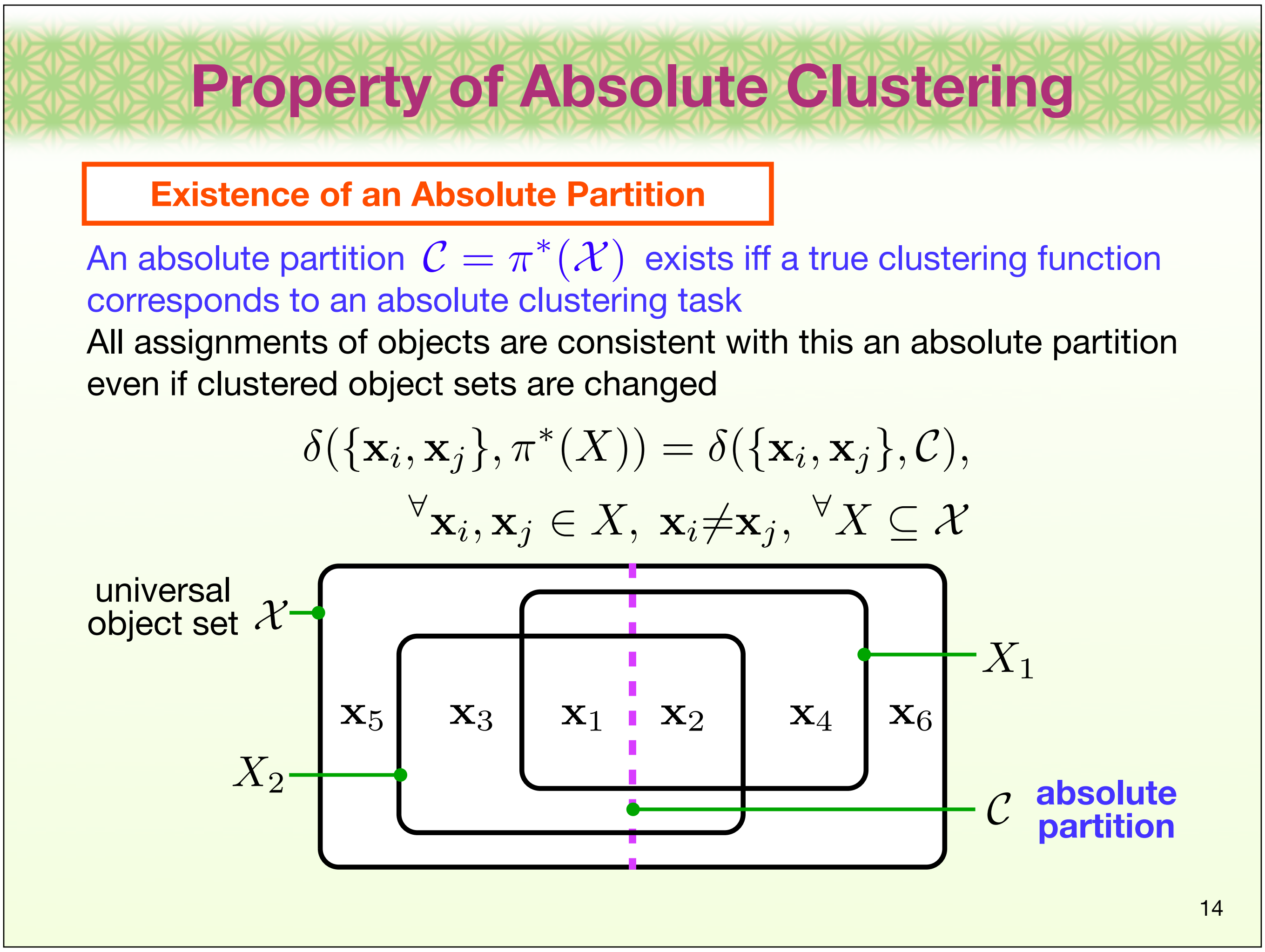
If the determination whether two objects are grouped together or separated is not influenced by the other objects, it is an absolute clustering task; otherwise, it is a relative clustering task

Formal Definition

If a true clustering function, $\pi^*(X)$, for the target task satisfies the following condition, the task is absolute clustering; otherwise, it is relative clustering

$$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi^*(X)) = \delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi^*(X')),$$
$$^\forall\mathbf{x}_i, \mathbf{x}_j \in X \cap X',\ \mathbf{x}_i{\neq}\mathbf{x}_j,\ ^\forall X, X' \subseteq \mathcal{X}$$

Intuitively speaking, if the determination whether two objects are grouped together or separated is influenced by the other objects, it is an absolute clustering task; otherwise, it is a relative clustering task.
This is formally defined like this.



