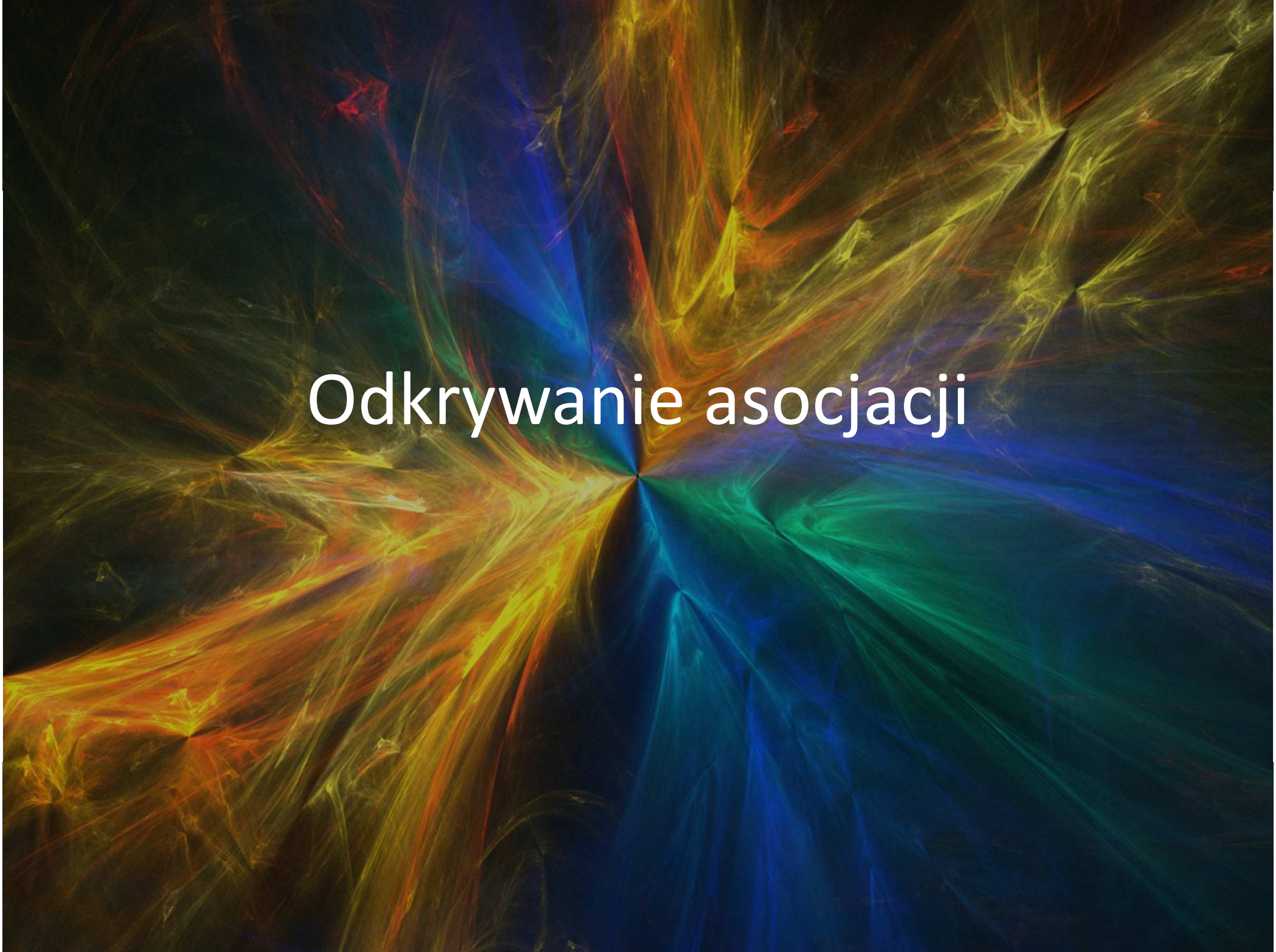
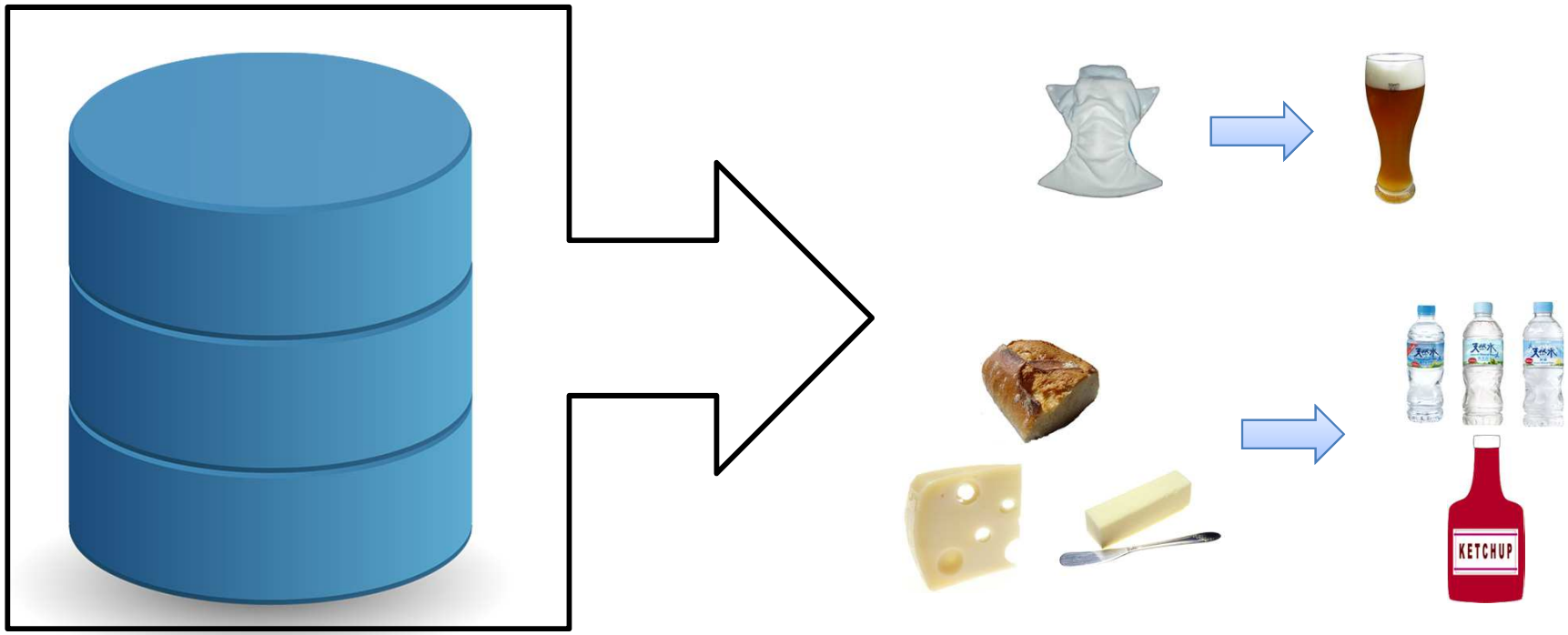

Odkrywanie asocjacji

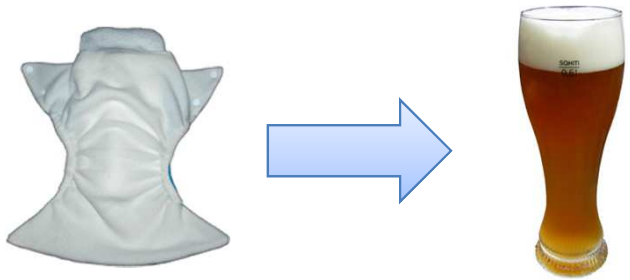
Cel odkrywania asocjacji

- Znalezienie interesujących zależności lub korelacji, tzw. *asocjacji*
  - Analiza dużych zbiorów danych
  - Wynik procesu: zbiór *reguł asocjacyjnych*



Geneza problemu - MBA

TS	ID_KLIENTA	Koszyk
12:57	1123	{mleko, pieluszki, piwo}
13:12	1412	{mleko, piwo, bułki, masło}
….	….	….



Tablica obserwacji

- Atrybuty A={A1,A2,…,An}, np. lista towarów
- Obserwacje T={T1,T2,…,Tm}, np. koszyki
- 0/1 – Czy dany towar jest w danym koszyku

Tr_id	A1	A2	A3	…	An
T1	1	0	0	…	1
T2	1	1	1	…	1
T3	1	0	1	…	0
T4	0	0	1	…	1
T5	0	1	1	…	1
…	…	…	…	…	…
Tm	1	1	1	…	0

Reguły asocjacyjne

- Postać:

$$\theta \rightarrow \varphi$$

  - θ nazywamy poprzednikiem reguły
  - φ nazywamy następnikiem reguły
- Poprzednik i następnik to koniunkcje warunków Ai=1
  - $\theta = \{(Ax = 1) \Lambda (Ay = 1) \Lambda (Az = 1)\}$
  - $\varphi = \{(Ak = 1) \Lambda (Al = 1)\}$
- Cała reguła:

$$\{(Ax = 1) \Lambda (Ay = 1) \Lambda (Az = 1)\} \rightarrow \{(Ak = 1) \Lambda (Al = 1)\}$$

- Interpretacja:

  Jeżeli klient kupił Ax, Ay i Az, to prawdopodobnie kupi też Ak i Al.

