



# Zespoły modeli – przegląd literatury

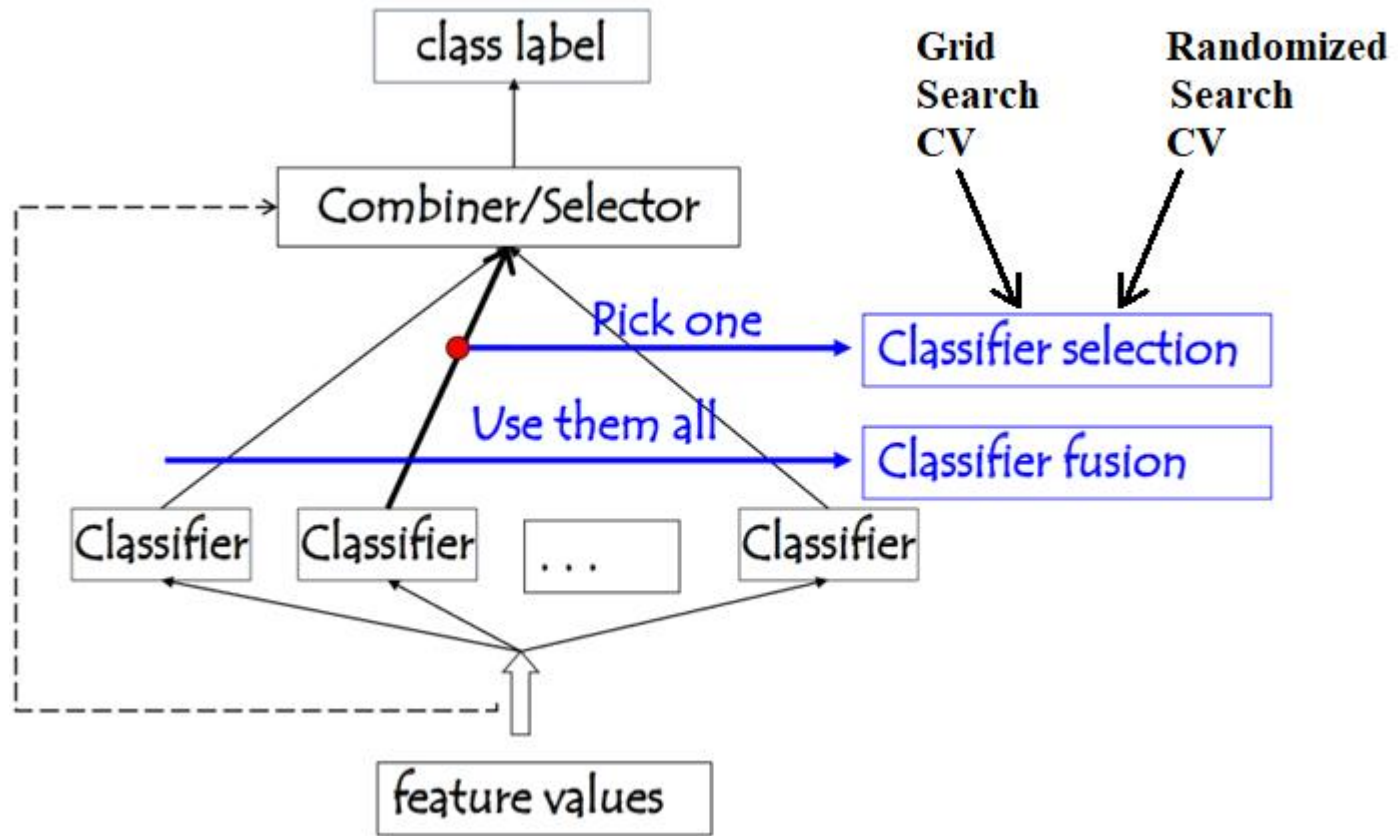
Stanisław Kaźmierczak



# Agenda

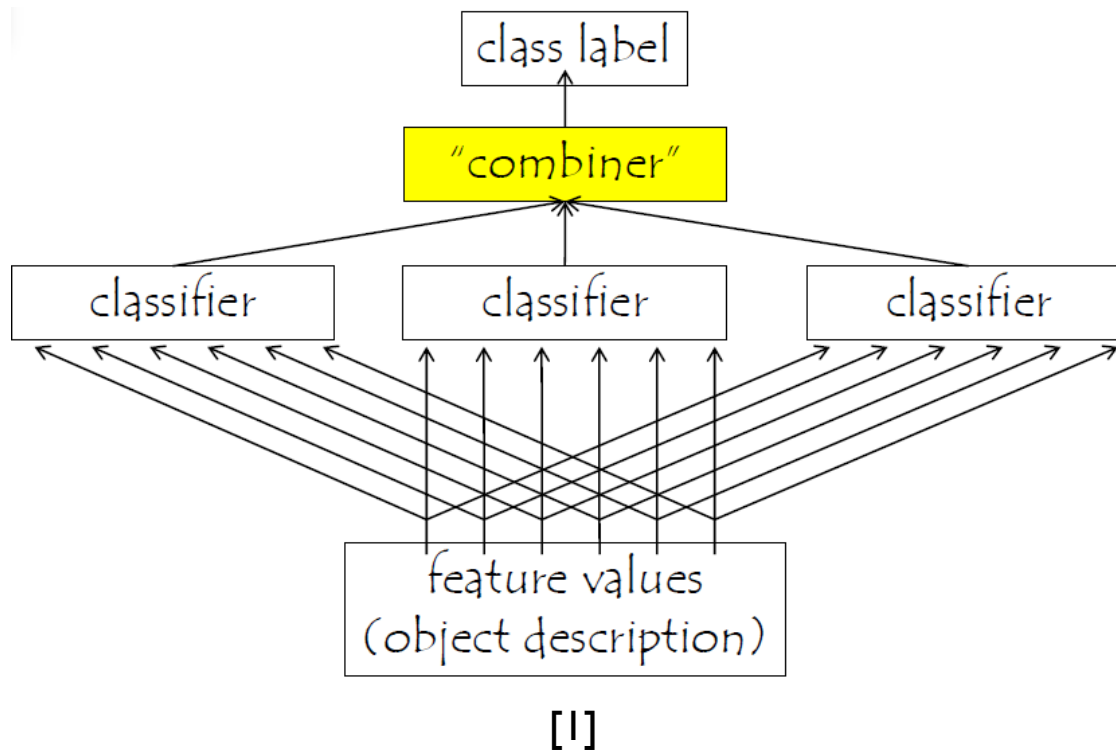
- Motywacja
- Obszary tworzenia zespołów
- Faza wyboru instancji
- Faza selekcji cech
- Faza tworzenia modeli
- Faza łączenia modeli
- Różnorodność

# Łączenie vs selekcja (1)



[1]