Property of Absolute Clustering

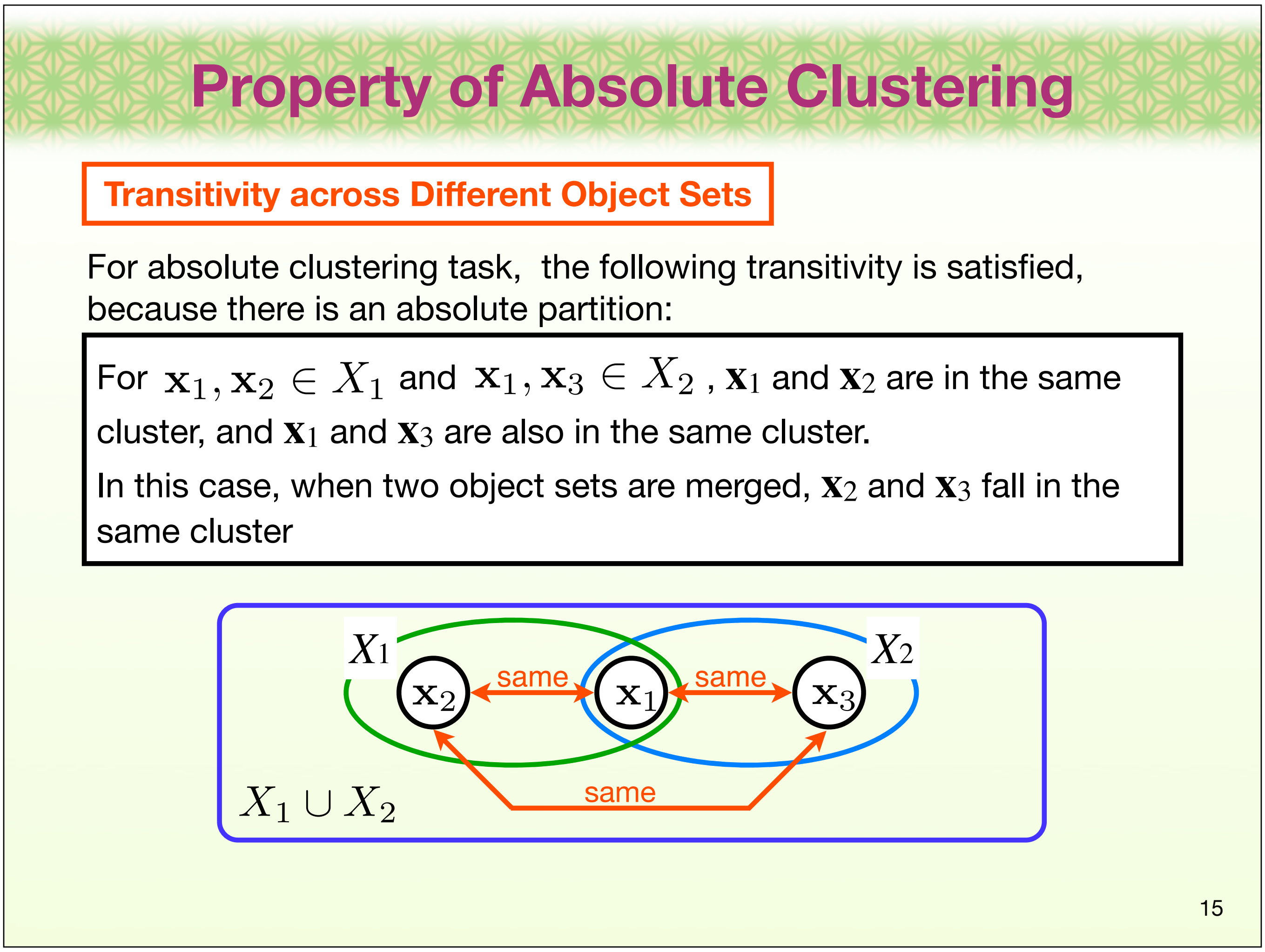
Existence of an Absolute Partition

An absolute partition $\mathcal{C} = \pi^*(\mathcal{X})$ exists iff a true clustering function corresponds to an absolute clustering task

All assignments of objects are consistent with this an absolute partition even if clustered object sets are changed

$$\delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \pi^*(X)) = \delta(\{\mathbf{x}_i, \mathbf{x}_j\}, \mathcal{C}),$$
$$^\forall\mathbf{x}_i, \mathbf{x}_j \in X,\ \mathbf{x}_i{\neq}\mathbf{x}_j,\ ^\forall X \subseteq \mathcal{X}$$

An absolute clustering task has two special properties.
The first property is the existence of absolute clustering.
An absolute partition exists if and only if a true clustering function corresponds to an absolute clustering task.
All assignments of objects are consistent with this an absolute partition even if clustered object sets are changed.
That’s why we call this property by absolute clustering.