Reguły asocjacyjne

- Regułę

$$\{(Ax = 1) \wedge (Ay = 1) \wedge (Az = 1)\} \rightarrow \{(Ak = 1) \wedge (Al = 1)\}$$

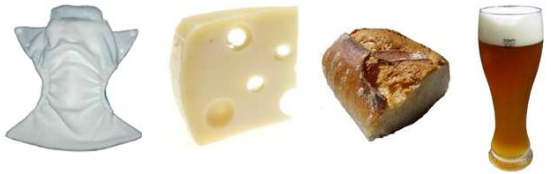
- Można zapisać bardziej zwięźle:

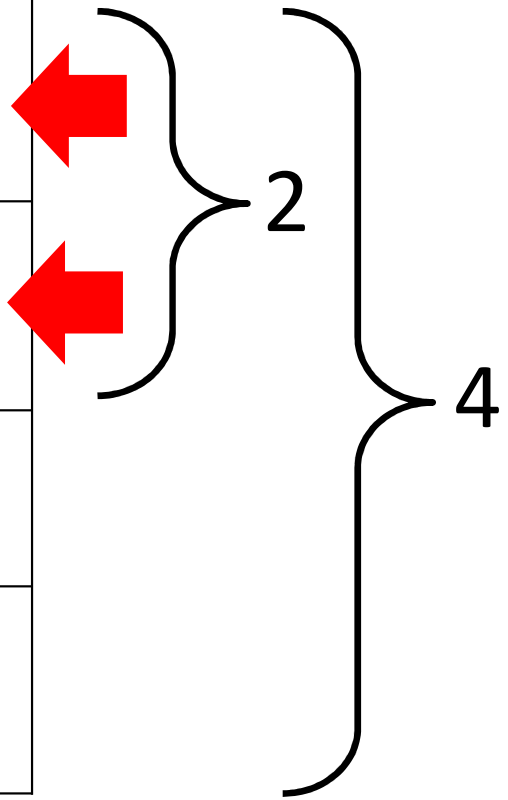
$$\{Ax, Ay, Az\} \rightarrow \{Ak, Al\}$$

- Miary oceny reguł asocjacyjnych
  - Wsparcie (ang. *support*) - $\sup(\theta \rightarrow \varphi)$
  - Ufność (ang. *confidence*) - $\mathrm{conf}(\theta \rightarrow \varphi)$

Wsparcie

$$\sup(\ \text{  } \rightarrow \text{  }\ ) = \frac{2}{4} = 50\%$$

Tr_id	Koszyk
1	
2	
3	
4	



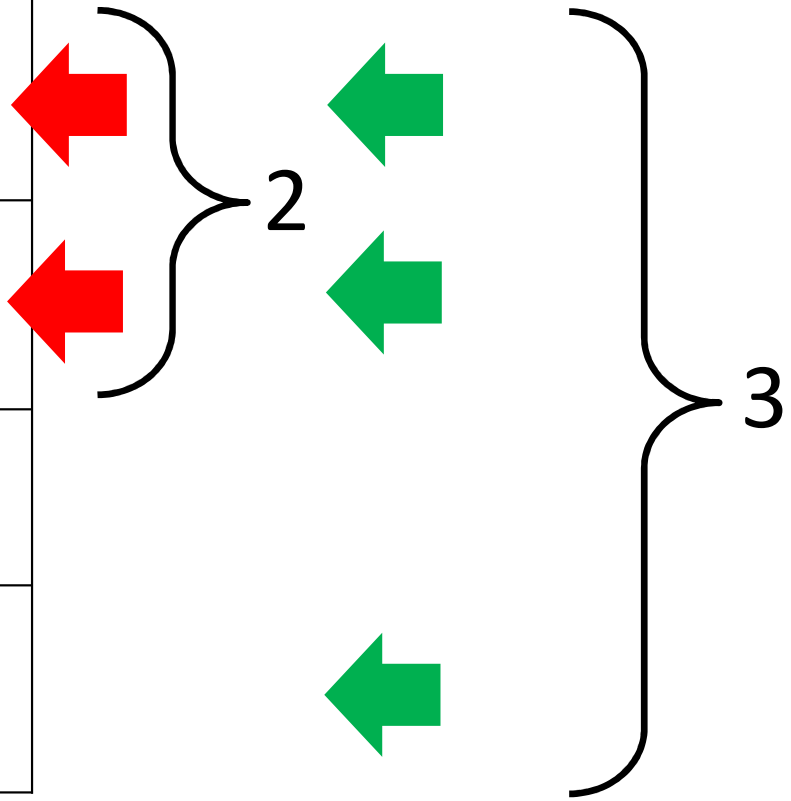
Wsparcie reguły $\theta \rightarrow \varphi$

- Stosunek liczby obserwacji spełniających warunek: $\theta \wedge \varphi$ do liczby wszystkich obserwacji
- Oszacowanie prawdopodobieństwa zajścia zdarzenia $\mathrm{p}(\theta \wedge \varphi)$.

Ufność

$$\text{conf}(\; \rightarrow \;) = \frac{2}{3} = 66,(6)\%$$

Tr_id	Koszyk
1	
2	
3	
4	



Ufność reguły $\theta \rightarrow \varphi$

- Stosunek liczby obserwacji spełniających warunek: $\theta \wedge \varphi$ do liczby obserwacji spełniających warunek θ.
- Oszacowanie prawdopodobieństwa warunkowego zajścia φ pod warunkiem zajścia θ - $\mathrm{p}(\varphi|\theta)$
- Ufność można również obliczyć ze wzoru:

$$conf(\theta \wedge \varphi) = \frac{\sup(\theta \wedge \varphi)}{\sup(\theta)}$$

Binarne reguły asocjacyjne

- Silna reguła asocjacyjna:
  - Jej wsparcie jest większe niż próg *minsup*
  - Jej ufność jest większa niż próg *minconf*
- Progi te są zadawane przez użytkownika
- Odkrywanie reguł asocjacyjnych:
  - Wejście: baza danych, *minsup*, *minconf*
  - Wynik: zbiór silnych reguł asocjacyjnych

Przykład

Tr_id	Koszyk
1	
2	
3	
4	
5	

$$minsup = 0.4$$

$$minconf = 0.6$$

Reguła	sup	conf
piwo → orzeszki	0,6	1
orzeszki → piwo	0,6	0,75
piwo ^ pieluszki → orzeszki	0,4	1
pieluszki ^ orzeszki → piwo	0,4	1
pieluszki → piwo ^ orzeszki	0,4	1
pieluszki → piwo	0,4	1
pieluszki → orzeszki	0,4	1
piwo ^ orzeszki → pieluszki	0,4	0,67

Naiwny algorytm

- Dane: I – zbiór wszystkich elementów (towarów), *minsup*, *minconf*.

1. Wygeneruj wszystkie podzbiory zbioru I. Oblicz wsparcie każdego podzbioru.
2. Pozostaw tylko podzbiory o wsparciu>*minsup.* Są to tzw. *zbiory częste.*
3. Z każdego pozostawionego podzbioru utwórz wszystkie możliwe reguły asocjacyjne. Oblicz ufność każdej reguły.
4. Pozostaw tylko reguły o ufności>*minconf*.

Naiwny algorytm

- Dlaczego jest do niczego?
  - Liczba podzbiorów zbioru I: $n = 2^{|I|} - 1$
- W praktyce $|\mathrm{I}| \approx 200000$, czyli $n = 2^{200000} - 1$
  - Liczba atomów w ciele człowieka: $7 * 2^{90}$
  - Liczba atomów w kuli ziemskiej: $1.3 * 2^{166}$
  - Liczba atomów we wszechświecie: 2^{266}
- Szukanie zbiorów częstych jest zatem bardzo wolne.
- Algorytm APRIORI

Własność monotoniczności

- Niech będzie dany zbiór elementów I oraz jego zbiór potęgowy J = 2|L|. Mówimy, ze miara f jest monotoniczna na zbiorze J, jeżeli:

$$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(X) \leq f(Y)$$

- Własność monotoniczności miary f oznacza, ze jeżeli X jest podzbiorem zbioru Y, to wartość miary f (X) nie może być większa od f (Y)

Własność antymonotoniczności

- Niech będzie dany zbiór elementów I oraz jego zbiór potęgowy J = $2^{|L|}$. Mówimy, ze miara f jest antymonotoniczna na zbiorze J, jeżeli:

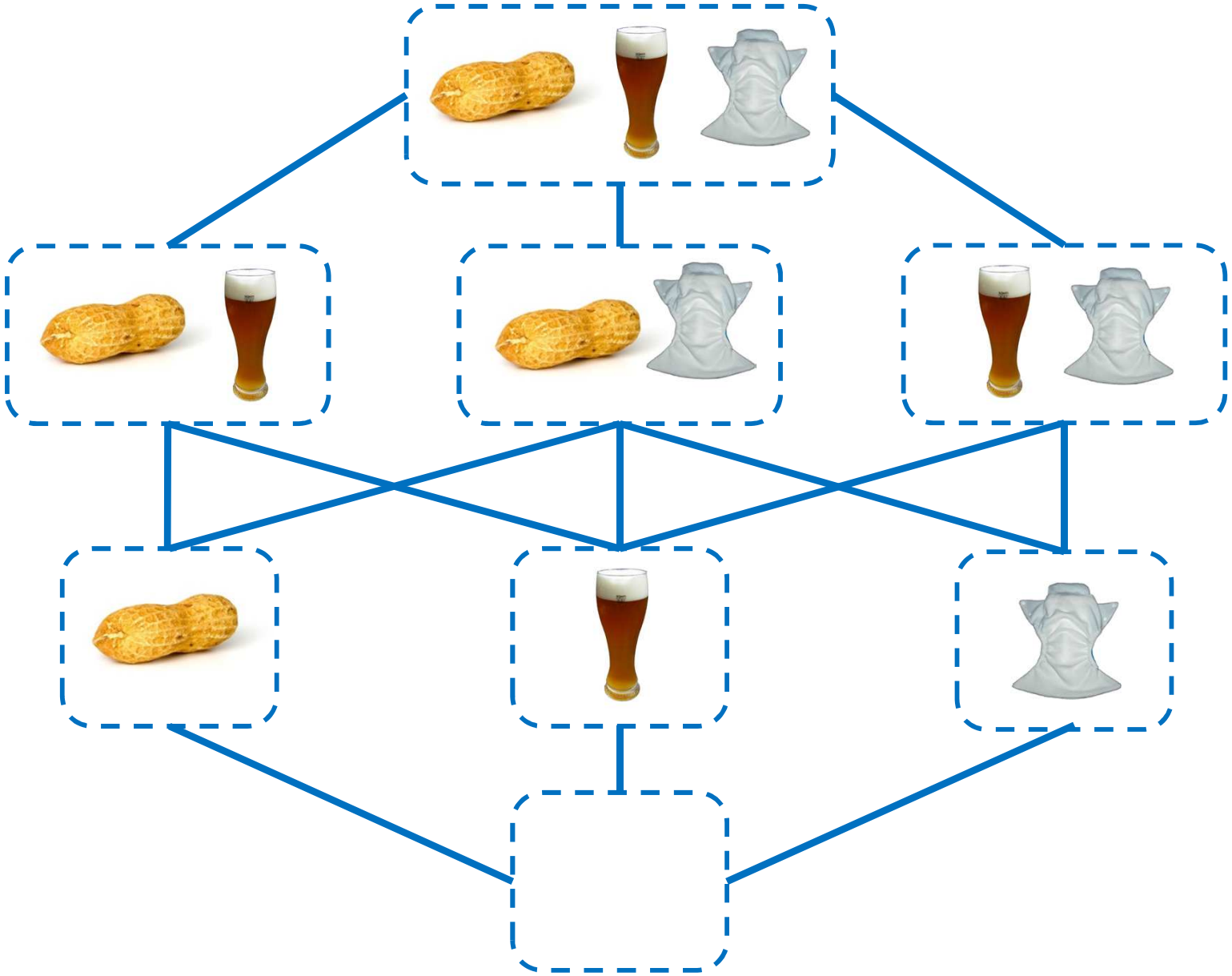
$$\forall X, Y \in J: (X \subseteq Y) \rightarrow f(X) \geq f(Y)$$