# Łączenie vs selekcja (2)



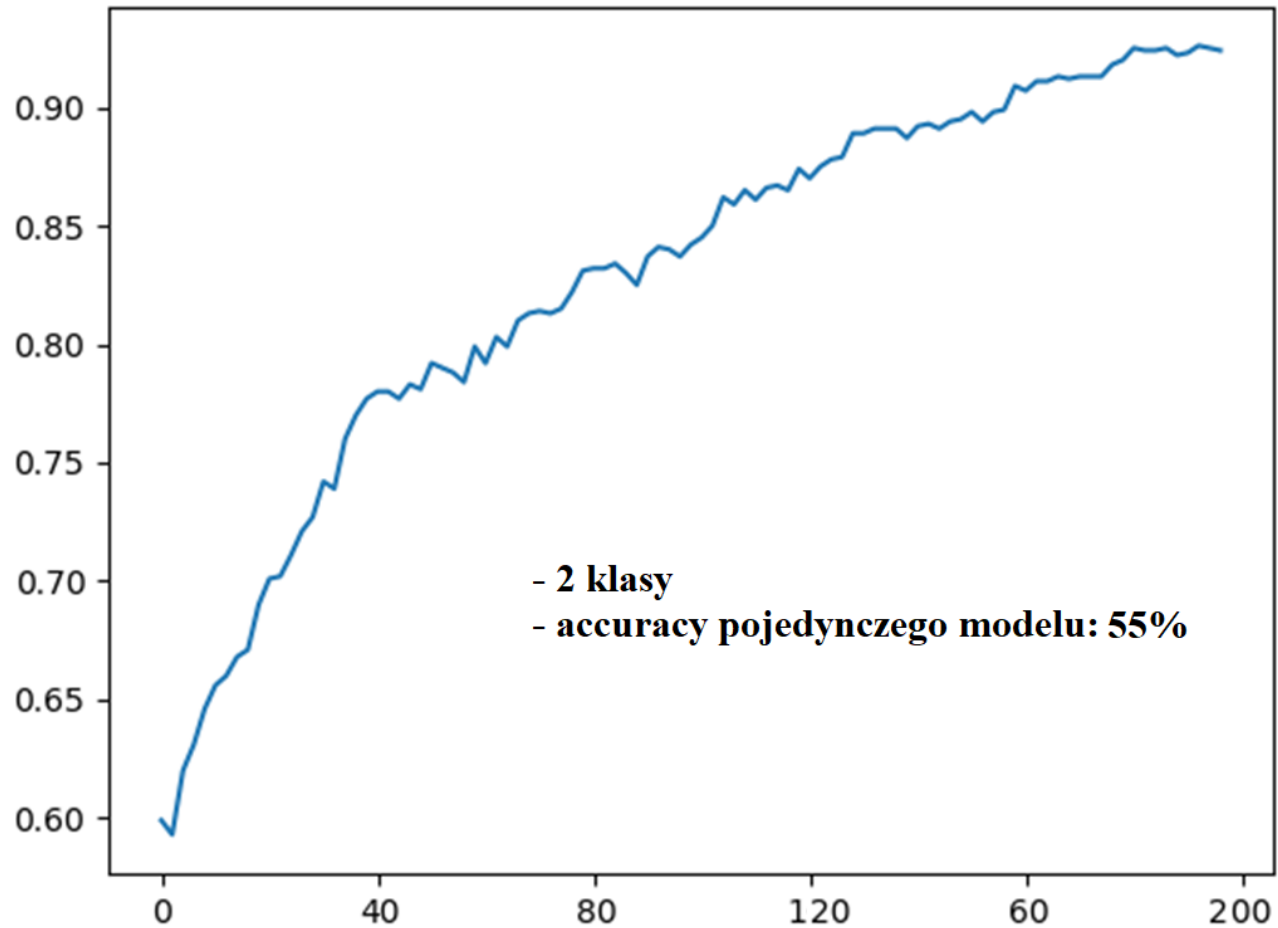
- Łączenie modeli w zespół prawie zawsze daje lepsze wyniki niż najlepszy z pojedynczych modeli
  - Analogia do zespołu ekspertów, konsylium lekarskiego

# Intuicja (1)

Poprawna klasa	Model 1	Model 2	Model 3	Zespół
A	A	B	A	A
A	A	A	B	A
B	A	B	B	B
A	B	B	B	B
B	B	B	A	B
B	A	B	B	B
B	B	A	B	B
A	B	A	A	A
B	B	B	A	B
A	A	A	A	A
Accuracy	60%	70%	60%	90%

# Intuicja (2)

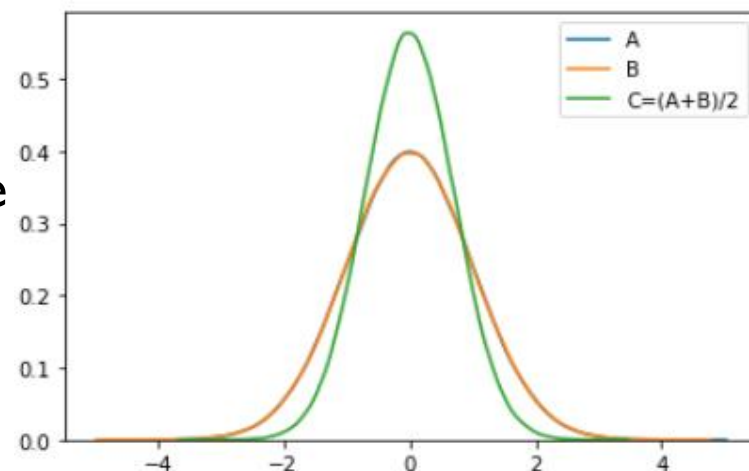
- *Weak learner* – model dający nieco lepszy rezultat niż losowa predykcja