Property of Absolute Clustering

Transitivity across Different Object Sets

For absolute clustering task, the following transitivity is satisfied, because there is an absolute partition:

For $\mathbf{x}_1, \mathbf{x}_2 \in X_1$ and $\mathbf{x}_1, \mathbf{x}_3 \in X_2$, $\mathbf{x}_1$ and $\mathbf{x}_2$ are in the same cluster, and $\mathbf{x}_1$ and $\mathbf{x}_3$ are also in the same cluster.

In this case, when two object sets are merged, $\mathbf{x}_2$ and $\mathbf{x}_3$ fall in the same cluster

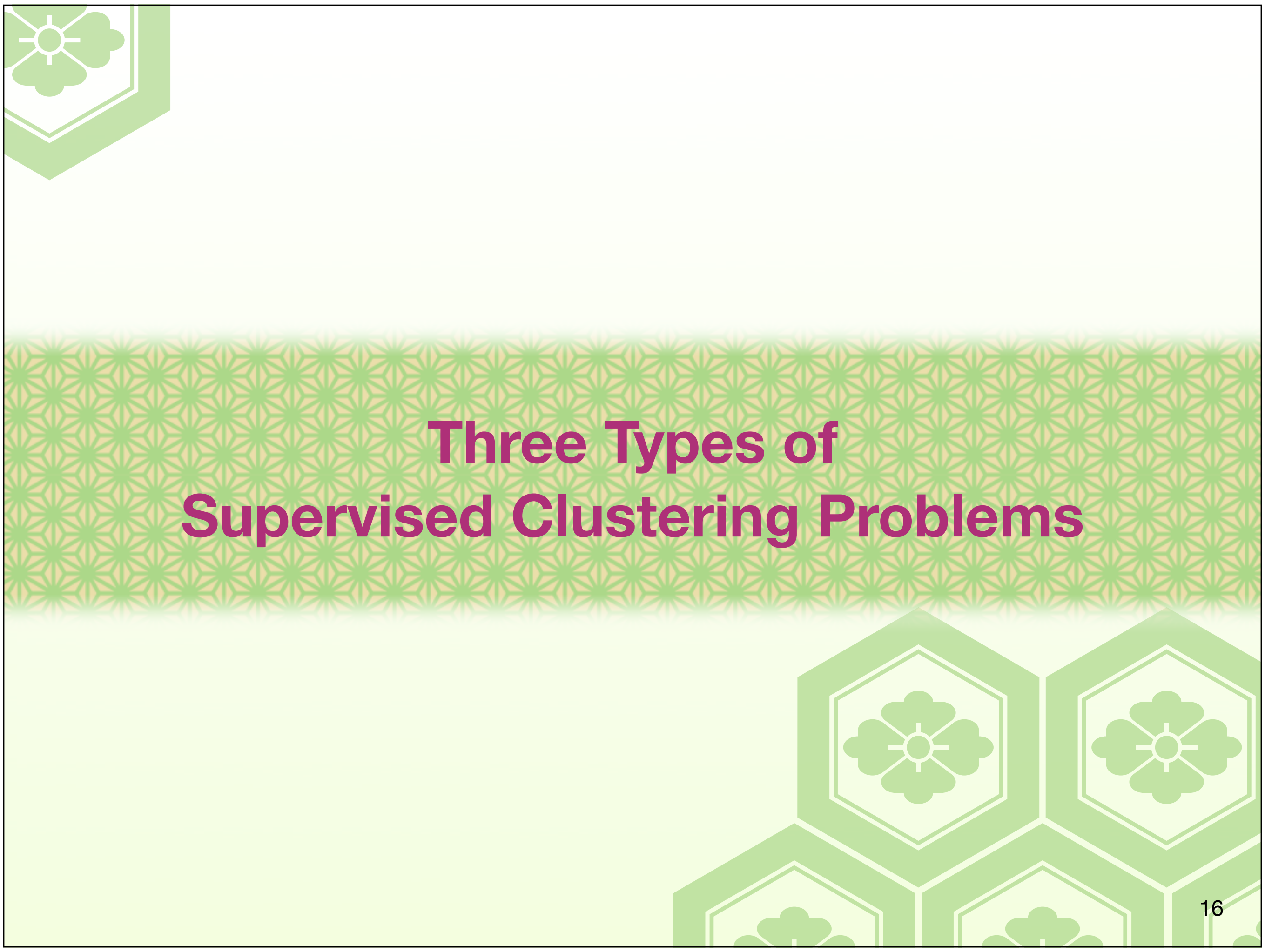
The second property is the transitivity across different object sets.

For absolute clustering task, the following transitivity is satisfied, because there is an absolute partition.