- Własność antymonotoniczności miary *f* oznacza, ze jeżeli X jest podzbiorem zbioru *Y*, to wartość miary f (Y) nie może być większa od f (X)

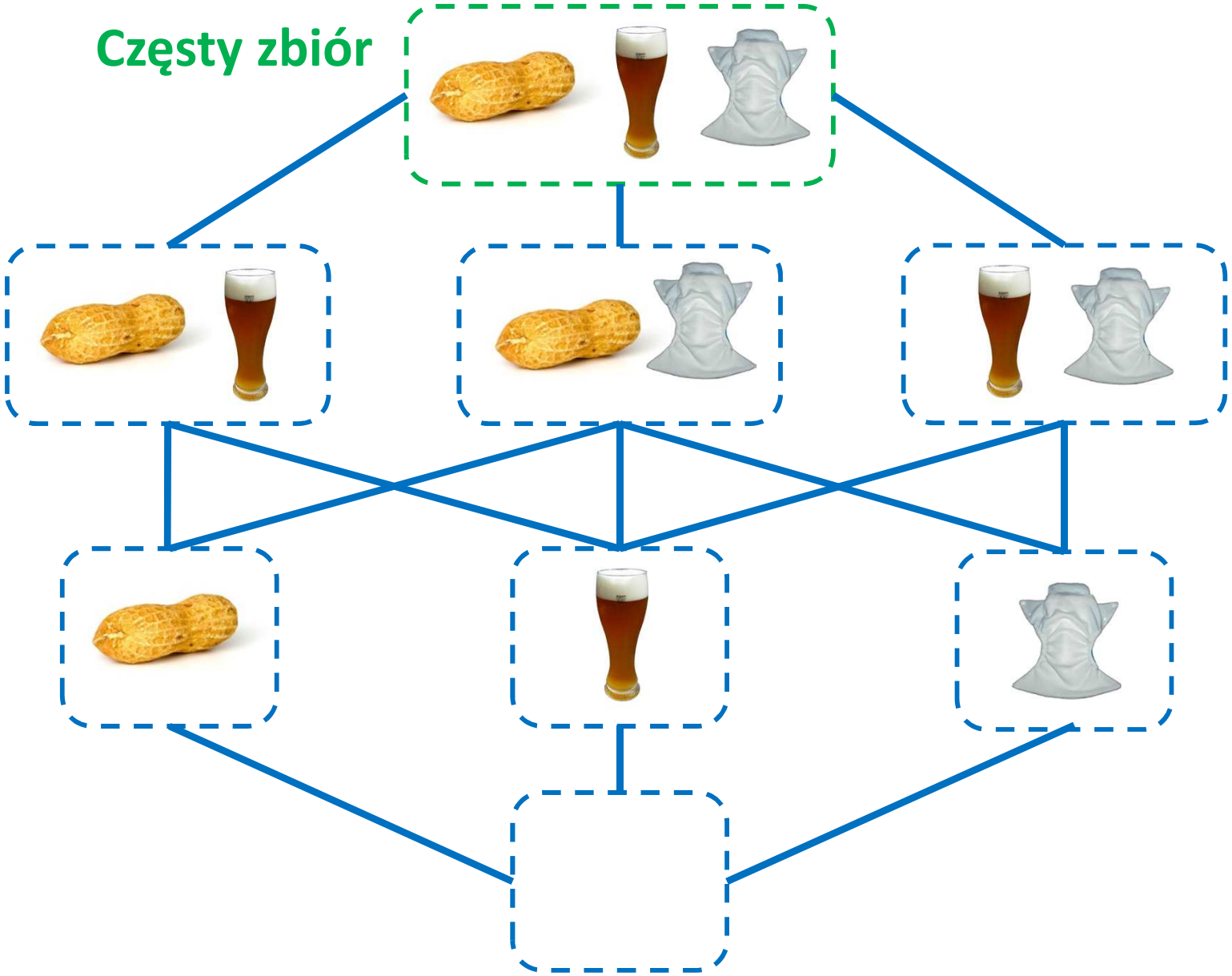
Antymonotoniczność miary wsparcia

- Własność antymonotoniczności - wszystkie podzbiory zbioru częstego muszą być częste, innymi słowy, jeżeli Y jest zbiorem częstym i $X \subseteq Y$, to X jest również zbiorem częstym
- Wniosek: jeżeli zbiór X nie jest zbiorem częstym, to żaden nadzbiór Y zbioru X, $X \subseteq Y$, nie będzie zbiorem częstym
- Nie musimy rozważać wsparcia zbioru X, którego jakikolwiek podzbiór nie jest zbiorem częstym