# Wzór Bienaymé

- Wartości zmiennych losowych mogą symulować wielkość błędów, które popełniają modele
- Przy założeniu, że wartość oczekiwana zmiennej losowej wynosi 0, wartość błędu średniokwadratowego (MSE) jest równa wariancji błędu (wariancji zmiennej losowej)
- Zmniejszenie wariancji jest tożsame ze spadkiem MSE
- Dla zbioru nieskorelowanych zmiennych losowych  $X_i$  o tym samym rozkładzie zachodzi:

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$



# Modele w pełni skorelowane

Poprawna klasa	Model 1	Model 2	Model 3	Zespół
A	A	A	A	A
A	A	A	A	A
B	A	A	A	A
A	B	B	B	B
B	A	A	A	A
B	B	B	B	B
B	B	B	B	B
A	B	B	B	B
B	B	B	B	B
A	A	A	A	A
Accuracy	60%	60%	60%	60%



# Modele częściowo skorelowane

- Rzeczywiste modele są z reguły dość mocno (ale nie w pełni) skorelowane
  - Trudne instancje są trudne dla (prawie) wszystkich modeli
  - Podobnie wygląda sytuacja dla prostych instancji

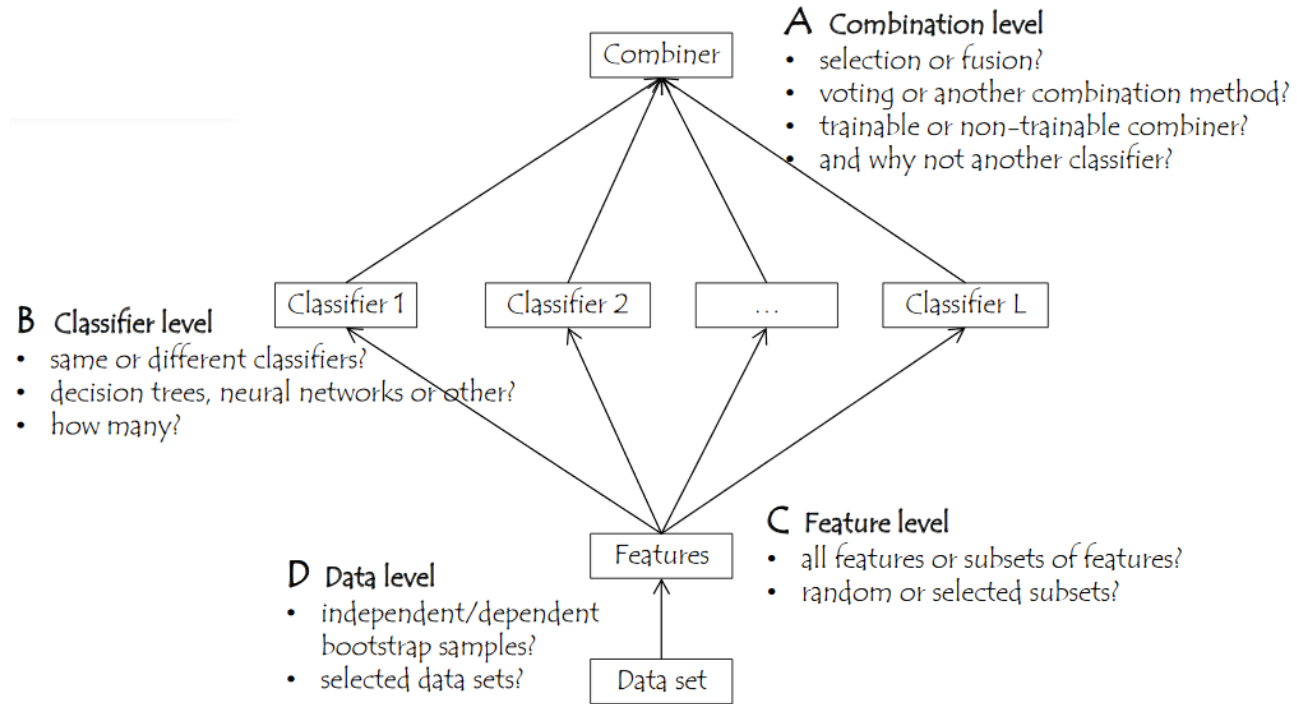
Poprawna klasa	Model 1	Model 2	Model 3	Zespół
A	A	A	A	A
A	A	A	A	A
B	A	A	A	A
A	B	B	B	B
B	B	B	A	B
B	A	B	B	B
B	B	A	B	B
A	B	B	B	B
B	B	B	B	B
A	A	A	A	A
Accuracy	60%	60%	60%	70%



