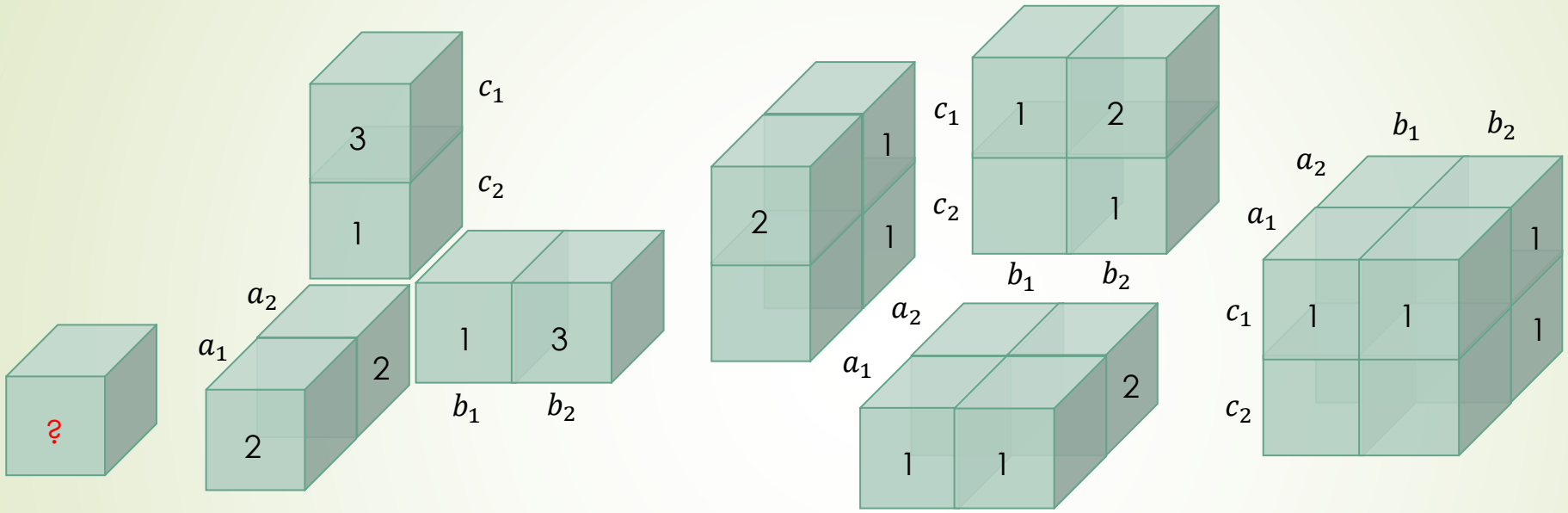
香港城市大學
City University of Hong Kong
專業 創新 胸懷全球
Professional·Creative
For The World

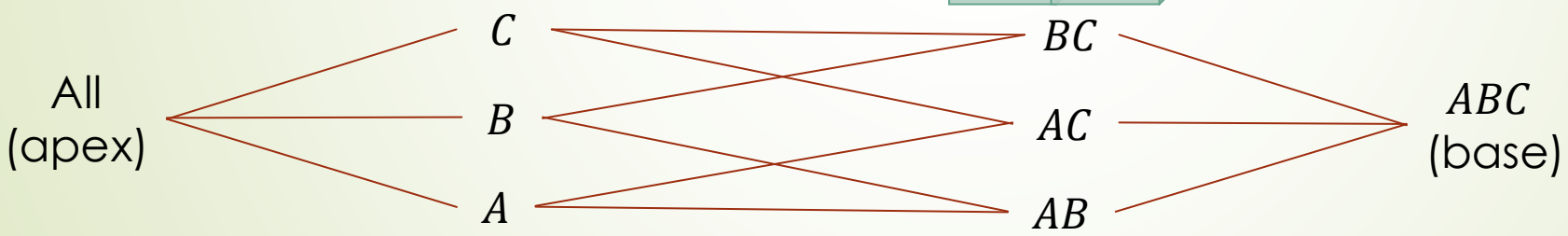
Data Cube Computation:
Closed Cube and Iceberg Cube