These objects, x_1 and x_2, of the object set, X_1 are in the same cluster.

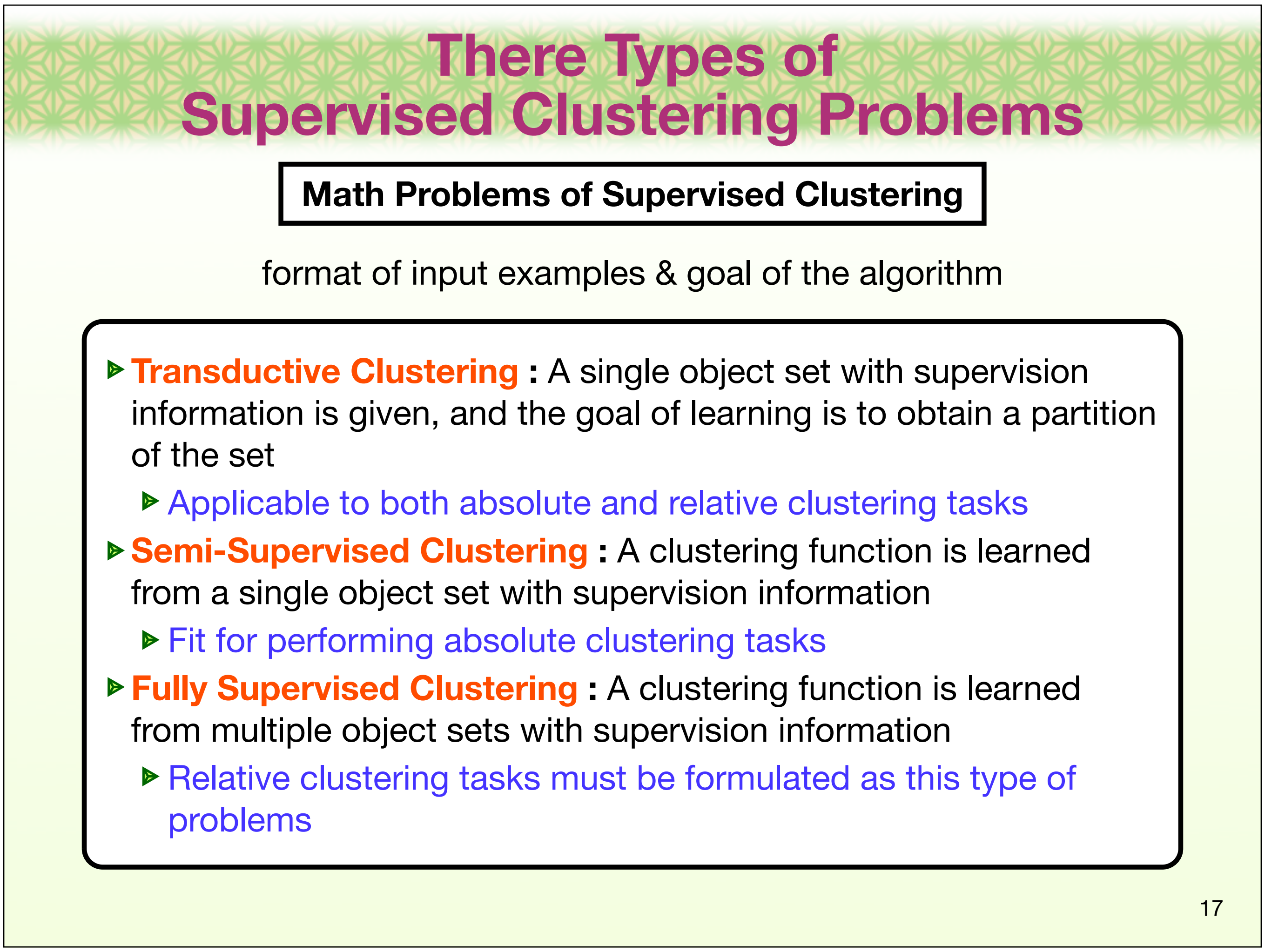
These objects, x_1 and x_3, of the object set, X_2 are in the same cluster.

In this case, when two object sets are merged, this object, x_2, and this object, x_3, fall in the same cluster.