# Jakość zespołu

- O sile zespołu modeli decydują dwa główne czynniki
  - Zróżnicowanie modeli składowych
  - Jakość modeli składowych
- *Tradeoff* pomiędzy zróżnicowaniem i jakością pojedynczych modeli
- W przypadku *weak learners* zespoły osiągają większe zróżnicowanie modeli składowych, ale zespół musi być liczniejszy

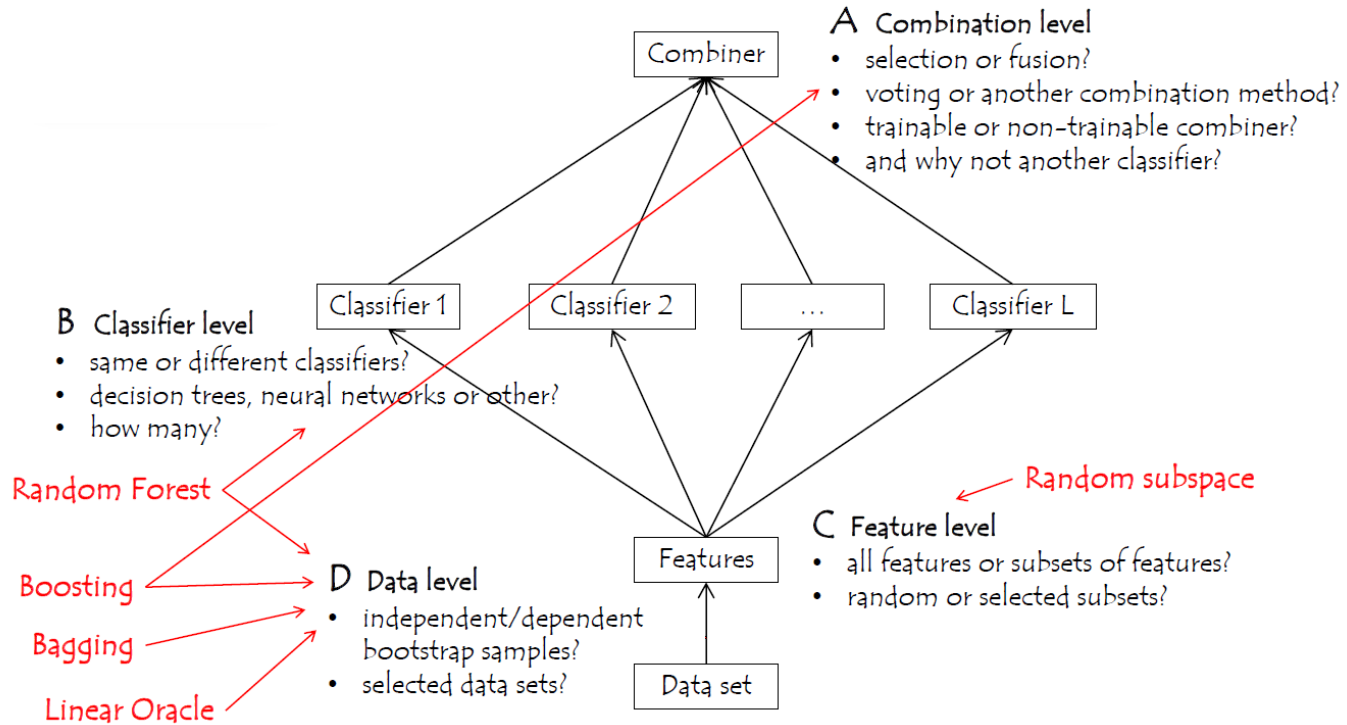
# Budowanie modelu (1)