CS5483 Data Warehousing and Data Mining

Data cube for count

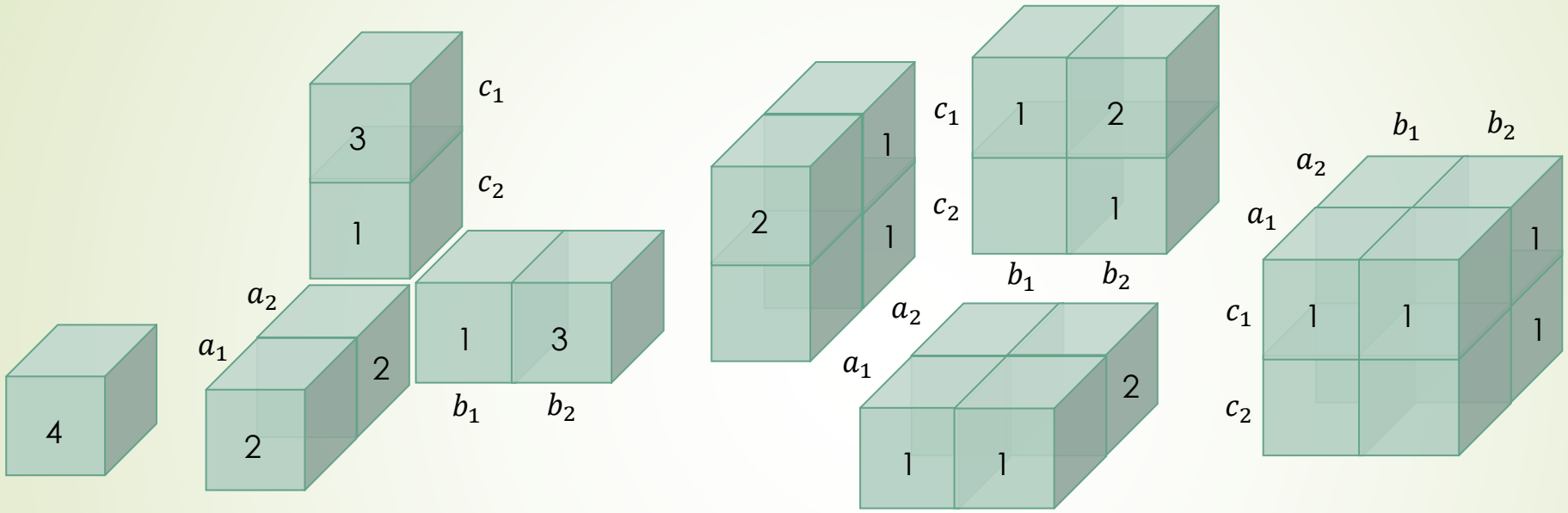


A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1



- How to avoid drawing the cubes?