Three Types of Supervised Clustering Problems

We next discuss three types of supervised clustering problems



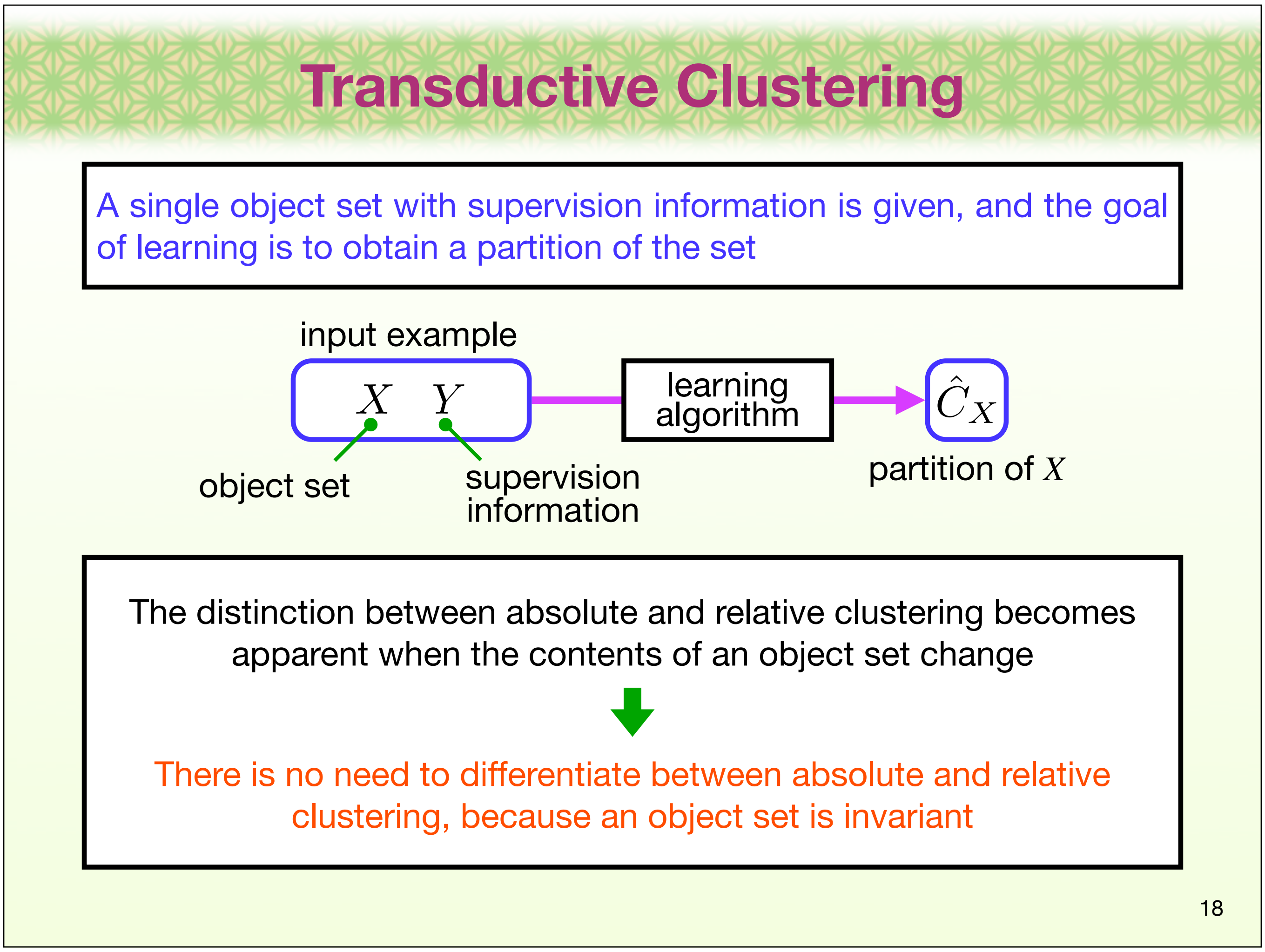
There Types of Supervised Clustering Problems

Math Problems of Supervised Clustering

format of input examples & goal of the algorithm

- **Transductive Clustering :** A single object set with supervision information is given, and the goal of learning is to obtain a partition of the set
    - Applicable to both absolute and relative clustering tasks
- **Semi-Supervised Clustering :** A clustering function is learned from a single object set with supervision information
    - Fit for performing absolute clustering tasks
- **Fully Supervised Clustering :** A clustering function is learned from multiple object sets with supervision information
    - Relative clustering tasks must be formulated as this type of problems

Mathematical problems of supervised clustering can be classified into three types based on the formats of input examples and the goal of the algorithm.
These are transductive clustering, semi-supervised clustering, and fully supervised clustering, and have relation with notions of absolute and relative clustering.
We sequentially show these problems.



