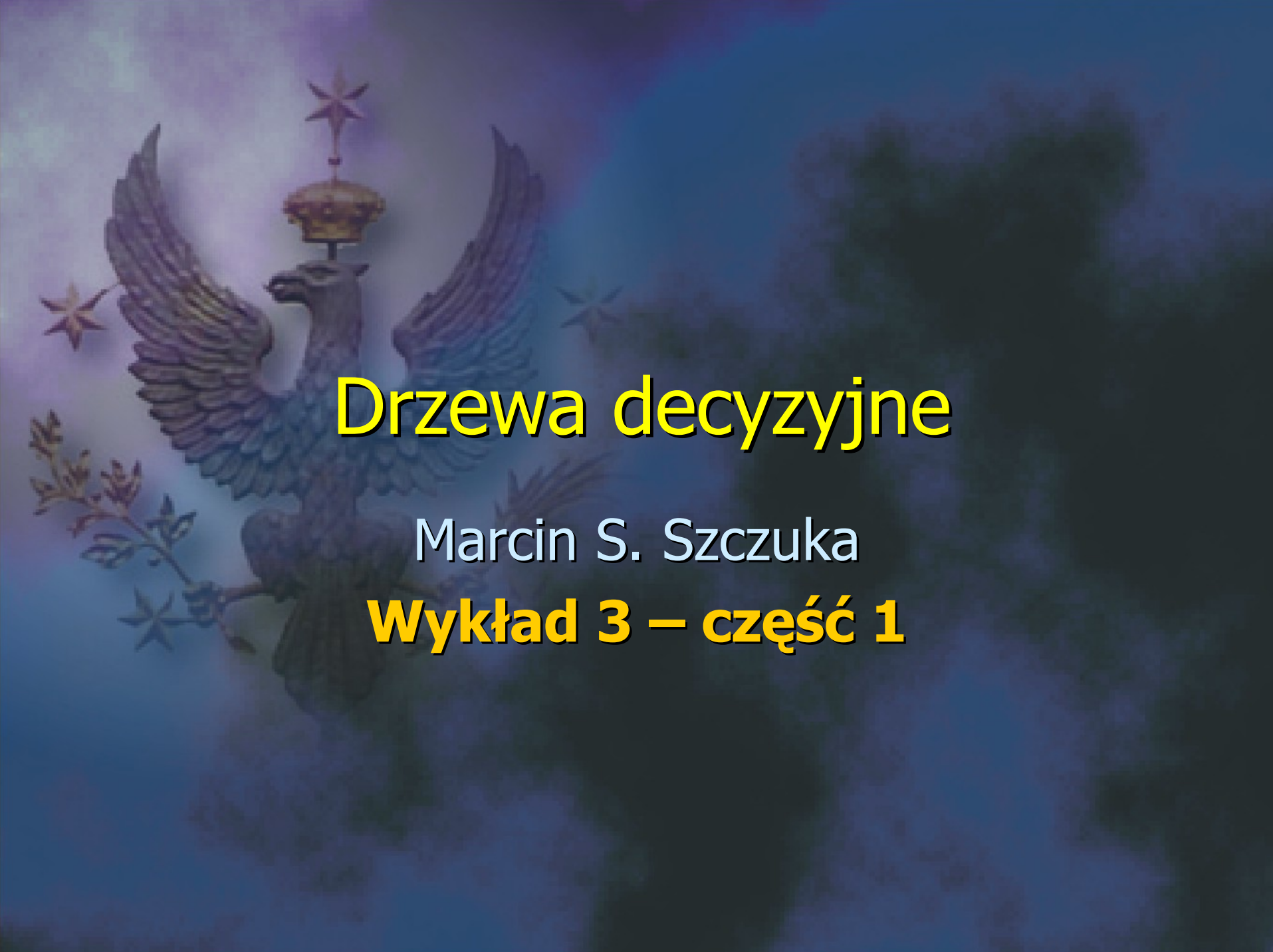


Drzewa decyzyjne

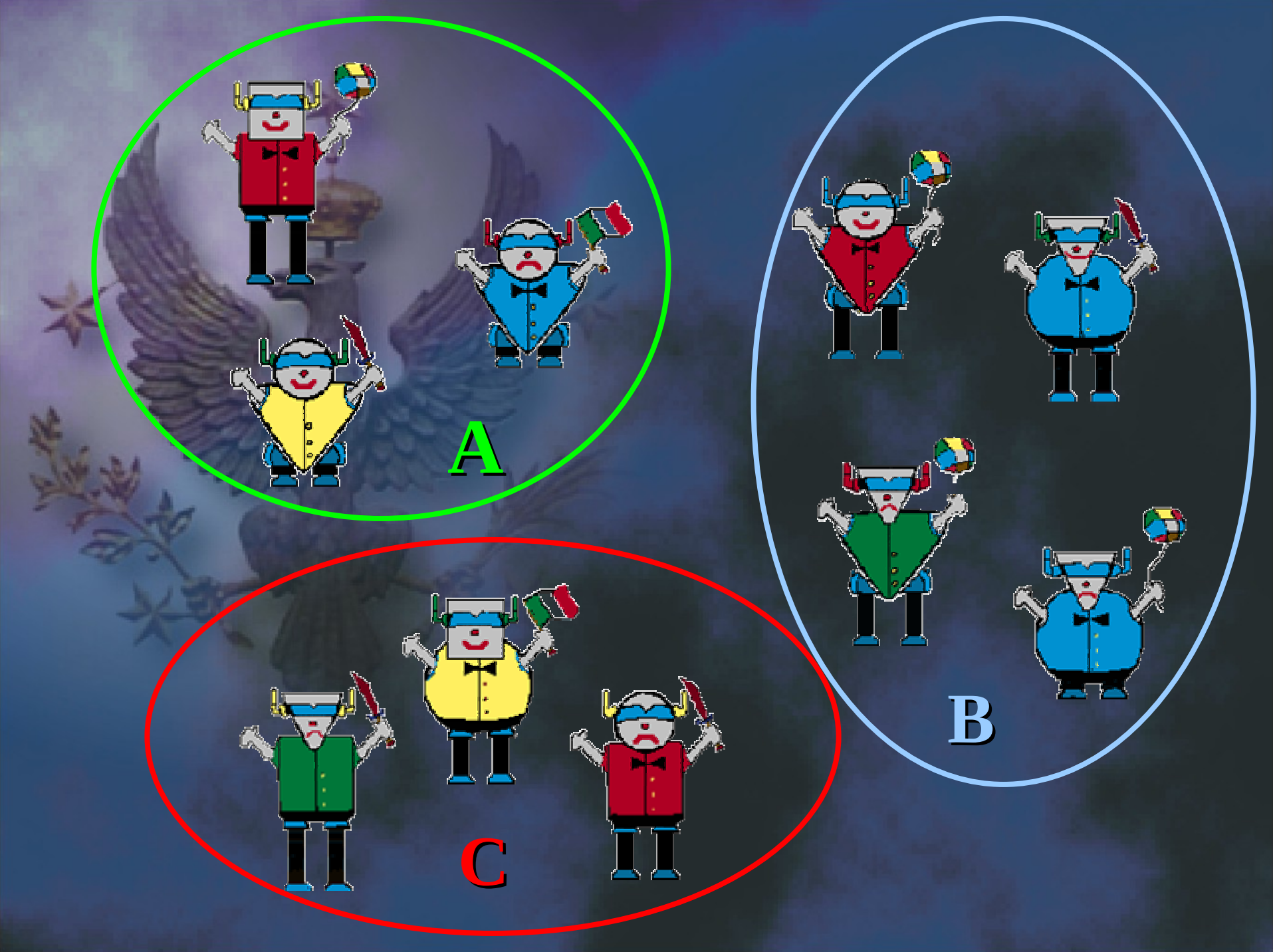
Marcin S. Szczuka

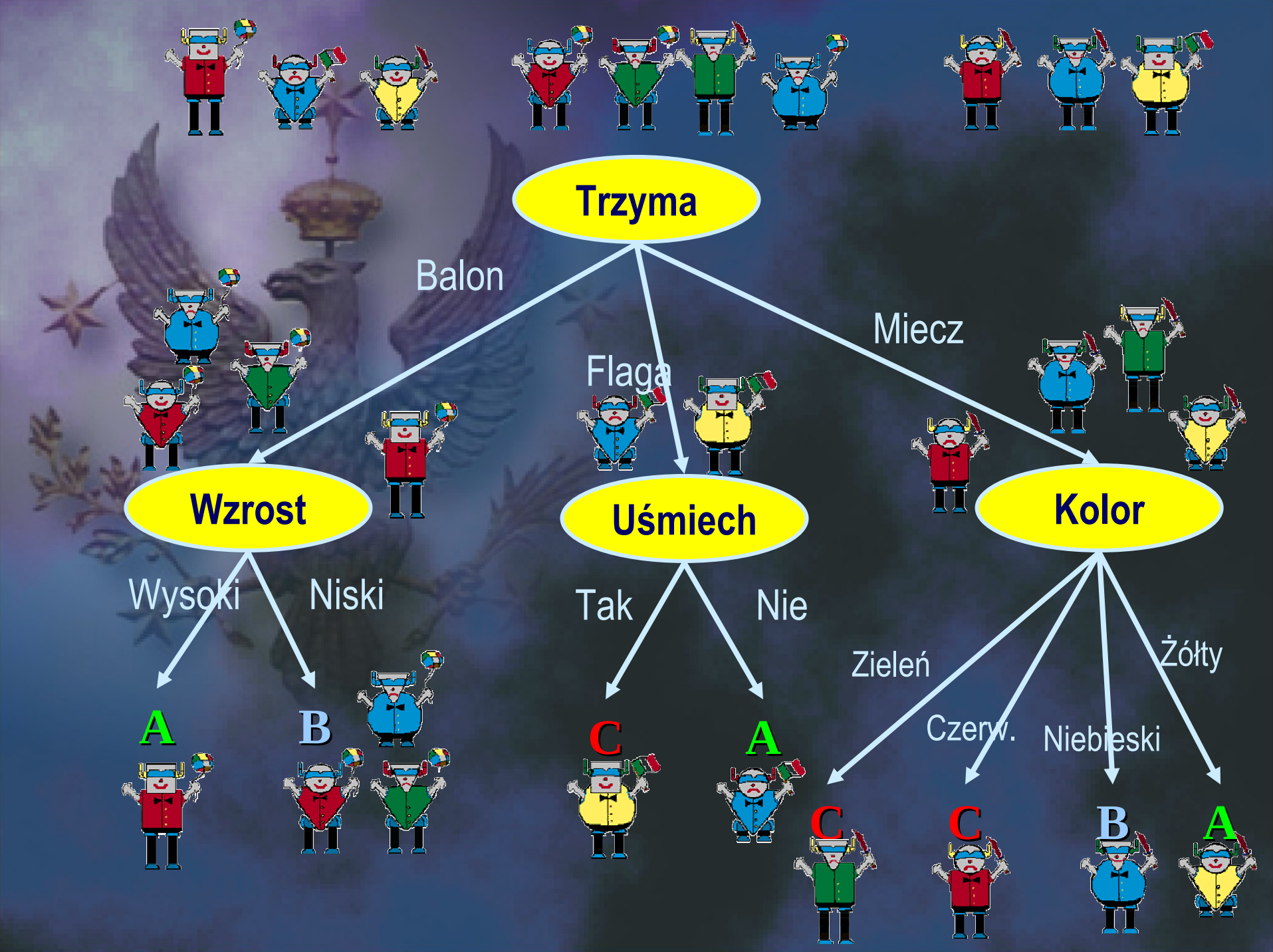
Wykład 3 – część 1

W teorii nie ma różnicy między teorią, a praktyką. W praktyce jest.

Drzewa decyzyjne

- Drzewo – skierowany, acykliczny graf planarny.
- Liść – wierzchołek bez wychodzących krawędzi.
- Wierzchołki wewnętrzne (nie liście) reprezentują testy.
- Każda krawędź odpowiada wynikowi odpowiedniego testu.
- Liście odpowiadają decyzjom (wg. hipotezy).