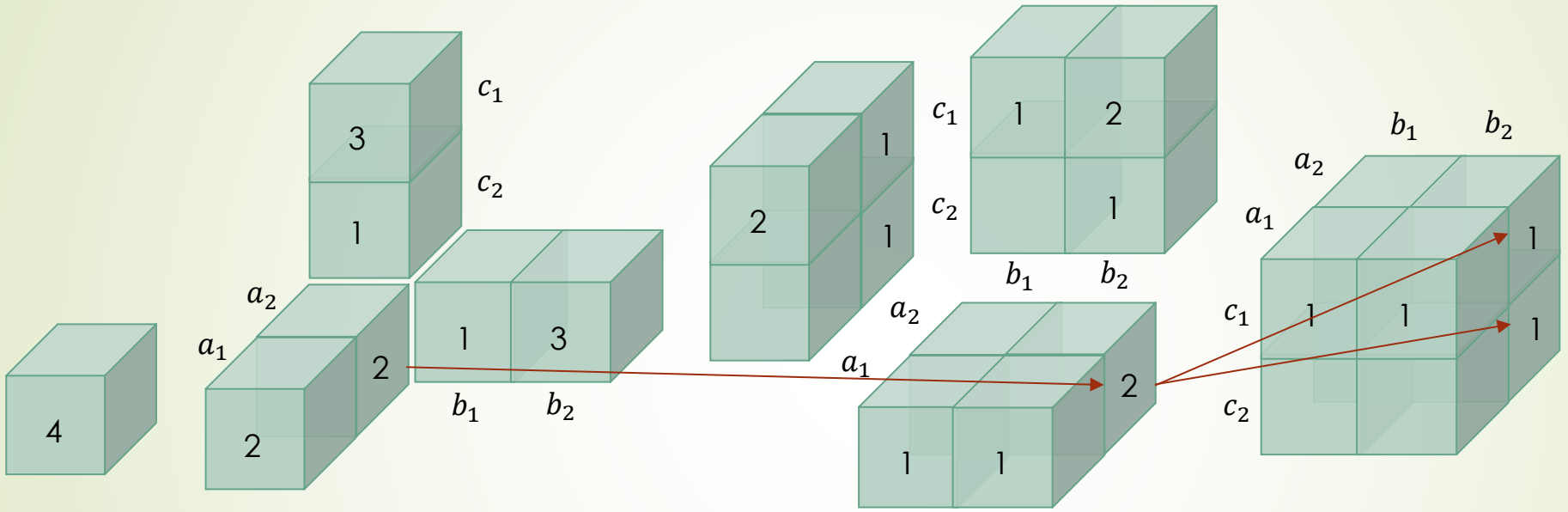
Notations



A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1

- A cell in the base cuboid is written as (a_i, b_j, c_k): fact.
  - E.g., (a_1, b_2, c_2):_____.
- For cells in other cuboid, use the star/wildcard * to match any attribute values.
  - E.g., $(a_2, b_2, *)$: 2 is in the AB-cuboid, _____ denotes the apex.

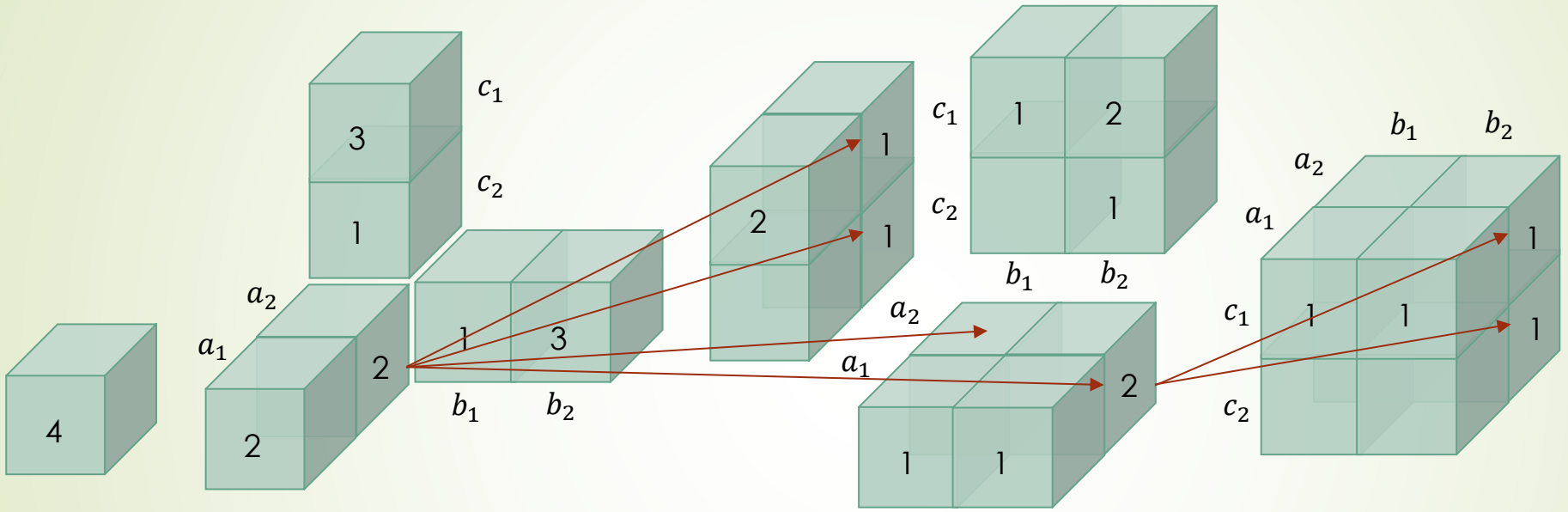
Ancestor-descendant relationship



A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1

- $(a_2, b_2, *)$ is the **p**_____ of (a_2, b_2, c_1) and (a_2, b_2, c_2).
- $(a_2, b_2, *)$ is a **c**_____ of $(a_2, *, *)$.
- $(a_2, *, *)$ is an **a**_____ of its **d**_____ (a_2, b_2, c_1) and (a_2, b_2, c_2).
- Why consider such relationships?