Transductive Clustering

A single object set with supervision information is given, and the goal of learning is to obtain a partition of the set

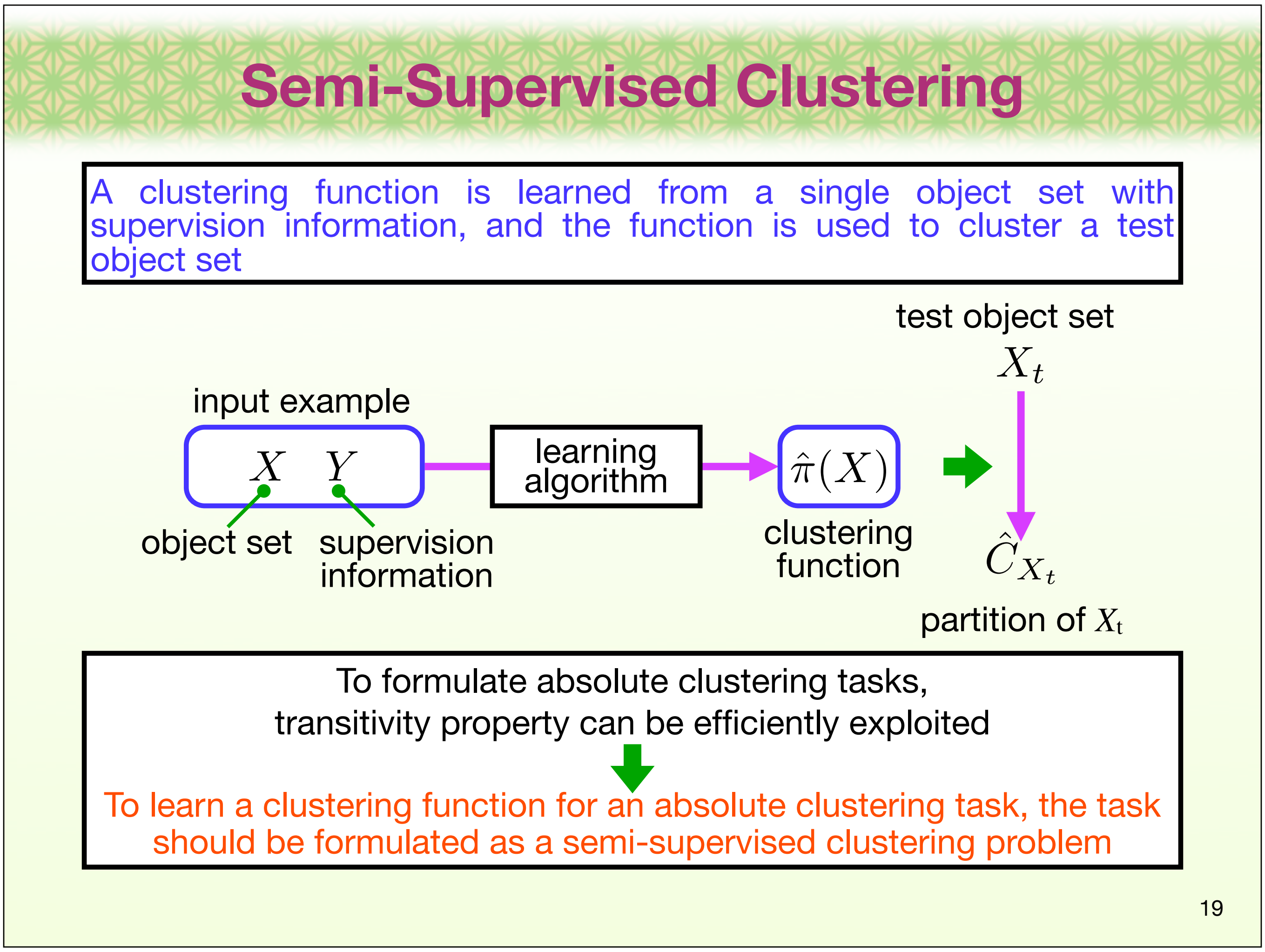
The distinction between absolute and relative clustering becomes apparent when the contents of an object set change

There is no need to differentiate between absolute and relative clustering, because an object set is invariant

In a case of transductive clustering, a single object set with supervision information is given, and the goal of learning is to obtain a partition of the set.

The distinction between absolute and relative clustering becomes apparent when the contents of an object set change.

There is no need to differentiate between absolute and relative clustering, because an object set is invariant.