A
B
C


Trzyma
Balon
Flaga
Miecz
Wzrost
Uśmiech
Kolor
Wysoki
Niski
Tak
Nie
Zieleń
Czerw.
Niebieski
Żółty
A
B
C
A
C
C
B
A

Trzyma

Balon

Flaga

Miecz

Wzrost

Uśmiech

Kolor

Wysoki

Niski

Tak

Nie

Zieleń

Czerw.

Niebieski

Żółty

A

B

C

A

C

C

B

A

Zadanie

Mając zbiór etykietowanych przykładów skonstruuj drzewo, które najlepiej przybliża proces podejmowania decyzji dla tych przykładów.

Notacja

Każdy przykład jest opisany przez zbiór atrybutów:

$$(a_1(x),\ldots,a_n(x))$$

gdzie $x \in X$ i $a_1,\ldots,a_n$ są atrybutami (cechami) takimi że:

$$A_i : X \rightarrow V_i$$

V_i nazywamy przestrzenią (wartości) atrybutu.

Np. atrybut **Kolor** ma przestrzeń

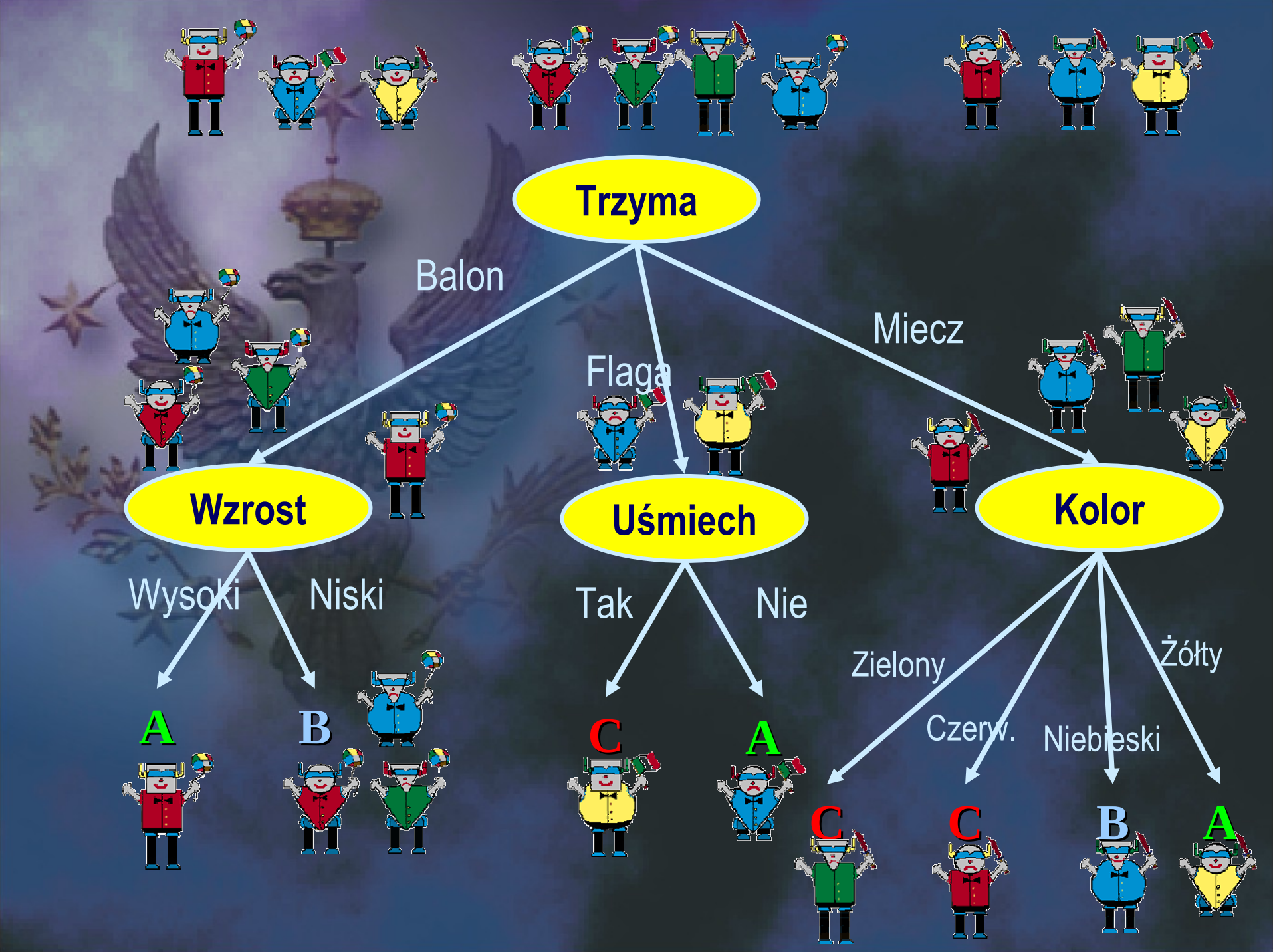
{Czerwony, Zielony, Niebieski, Żółty}

Przykład tablicy decyzyjnej

Trzyma a_1	**Uśmiech** a_2	**Wzrost** a_3	**Kolor** a_4	**Klub** d
Flaga	Nie	Niski	Niebieski	**A**
Miecz	Nie	Wysoki	Zielony	**B**
Flaga	Tak	Wysoki	Żółty	**C**

Budowanie drzewa – ogólna idea

- Podejście „dziel i rządź”.
  - Wybierz „najlepszy” atrybut i ustaw jako test w korzeniu.
  - Stwórz gałąź dla każdej wartości atrybutu. Usuń atrybut z dalszych rozważań.
  - Na końcu każdej gałęzi konstruuj (rekurencyjnie) drzewo z przykładów odpowiadających tej gałęzi.


Trzyma
Balon
Flaga
Miecz
Wzrost
Uśmiech
Kolor
Wysoki
Niski
Tak
Nie
Zielony
Czerw.
Niebieski
Żółty
A
B
C
A
C
C
B
A

ID3 - Notacja

S – próbka treningowa, zbiór przykładów