Aggregation for count



A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1

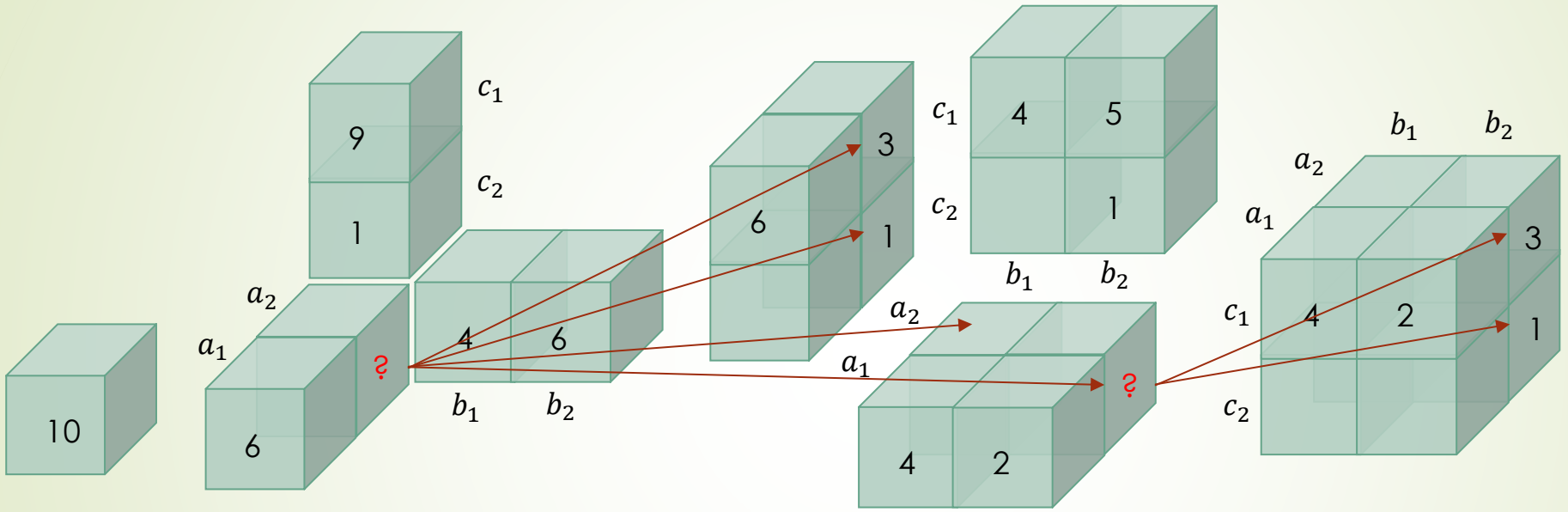
- $\text{count}(a_2, b_2, *) = \text{count}(\_\_\_\_\_) + \text{count}(\_\_\_\_\_)$
- $\text{count}(a_2, *, *) = \text{count}(a_2, b_1, *) + \text{count}(a_2, b_2, *)$

  $= \text{count}(\_\_\_\_\_) + \text{count}(\_\_\_\_\_)$
- What other functions can be computed this way?

Distributive aggregate functions

- Functions that can be computed in a distributive manner:
  - The fact of a parent cell can be obtained from the facts of its children.
  - E.g., count, sum, min and max.