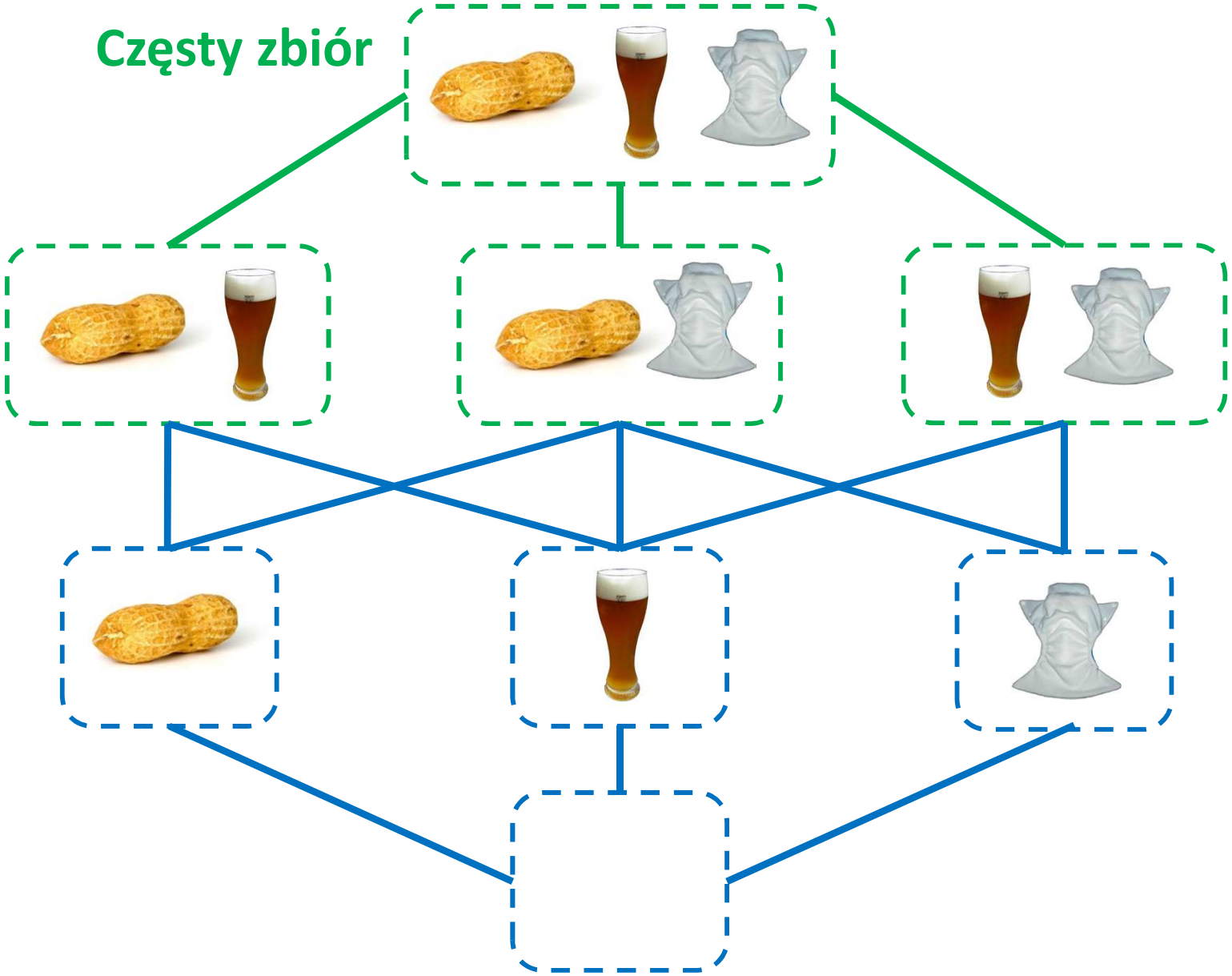
Algorytm APRIORI



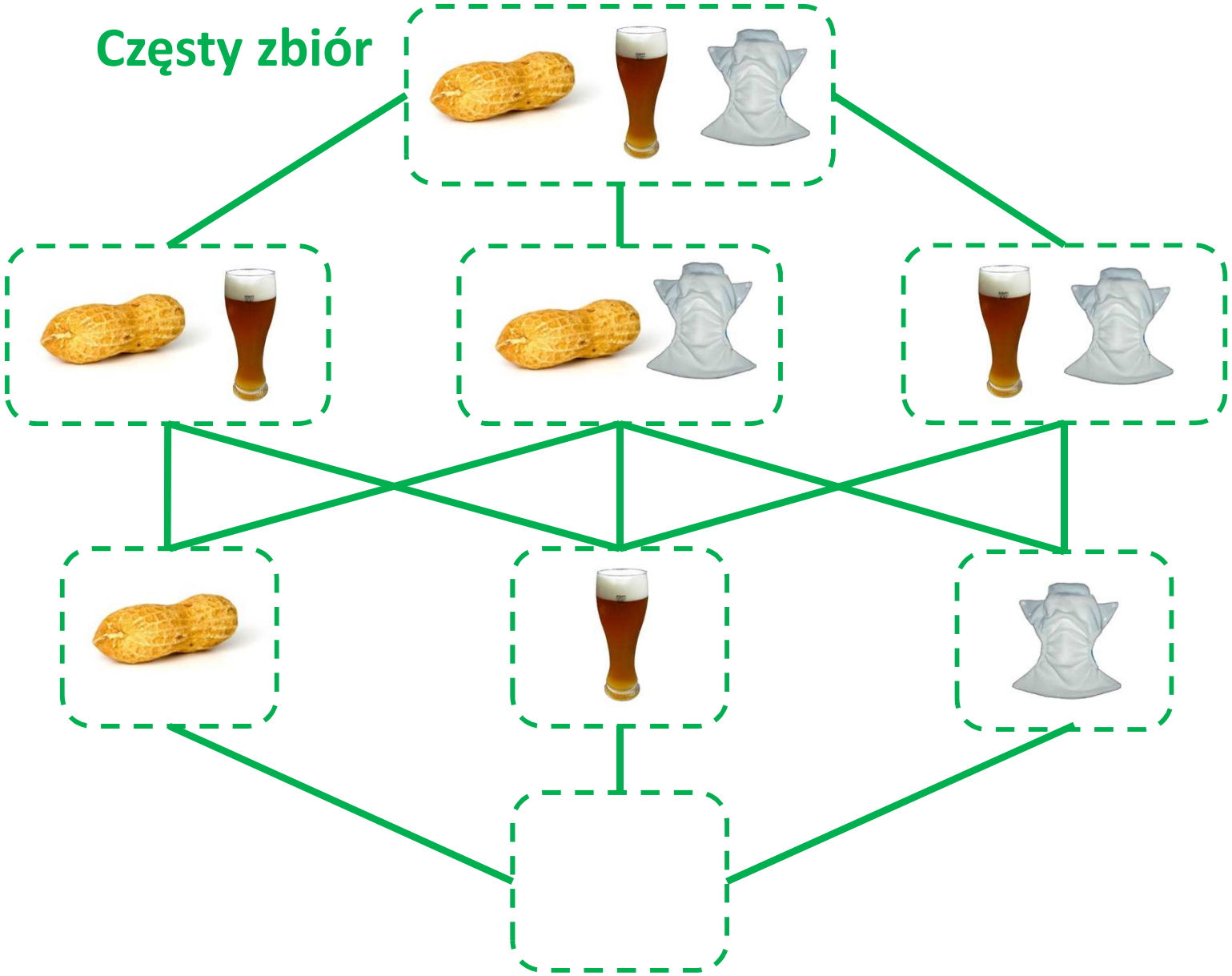
Algorytm APRIORI



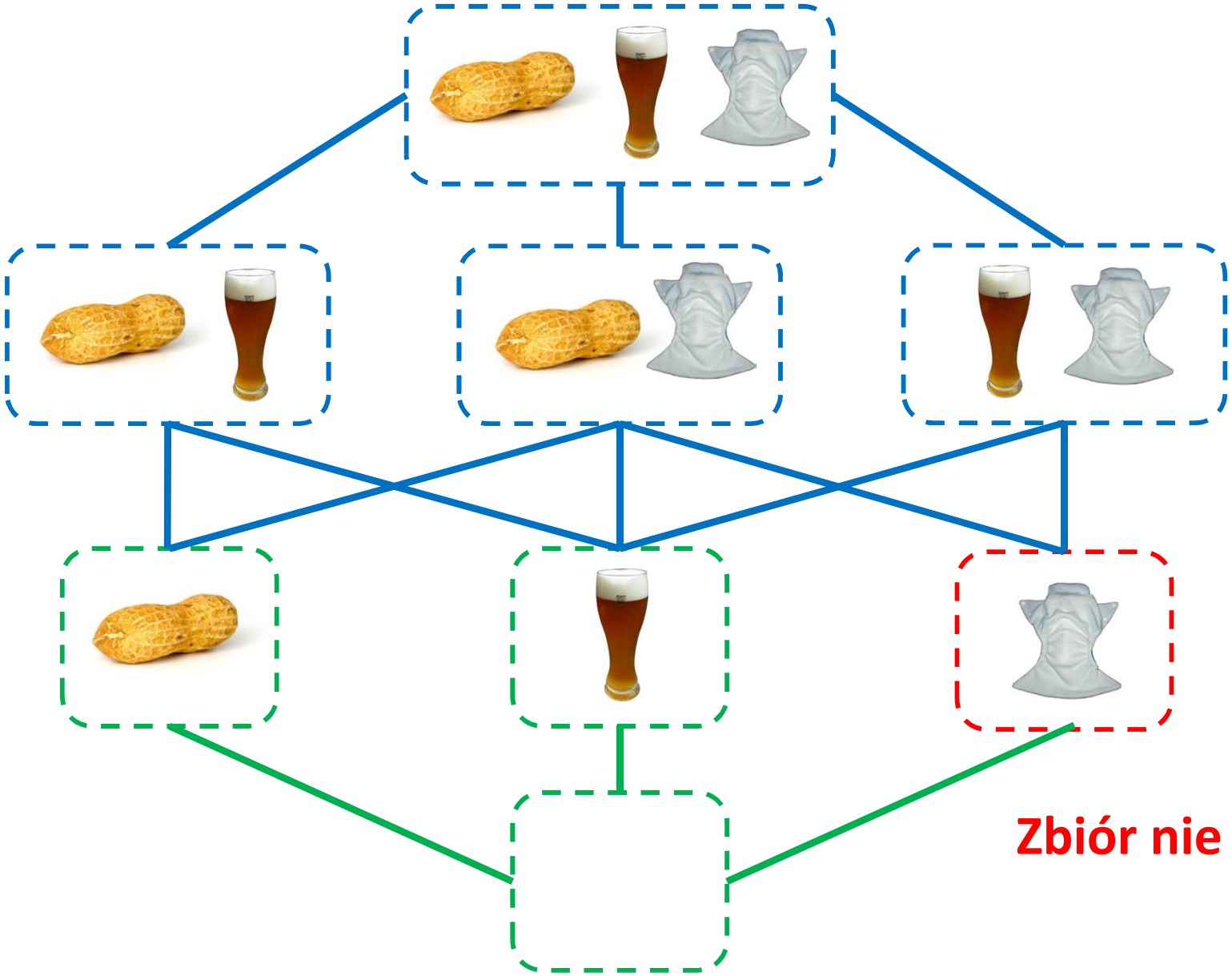
Algorytm APRIORI



Algorytm APRIORI

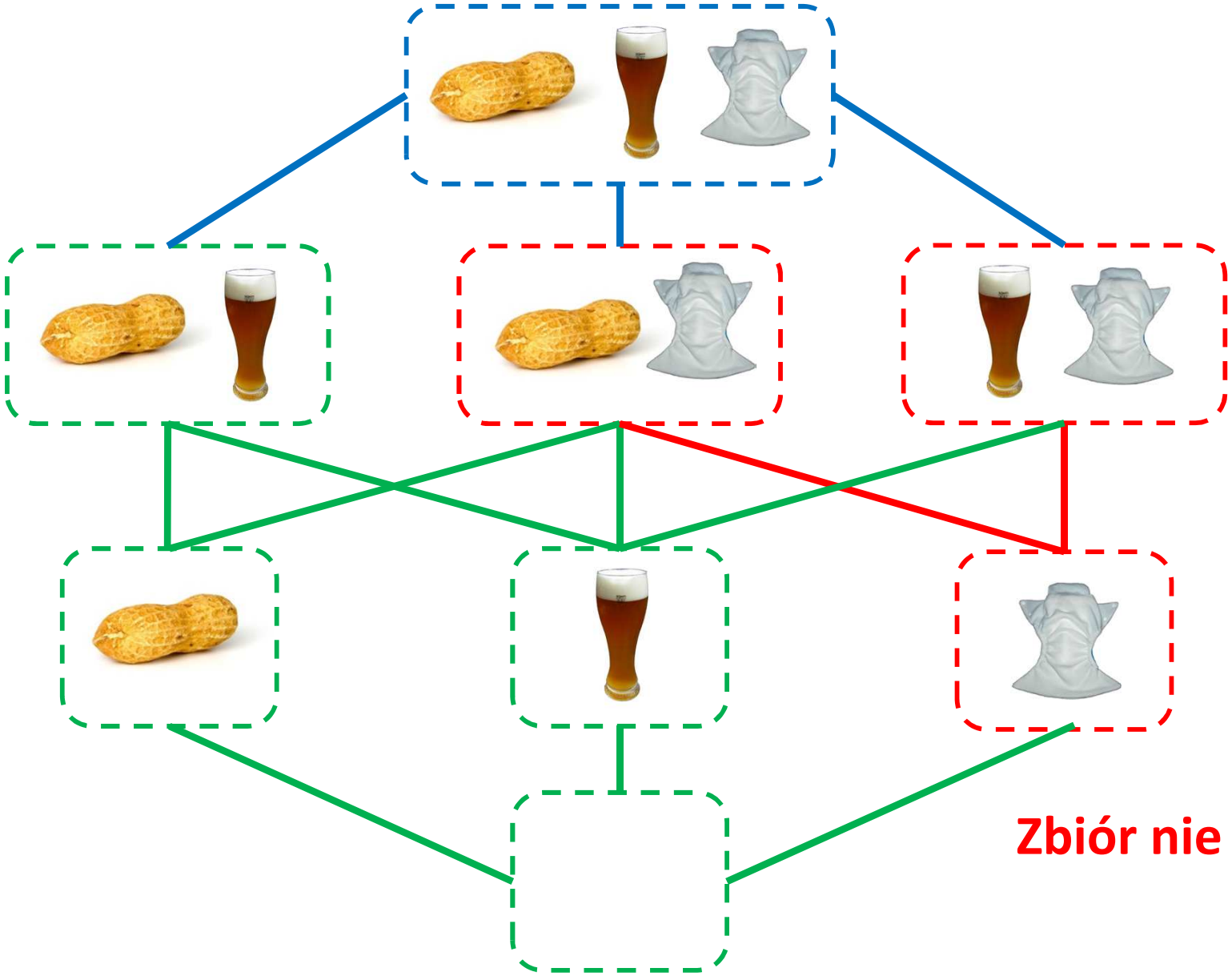


Algorytm APRIORI



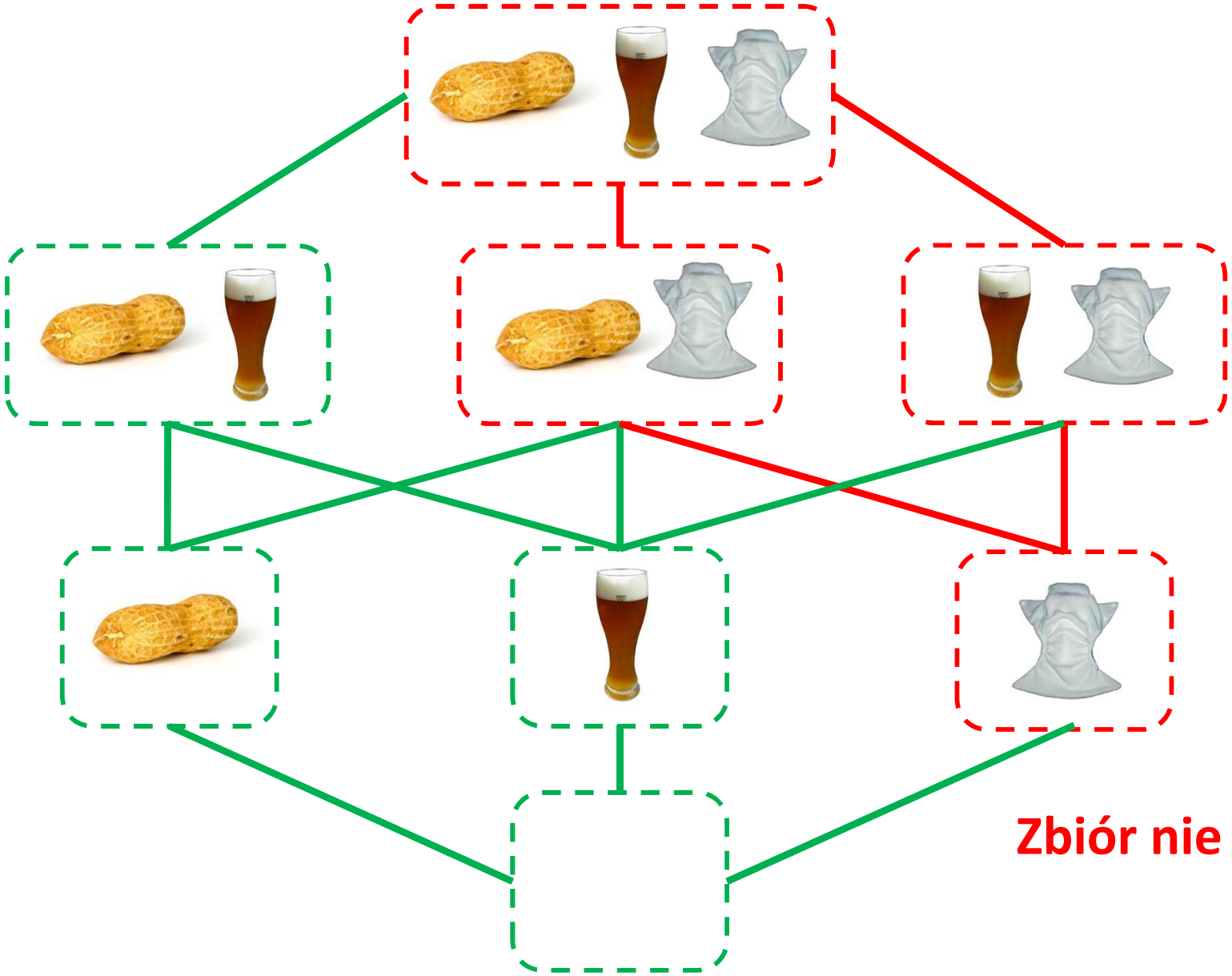
Zbiór nie częsty

Algorytm APRIORI



Zbiór nie częsty

Algorytm APRIORI



Zbiór nie częsty

Oznaczenia

- Poszczególne „towary” oznaczmy przez litery
- Zakładamy, że zbiory są posortowane, np.: $\mathrm{T1} = \{A, C, D\}$
- Przez L_k oznaczamy kolekcję zbiorów częstych o rozmiarze k.
- Przez C_k oznaczamy kolekcję zbiorów o rozmiarze k. Są to tzw. *zbiory kandydujące*.

Schemat algorytmu

- Znajdź kolekcję zbiorów częstych L1.
- Na podstawie L1 wygeneruj zbiór kandydatów C2.
- Oblicz wsparcie kandydatów ze zbioru C2.
- Usuń kandydatów, których wsparcie<*minsup.* W ten sposób powstaje kolekcja L2.
- Na podstawie L2, wygeneruj zbiór kandydatów C3….

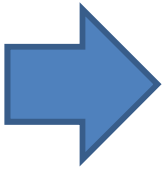
Generowanie kandydatow (1)

L1

A

B

C



C2

AB

AC

BC

L2

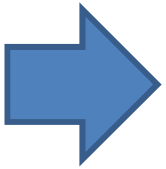
AB

AC

AD

BC

BD



C3

ABC

ABD

ACD

BCD

L3

ABC

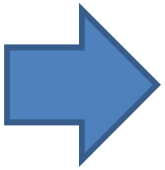
ABD

ABE

ABF

BCD

BDE



C4

ABCD

ABCE

ABCF

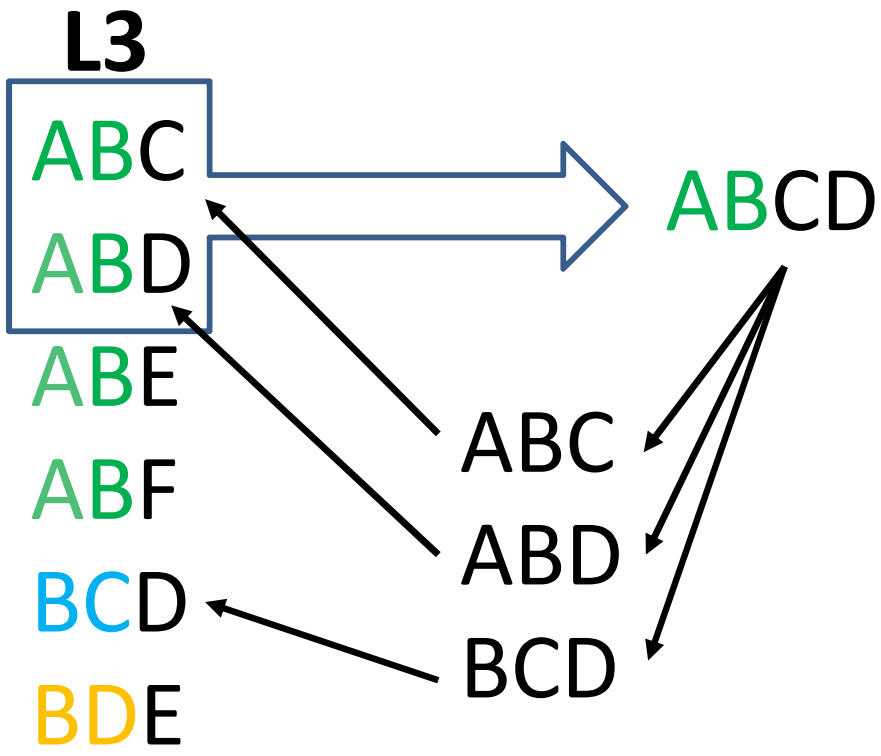
ABDE

ABDF

ABCF

Generowanie kandydatow (2)

Każdy podzbior musi być częsty!