$S_{a=v}$ – podzbiór przykładów mających v jako wartość na atrybucie a

$d_{\mathbf{l}}$ – decyzja dla liścia $\mathbf{l}$ w drzewie

$a_{\mathbf{n}}$ – test w wierzchołku wewnętrznym drzewa $\mathbf{n}$

A – zbiór dostępnych atrybutów (testów)

dec(S,d) – zwraca najczęstszą decyzję w S lub wartość domyślną d, gdy S pusty.

ID3 - algorytm

ID3(*S,A,d*)
create root;
if (*S pusty*) **or** (*wszystkie przykłady z tą samą decyzja*) **or**
(*A pusty*) **then**
create-leaf l; d_l:=**dec**(*S,d*);
return(*drzewo z pojedynczym liściem* **l**);
endif;
create node n;
*a***n** :=**choose-attribute**(*S,A*); **root:=n;**
d:=**dec**(*S,d*);
forall $v \in V_{an}$ **do**
add-subtree ID3($S_{an=v}$, *A*-{*a***n**}, *d*);
end;
return(*tree*);

ID3 – wybór atrybutu

`choose-attribute(`S,A**`);`**

Zwraca atrybut z A, który prowadzi do najlepszego podziału S tj. atrybut będący najlepszym dyskryminatorem ze względu na decyzję.

ID3 - dyskusja

- Prosty ID3 jest zwykle za prosty. W rzeczywistości korzysta się z jego rozszerzeń np. C4.5 i C5 .
- Złożoność jest rozsądna – $O(|S|n\log n)$.
- Algorytm jest dokładny (kompletny). Każda hipoteza może być skonstruowana.

Drzewa – co dalej

Na następnym wykładzie:

- Wybieranie atrybutu dla drzewa na podstawie miary entropijnej.
- Rozszerzenia i uzupełnienia algorytmu konstrukcji drzewa.
- Więcej o złożoności i rozwiązywaniu problemów związanych z nią.