Aggregation for sum

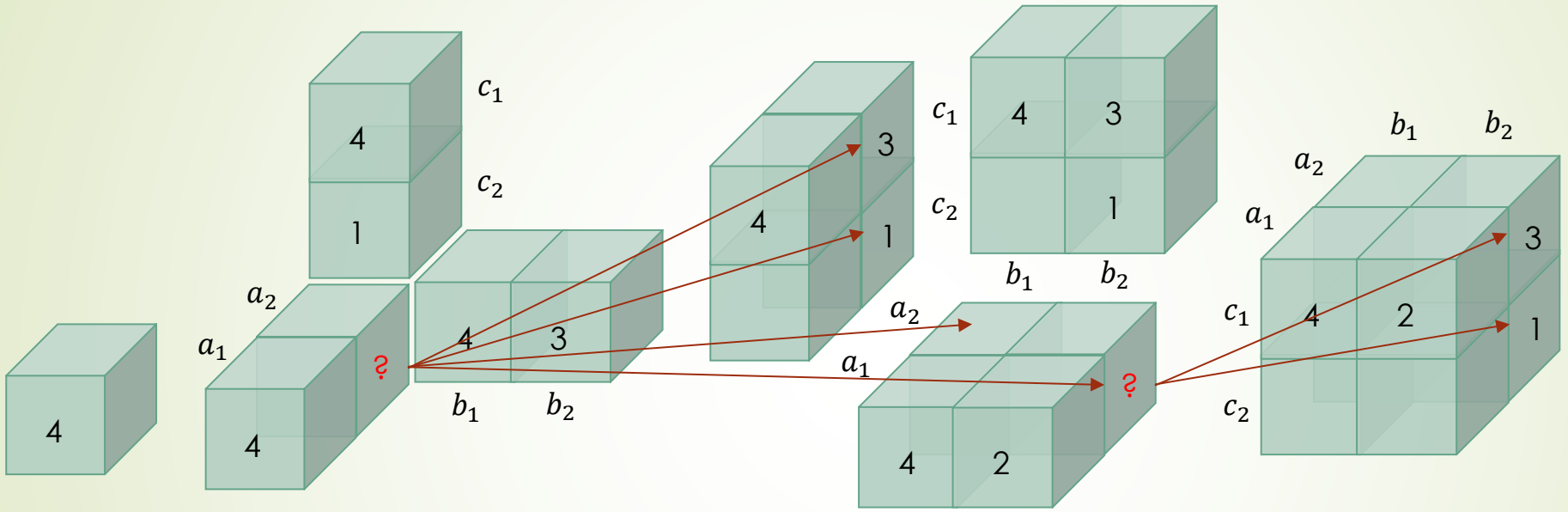


A	B	C	$
a_2	b_2	c_2	1
a_1	b_2	c_1	2
a_2	b_2	c_1	3
a_1	b_1	c_1	4

- $\text{sum}(a_2, b_2, *) = \text{sum}(a_2, b_2, c_1) + \text{sum}(a_2, b_2, c_2)$
- $\text{sum}(a_2, *, *) = \text{sum}(a_2, b_1, *) + \text{sum}(a_2, b_2, *)$

  $= \text{sum}(a_2, *, c_1) + \text{sum}(a_2, *, c_2)$
- Sum reduces to count when _____.

Aggregation for max



A	B	C	$
a_2	b_2	c_2	1
a_1	b_2	c_1	2
a_2	b_2	c_1	3
a_1	b_1	c_1	4

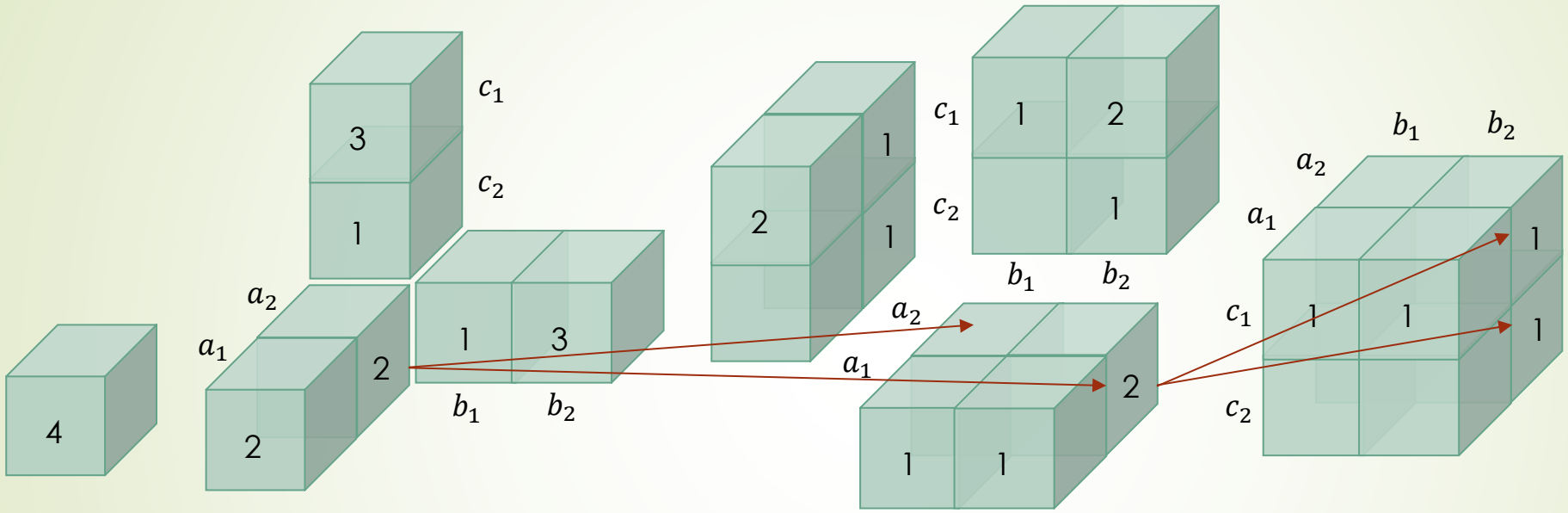
- $\max(a_2, b_2, *) = \max\{\max(a_2, b_2, c_1), \max(a_2, b_2, c_2)\}$
- $\max(a_2, *, *) = \max\{\max(a_2, b_1, *), \max(a_2, b_2, *)\}$