Podzbiory 3-elementowe

Czy należy sprawdzać 2 elementowe podzbiory?

Przykład (1/3)

Tr_id	Koszyk
1	
2	
3	
4	
5	

$$minsup = 0.4$$

A **B** **C** **D**

Tr_id	Koszyk
1	AB
2	ACD
3	B
4	ABC
5	ACD

Przykład (2/3)

Tr_id	Koszyk
1	AB
2	ACD
3	B
4	ABC
5	ACD

$$minsup = 0.4$$

$$\sup(A) = 0.8$$

$$\sup(B) = 0.6$$

$$\sup(C) = 0.4$$

$$\sup(D) = 0.4$$

Kandydaci:

Kandydat	Wsparcie	Kandydat	Wsparcie	Kandydat	Wsparcie
AB	0.4				
AC	0.4	~~BC~~	0.2		
AD	0.4	~~BD~~	0	CD	0.4

Przykład (3/3)

Tr_id	Koszyk
1	AB
2	ACD
3	B
4	ABC
5	ACD

$$minsup = 0.4$$

$$\sup(AB) = 0.4$$

$$\sup(AC) = 0.4$$

$$\sup(AD) = 0.4$$

$$\sup(CD) = 0.4$$

Kandydaci:

Kandydat	Wsparcie
~~ABC~~	Nie wszystkie podzbiory są częste
~~ABD~~	Nie wszystkie podzbiory są częste
ACD	0.4

Przykład - podsumowanie

Tr_id	Koszyk
1	AB
2	ACD
3	B
4	ABC
5	ACD

Zbiór częsty	Wsparcie
A	0.8
B	0.6
C	0.4
D	0.4
AB	0.4
AC	0.4
AD	0.4
CD	0.4
ACD	0.4

Wyszukiwanie reguł

- Reguły powstają ze zbiorów częstych
- Miarą oceny reguły jest miara *confidence*

$$conf(\theta \rightarrow \varphi) = \frac{\sup(\theta \wedge \varphi)}{\sup(\theta)}$$

- Niech β będzie zbiorem częstym i niech

$$\theta \subset \beta$$

- Wówczas regułę można utworzyć następująco:

$$\theta \rightarrow (\beta - \theta)$$

Ufność reguł opartych o jeden zbiór

- Niech będzie dana reguła

$$\theta \rightarrow (\beta - \theta)$$

- Taka, że:

$$conf\left(\theta \rightarrow (\beta - \theta)\right) < minconf$$

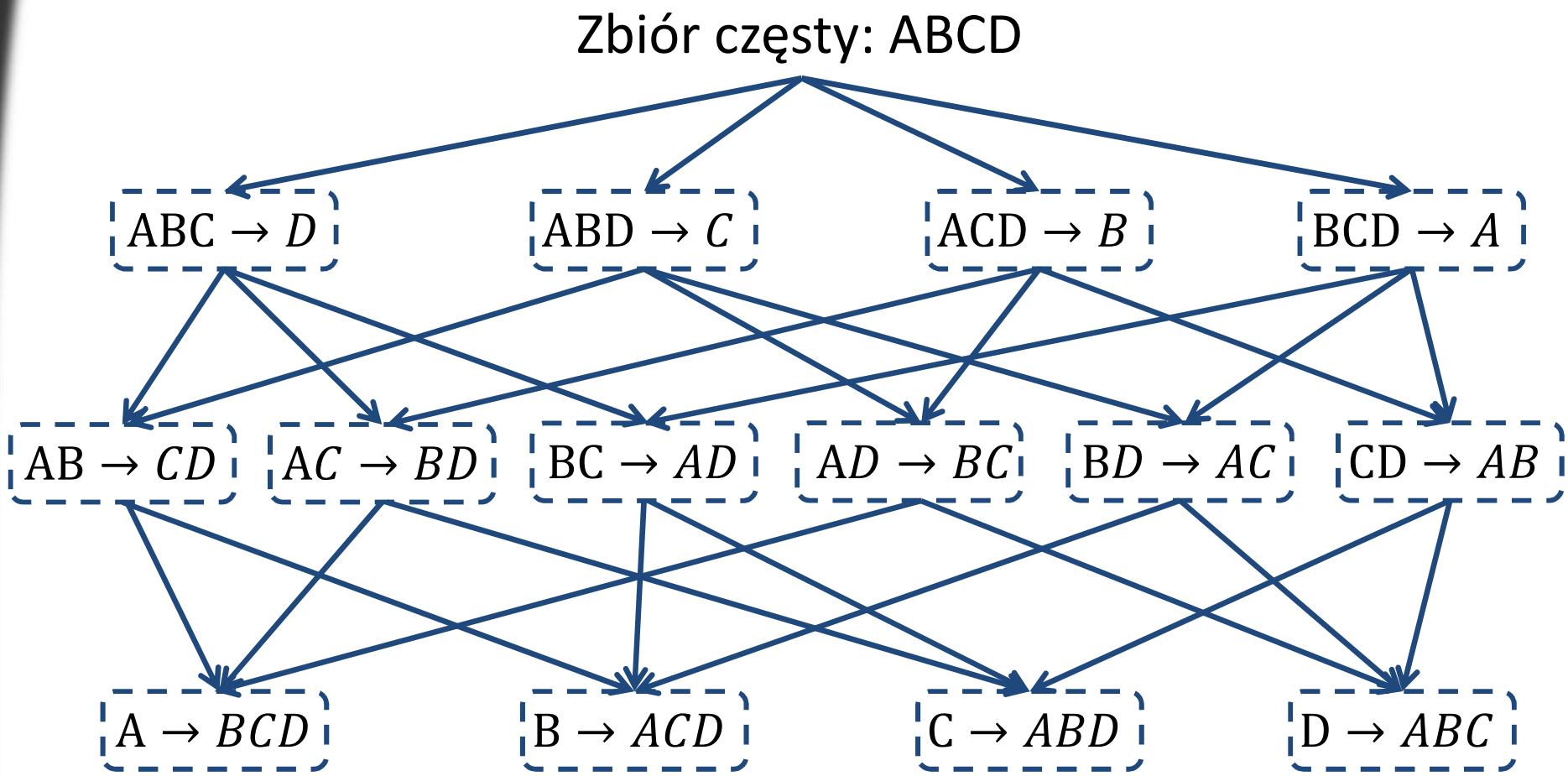
- Wówczas dowolna reguła

$$\lambda \rightarrow (\beta - \lambda)$$

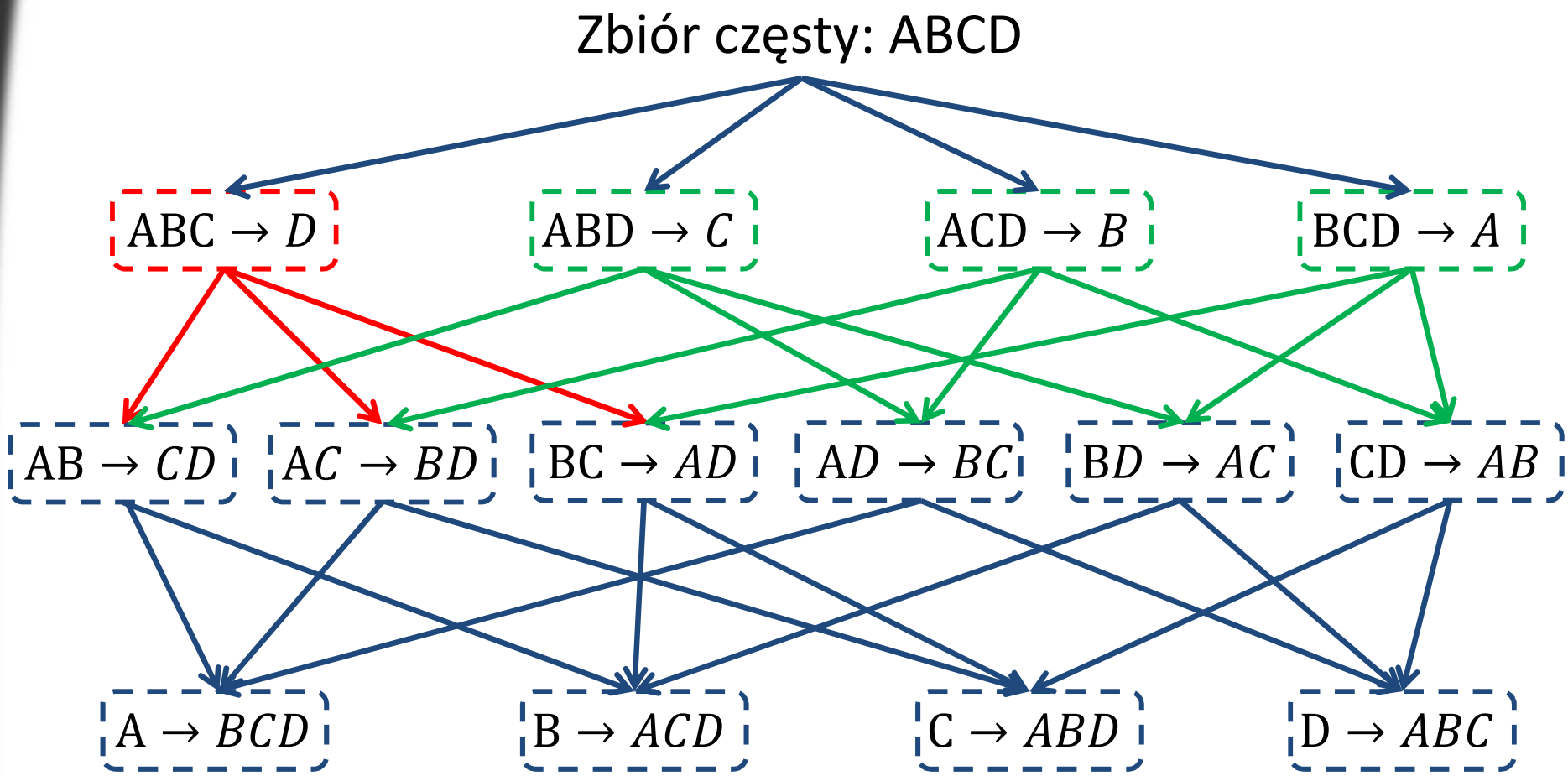
- Gdzie: $\lambda \subseteq \theta$
- Spełnia warunek:

$$conf\left(\lambda \rightarrow (\beta - \lambda)\right) < minconf$$

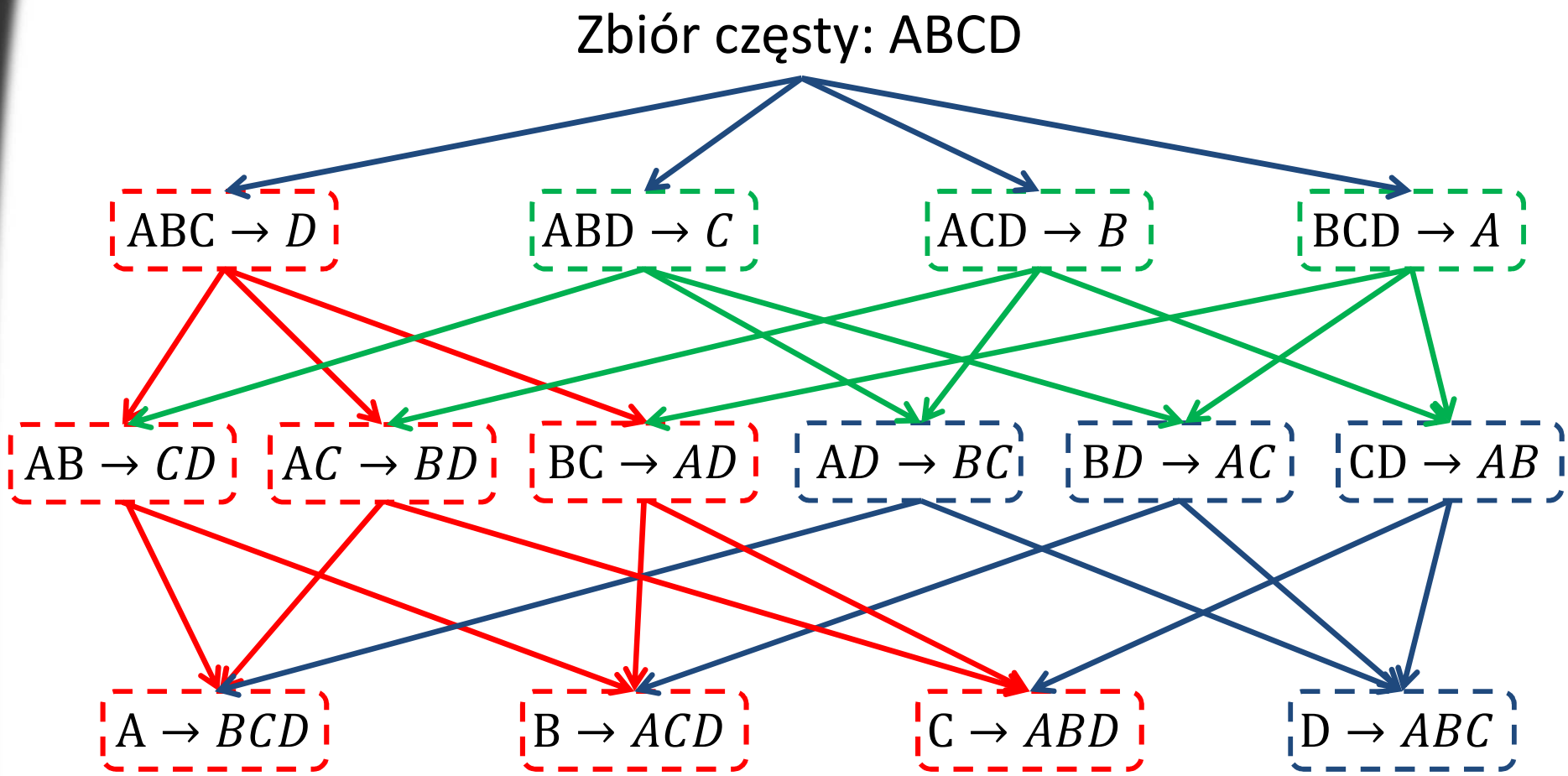
Eliminacja na kracie



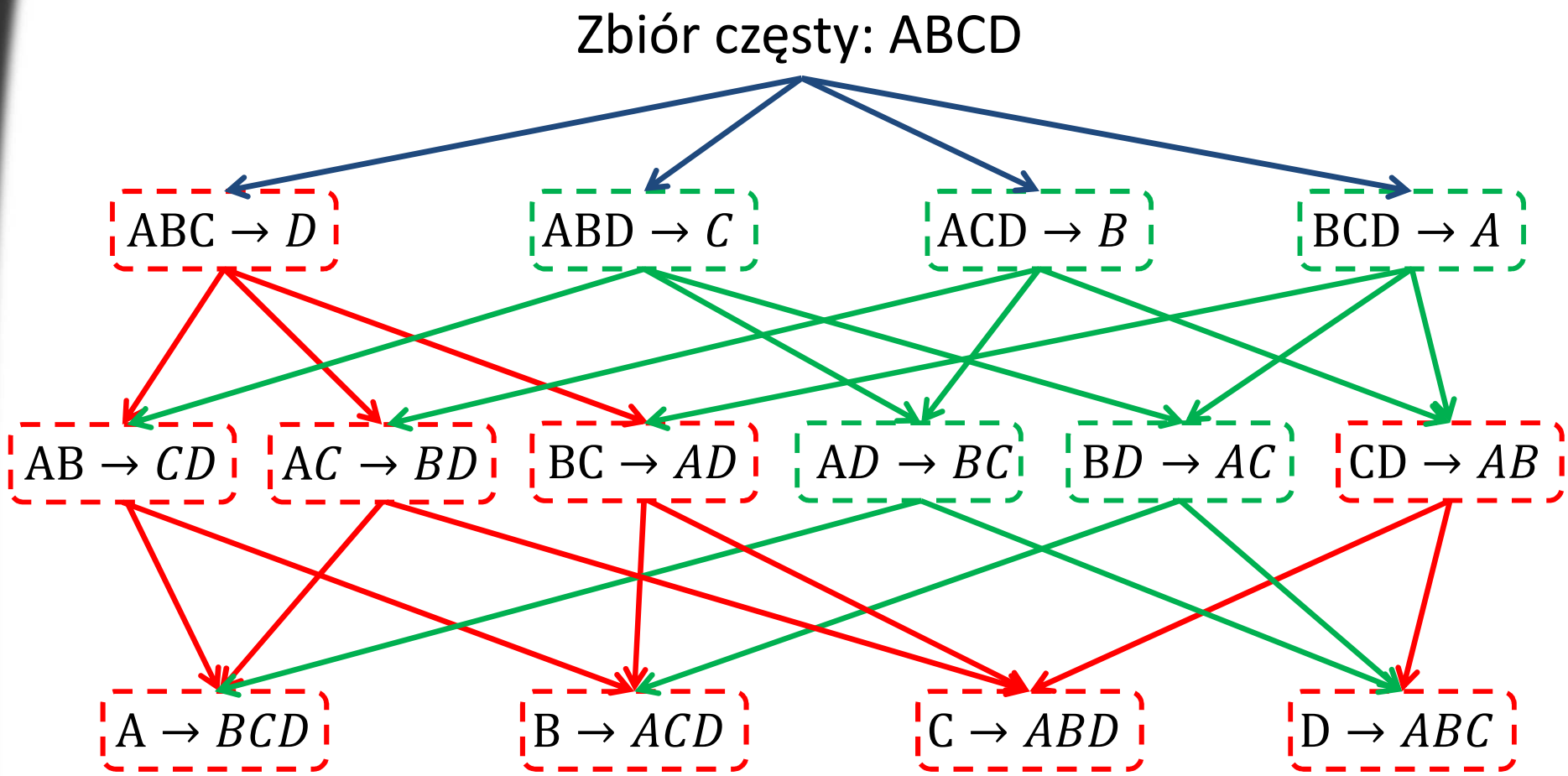
Eliminacja na kracie



Eliminacja na kracie



Eliminacja na kracie

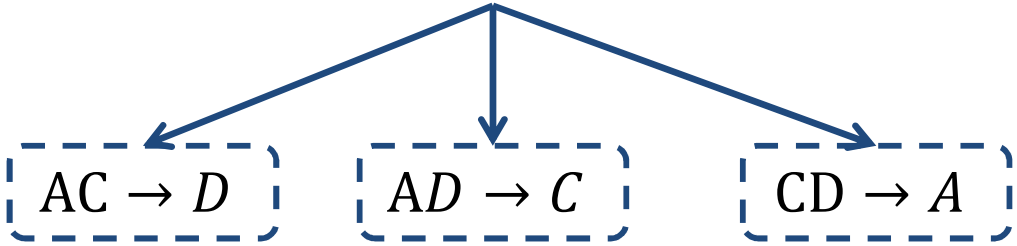


Przykład (1/3)

Zbiór częsty	Wsparcie
A	0.8
B	0.6
C	0.4
D	0.4
AB	0.4
AC	0.4
AD	0.4
CD	0.4
ACD	0.4

$$minconf = 0.6$$

Zbiór częsty: ACD



$$\text{conf}(\text{AC} \rightarrow D) = \frac{\sup(ACD)}{\sup(AC)} = \frac{0.4}{0.4} = 1$$

$$\text{conf}(\text{A}D \rightarrow C) = \frac{\sup(ACD)}{\sup(AD)} = \frac{0.4}{0.4} = 1$$

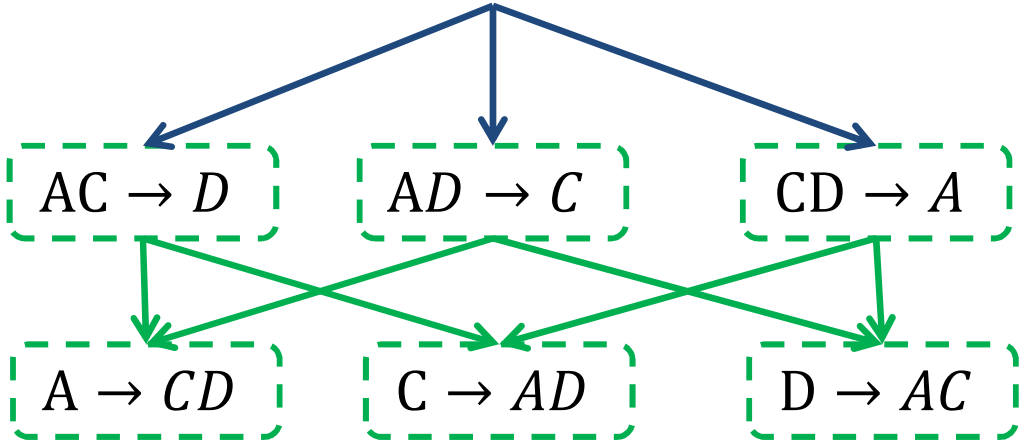
$$\text{conf}(\text{CD} \rightarrow A) = \frac{\sup(ACD)}{\sup(CD)} = \frac{0.4}{0.4} = 1$$

Przykład (2/3)

Zbiór częsty	Wsparcie
A	0.8
B	0.6
C	0.4
D	0.4
AB	0.4
AC	0.4
AD	0.4
CD	0.4
ACD	0.4

$$minconf = 0.6$$

Zbiór częsty: ACD



$$\text{conf}(\text{A} \rightarrow CD) = \frac{\sup(ACD)}{\sup(A)} = \frac{0.4}{0.8} = 0.5$$

$$\text{conf}(\text{C} \rightarrow AD) = \frac{\sup(ACD)}{\sup(C)} = \frac{0.4}{0.4} = 1$$

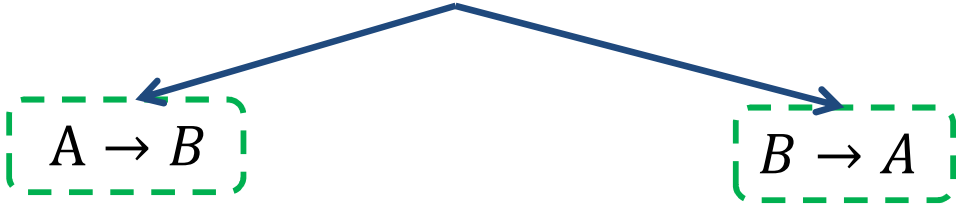
$$\text{conf}(\text{D} \rightarrow AC) = \frac{\sup(ACD)}{\sup(D)} = \frac{0.4}{0.4} = 1$$

Przykład (3/3)

Zbiór częsty	Wsparcie
A	0.8
B	0.6
C	0.4
D	0.4
AB	0.4
AC	0.4
AD	0.4
CD	0.4
ACD	0.4

$$minconf = 0.6$$

Zbiór częsty: AB



$$\text{conf}(\text{A} \rightarrow B) = \frac{\sup(AB)}{\sup(A)} = \frac{0.4}{0.8} = 0.5$$

$$\text{conf}(\text{B} \rightarrow A) = \frac{\sup(AB)}{\sup(B)} = \frac{0.4}{0.6} = 0.6(6)$$

Wielowymiarowe reguły asocjacyjne

- Regułę asocjacyjną nazywamy wielowymiarową, jeżeli dane występujące w regule reprezentują różne dziedziny wartości (różne wymiary analizy)
- Reguły wielowymiarowe określają współwystępowanie wartości danych ciągłych i/lub kategorycznych

Przykład

tid	wiek	dochód	stan-cywilny	płeć	partia
100	44	30 000	żonaty	męska	A
200	55	45 000	żonaty	męska	A
300	45	50 000	panna	żeńska	A
400	34	44 000	panna	żeńska	B
500	45	38 000	żonaty	męska	A
600	32	43 000	kawaler	męska	B
700	32	41 000	kawaler	męska	B
800	33	44 000	panna	żeńska	A

<wiek: 44..55> $\wedge$ <Stan-cywilny: żonaty> $\rightarrow$ <Partia: A>
(sup = 37,5%, conf = 100%)

<Stan-cywilny: panna> $\rightarrow$ <Partia: A>
(sup = 25%, conf = 66,6%)

Binaryzacja danych

- Zamiana atrybutu na zbiór atrybutów binarnych
- Atrybuty ciągłe (2 etapy)
  - Dyskretyzacja, np.: wiek [20-29][30-39]…
  - Zamiana na zbiór atrybutów binarnych:
    - Wiek20-29: 0/1, Wiek30-39: 0/1….
- Atrybuty porządkowe i kategoryczne
  - Każda wartość to osobny atrybut: np.: kolor (red, green,blue)
  - kolorRed: 0/1, kolorGreen: 0/1, kolorBlue: 0/1

Algorytm

1. Transformuj każdy atrybut na postać binarną
2. Nadaj identyfikator (np. numer) każdemu atrybutowi binarnemu
3. Z każdego rekordu utwórz zbiór zawierający identyfikatory dla których rekord zawiera 1.
4. Odkryj reguły asocjacyjne
5. Zastąp identyfikatory w regułach definicjami atrybutów

Przykład (1/4)

tid	wiek	dochód	stan-cywilny	płeć	partia
100	44	30 000	żonaty	męska	A
200	55	45 000	żonaty	męska	A
300	45	50 000	panna	żeńska	A
400	34	44 000	panna	żeńska	B
500	45	38 000	żonaty	męska	A
600	32	43 000	kawaler	męska	B
700	32	41 000	kawaler	męska	B
800	33	44 000	panna	żeńska	A