Semi-Supervised Clustering

A clustering function is learned from a single object set with supervision information, and the function is used to cluster a test object set

input example

X Y

object set

supervision information

learning algorithm

$\hat{\pi}(X)$

clustering function

test object set

X_t

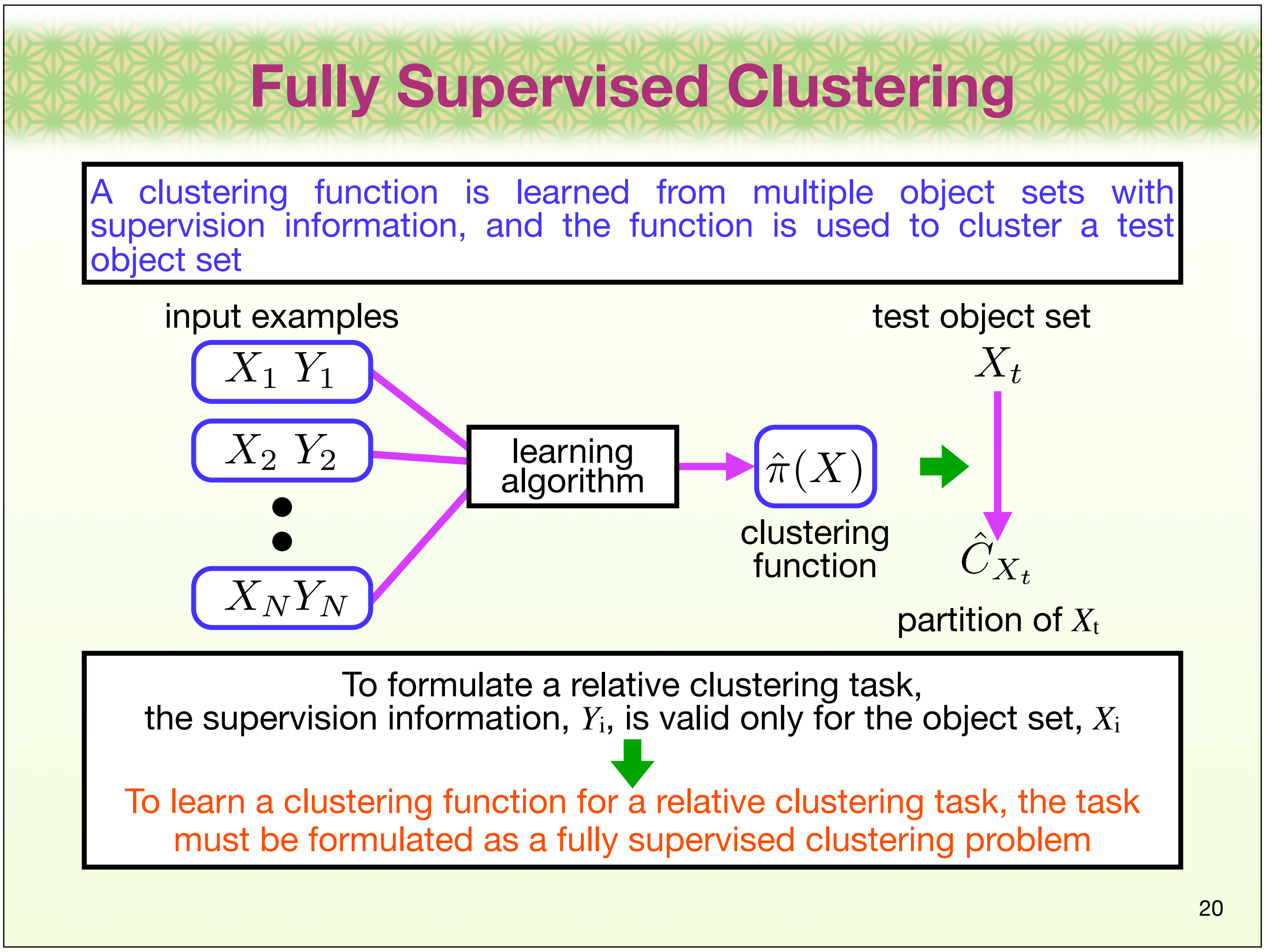
$\hat{C}_{X_t}$

partition of X_t

To formulate absolute clustering tasks,
transitivity property can be efficiently exploited

To learn a clustering function for an absolute clustering task, the task should be formulated as a semi-supervised clustering problem

In a case of semi-supervised clustering, A clustering function is learned from a single object set with supervision information, and the function is used to cluster a test object set.
To formulate absolute clustering tasks, transitivity property of absolute clustering can be efficiently exploited.
Therefore, to learn a clustering function for an absolute clustering task, the task should be formulated as a semi-supervised clustering problem.