  $= \max\{\max(a_2, *, c_1), \max(a_2, *, c_2)\}$

Algebraic aggregate functions

- **Algebraic** functions of a **constant number** of distributive functions called **subaggregates**:
  - avg = sum / count
  - std_dev = _____
- Non-algebraic (**h**_____) aggregate functions are difficult to compute:
  - Median, Mode, Rank, …
  - E.g.: with any $10000 < x < 20000$,
    Male income values: $10000, 10000, x$
    Female income values: $x, 30000, 30000$
    Overall median income: ____ which is sensitive to the distribution of incomes.

How to characterize a data cube efficiently?

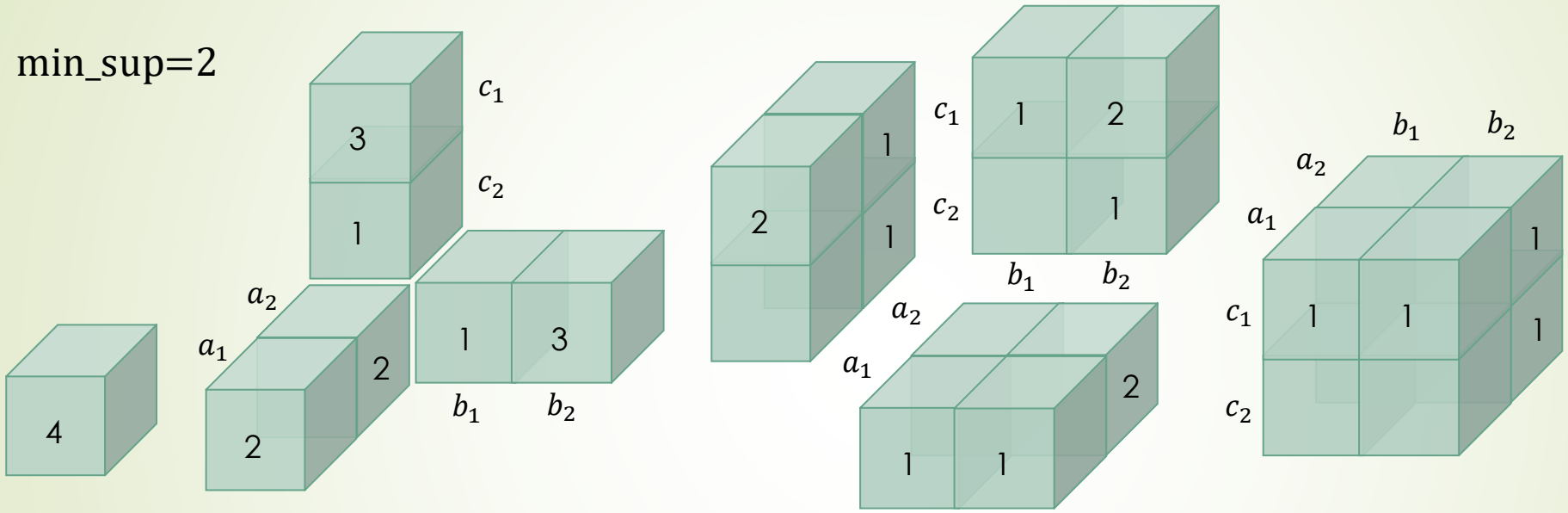


A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1

- **Closed cube**: keep only the **closed** cells having *no descendant with the same count*.
  - E.g., $(a_2, b_2, *)$ is closed because _____.