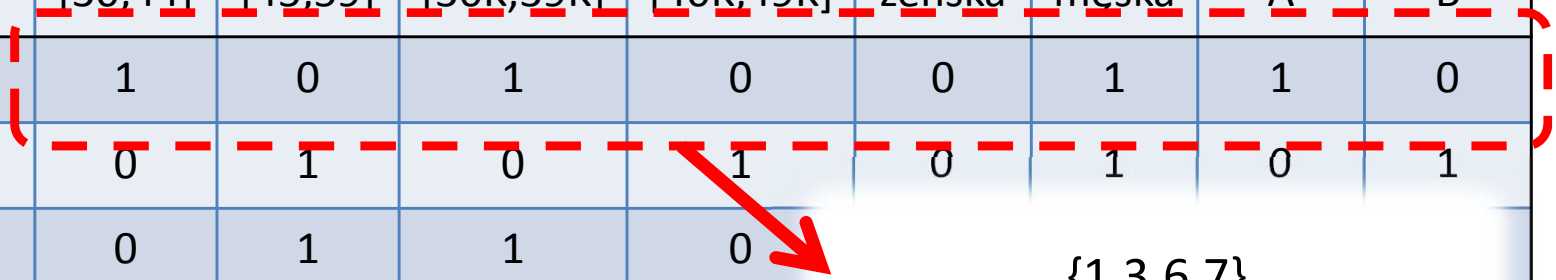
- Dziedzinę atrybutu *wiek* dzielimy na dwa przedziały:

  wiek $\in$ [30,44] oraz wiek $\in$ [45,59]
- Dziedzinę atrybutu *dochód* dzielimy na dwa przedziały:

  dochód$\in$ [30K,39K] oraz dochód$\in$ [40K,49K]

Przykład (2/4)

	1	2	3	4	5	6	7	8
tid	wiek	wiek	dochód	dochód	płeć	płeć	partia	partia
	[30,44]	[45,59]	[30K,39K]	[40K,49K]	żeńska	meska	A	B
100	1	0	1	0	0	1	1	0
200	0	1	0	1	0	1	0	1
300	0	1	1	0				
400	1	0	0	1				
500	0	1	1	0	0	1	1	0
600	1	0	0	1	1	0	0	1
700	0	1	1	0	1	0	0	1
800	1	0	0	1	1	0	1	0
	1	2	3	4	5	6	7	8



Przykład (3/4)

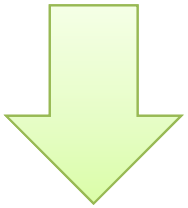
tid	transakcja
100	1, 3, 6, 7
200	2, 4, 6, 8
300	2, 3, 6, 7
400	1, 4, 5, 8
500	2, 3, 6, 7
600	1, 4, 5, 8
700	2, 3, 5, 8
800	1, 4, 5, 7

- Otrzymane reguły:

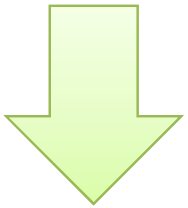
Reguła	Wsparcie	Ufność
$1 \wedge 4 \rightarrow 5$	*sup*= 37,5%	Conf=100%
$1 \wedge 5 \rightarrow 4$	*sup*= 37,5%	Conf=100%
$4 \wedge 5 \rightarrow 1$	*sup*= 37,5%	Conf=100%
$3 \wedge 6 \rightarrow 7$	*sup*= 37,5%	Conf=100%
$3 \wedge 7 \rightarrow 6$	*sup*= 37,5%	Conf=100%
$6 \wedge 7 \rightarrow 3$	*sup*= 37,5%	Conf=100%

Przykład (4/4)

$1 \wedge 4 \rightarrow 5$



	1	2	3	4	5	6	7	8
tid	wiek	wiek	dochód	dochód	płeć	płeć	partia	partia
	[30,44]	[45,59]	[30K,39K]	[40K,49K]	żeńska	męska	A	B
100	1	0	1	0	0	1	1	0



wiek $\in$ [30,44]^dochód $\in$ [40K,49K] $\rightarrow$ płeć=żeńska

Dyskretyzacja atrybutów ilościowych

- Dyskretyzacja może mieć charakter statyczny lub dynamiczny
- Dyskretyzacja statyczna – np. dyskretyzacja atrybutu na przedziały o równej szerokości lub gęstości
- Dyskretyzacja dynamiczna
  - w oparciu o rozkład wartości atrybutu
  - w oparciu o odległości pomiędzy wartościami atrybutu
  - w oparciu o wartości innego atrybutu

Dyskretyzacja atrybutów ilościowych

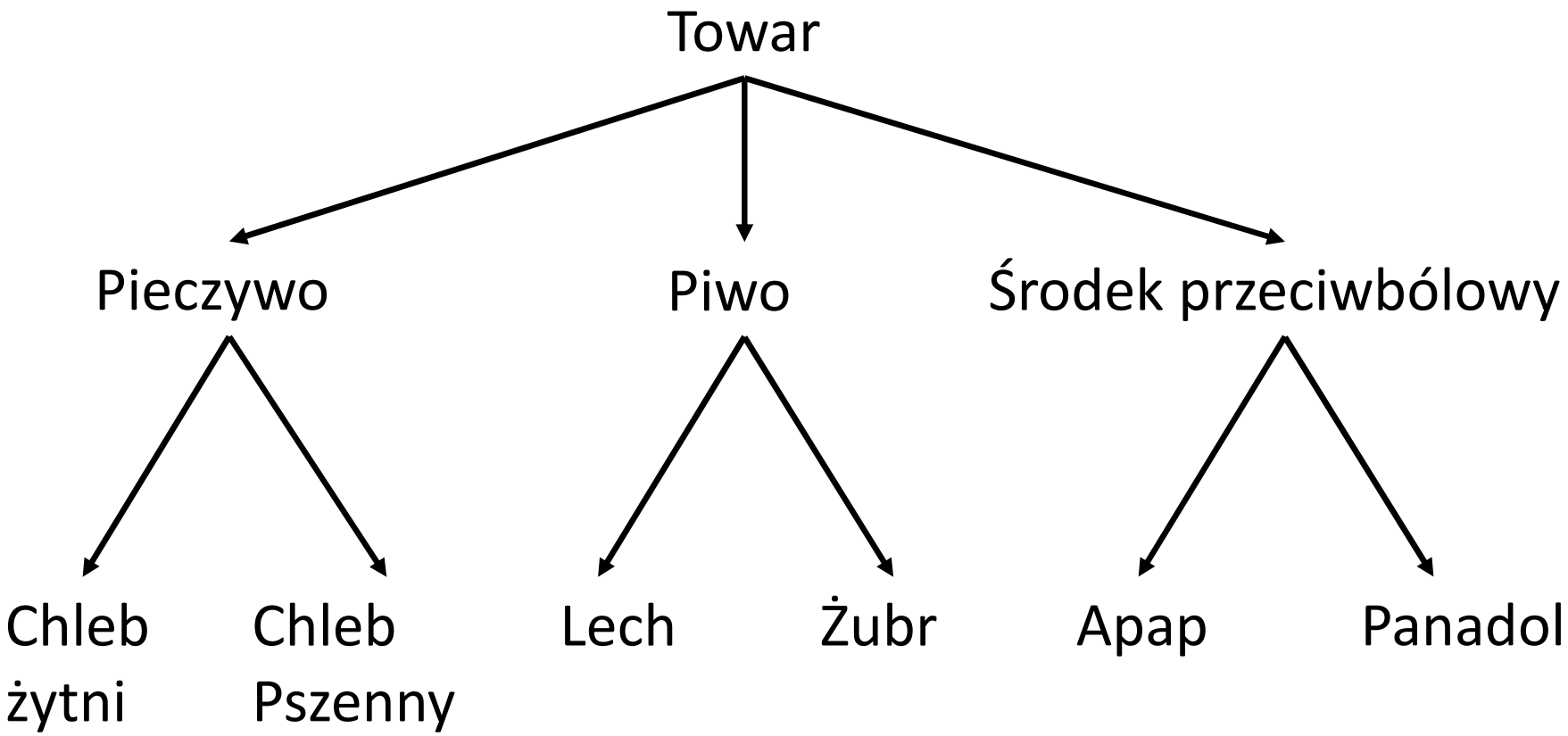
- Przedziały o równej szerokości – rozmiar każdego przedziału jest identyczny (np. przedziały 10K dla atrybutu „zarobek”)
- Przedziały o równej gęstości – każdy przedział posiada zbliżoną (równą) liczbę rekordów przypisanych do przedziału
- Dyskretyzacja poprzez grupowanie (ang. cluster-based) – przedziały odpowiadają skupieniom wartości dyskretyzowanego atrybutu

Wielopoziomowe reguły asocjacyjne

- Czy reguły:
  - {sobieski, żubr} $\rightarrow$ {panadol}
  - {żytnia, lech} $\rightarrow$ {apap}
- …to to samo czy nie?
- Wsparcie konkretnego produktu może być zbyt niskie aby odkryć regułę, ale wsparcie rodzaju produktu już może być wystarczające

Wielopoziomowe reguły asocjacyjne

- Aby rozwiązać problem należy rozbudować algorytm o obsługę hierarchii np.:



Algorytm

1. Uzupełnij każdy zbiór w bazie o poprzedników w hierarchii każdego towaru w zbiorze
2. Odkryj zbiory częste i reguły asocjacyjne standardowym algorytmem
3. Usuń trywialne reguły (wynikające z hierarchii, np. Apap $\rightarrow$ środek przeciwbólowy)

Problem oceny reguł asocjacyjnych

- Reguły o dużym wsparciu niekoniecznie muszą być interesujące
- Reguły o wysokim współczynniku ufności, najczęściej, są dobrze znane użytkownikom
- Tylko użytkownik potrafi ocenić na ile znaleziona reguła jest interesująca

Problem oceny reguł asocjacyjnych

- Miary wsparcia i ufności są niewystarczające do oceny atrakcyjności reguły asocjacyjnej, gdyż:
  - nie uwzględniają aspektu korelacji pomiędzy poprzednikiem i następnikiem reguły asocjacyjnej i tym samym nie dają użytkownikom możliwości rozróżnienia pomiędzy silnymi pozytywnymi regułami asocjacyjnymi pozytywnie i negatywnie skorelowanymi
  - eliminują możliwość znalezienia interesujących reguł asocjacyjnych o niewielkim wsparciu

Przykład

- Wynik badań preferencji klientów kafejki:

	kawa	nie-kawa	suma
herbata	20	5	25
nie-herbata	70	5	75
suma	90	10	100

- Reguła asocjacyjna herbata $\rightarrow$ kawa
  - Sup=20/100=20%
  - Conf=20/25=80%

Przykład

- „Kto lubi herbatę, najczęściej, lubi również kawę”
  - Reguła o dużym wsparciu i wysokiej ufności
- 90% ankietowanych lubi kawę!!!
  - Skąd różnica w wartości wsparcia?
    - Ludzie, którzy lubią herbatę najczęściej nie lubią kawy
    - Istnieje negatywna korelacja pomiędzy preferencją „lubię herbatę” i „lubię kawę”

nie-herbata $\rightarrow$ kawa

(sup = 70%, conf = 70/75=93%)