Fully Supervised Clustering

A clustering function is learned from multiple object sets with supervision information, and the function is used to cluster a test object set

input examples

$X_1\ Y_1$

$X_2\ Y_2$

$X_N Y_N$

learning algorithm

$\hat{\pi}(X)$

clustering function

test object set

X_t

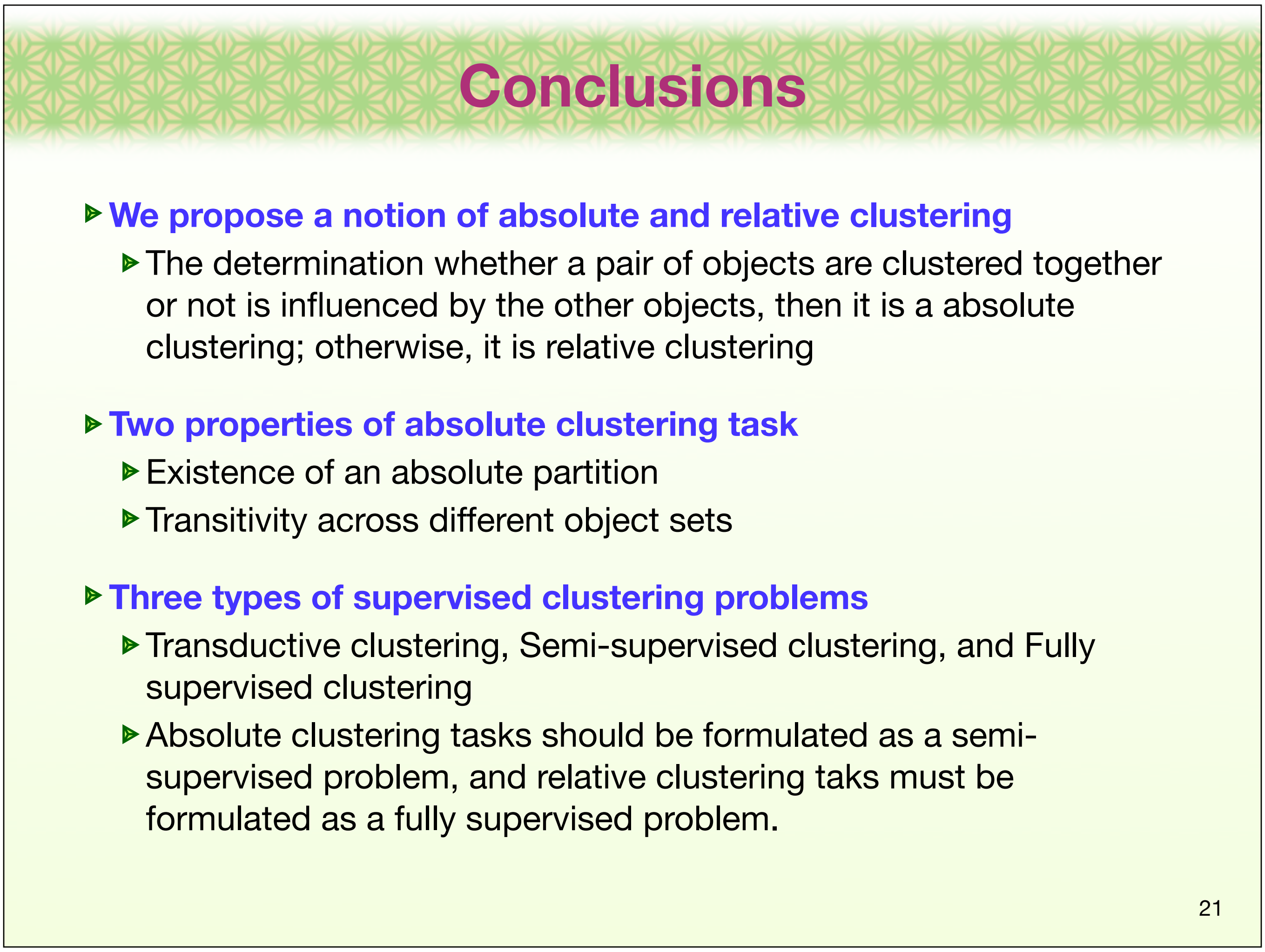
$\hat{C}_{X_t}$

partition of X_t

To formulate a relative clustering task,
the supervision information, Y_i, is valid only for the object set, X_i

To learn a clustering function for a relative clustering task, the task must be formulated as a fully supervised clustering problem

In a case of fully supervised clustering, a clustering function is learned from multiple object sets with supervision information, and the function is used to cluster a test object set.
To formulate relative clustering tasks, the supervision information is valid only for its corresponding object set
Therefore, to learn a clustering function for a relative clustering task, the task must be formulated as a fully supervised clustering problem.



Conclusions

- **We propose a notion of absolute and relative clustering**
  - The determination whether a pair of objects are clustered together or not is influenced by the other objects, then it is a absolute clustering; otherwise, it is relative clustering
- **Two properties of absolute clustering task**
  - Existence of an absolute partition
  - Transitivity across different object sets
- **Three types of supervised clustering problems**
  - Transductive clustering, Semi-supervised clustering, and Fully supervised clustering
  - Absolute clustering tasks should be formulated as a semi-supervised problem, and relative clustering taks must be formulated as a fully supervised problem.

Our conclusions are as follows.
That’s all I have to say, thank you for your attention.