  - E.g., $(a_2, *, *)$ is not closed because _____.
- The count of a cell not in the closed cube is the maximum/minimum of the counts of its ancestor/descendants in the closed cube.

Iceberg cube

min_sup=2

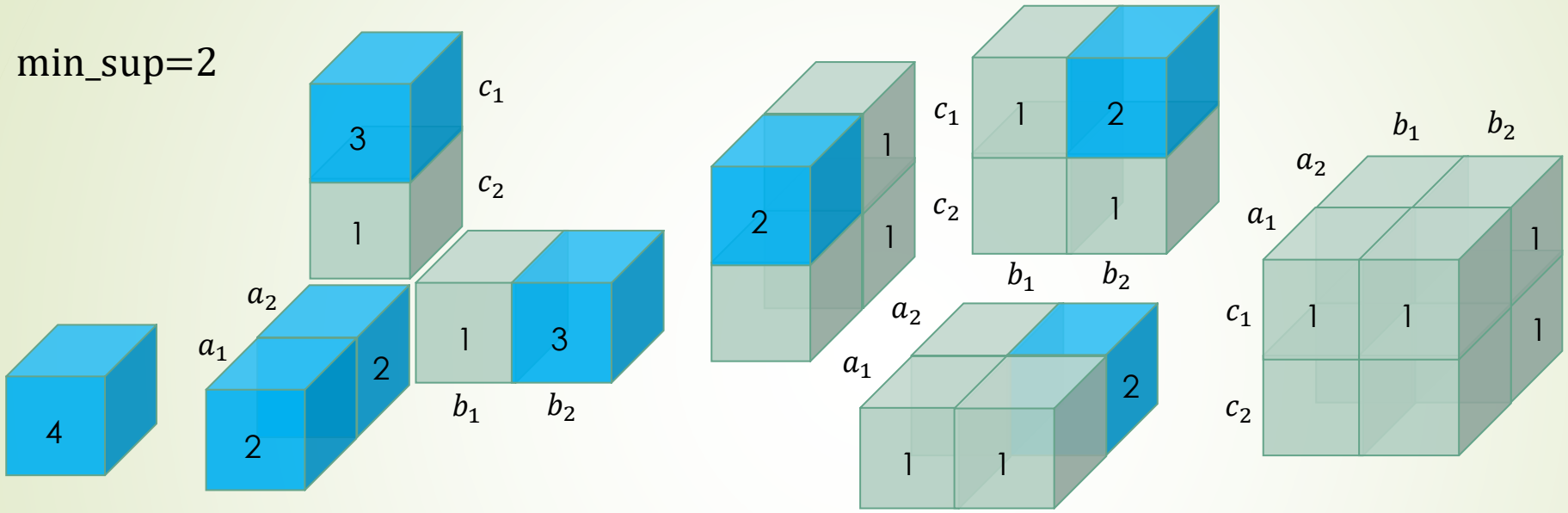


A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1

- **Iceberg cube**: Keep only cells with the **iceberg condition** counts ≥ min_sup.

Closed iceberg cube

min_sup=2



A	B	C
a_2	b_2	c_2
a_1	b_2	c_1
a_2	b_2	c_1
a_1	b_1	c_1

- Which cells in the iceberg cubes are not closed? _____
- Hass diagram representation:

$(*,*,*): 4$ — $(*,*,c_1): 3$, $(*,b_2,*): 3$

$(*,*,c_1): 3$ — $(*,b_2,c_1): 2$, $(a_1,*,c_1): 2$

$(*,b_2,*): 3$ — $(*,b_2,c_1): 2$, $(a_2,b_2,*): 2$

References

- 5.1 Data Cube Computation: Preliminary Concepts