[1]

- Większość technik z różnych obszarów można stosować niezależnie od siebie

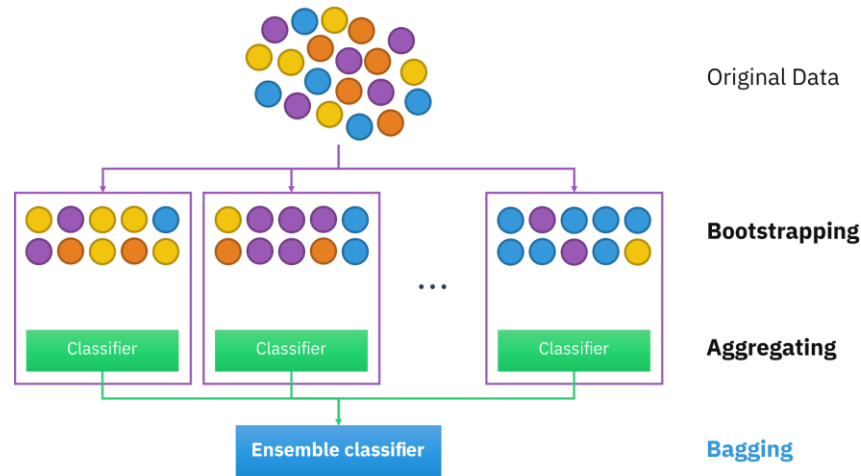
# Budowanie modelu (2)



[1]

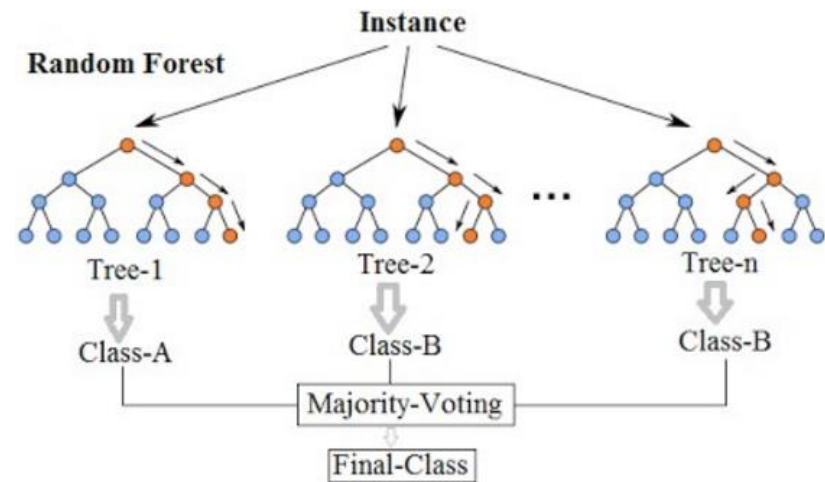
# Bagging

- *Bagging* = **bootstrap aggregating**
- Budowa zespołu bazuje na utworzeniu grupy modeli na różnych zestawach danych wygenerowanych z oryginalnego zbioru uczącego
- Każdy z zestawów zawiera tę samą liczbę rekordów i powstaje w wyniku losowania ze zwracaniem z oryginalnego zbioru uczącego
- Pojedynczy zestaw danych zawiera zdublowane rekordy, jak również nie zawiera pewnych rekordów z oryginalnego zbioru



# Lasy losowe (1)

- Zespół modeli budowany wg następujących założeń:
  - Modelami składowymi są drzewa decyzyjne
  - Dane treningowe służące do skonstruowania kolejnych drzew losowane są ze zwracaniem ze zbioru uczącego (*bootstrapping*)
  - Drzewa są jedynie konstruowane, nie jest stosowane ich przycinanie
- Las losowy jest więc formą baggingu, w którym pojedyncze modele są lekko zmodyfikowanymi drzewami decyzyjnymi
- Las losowy może być wykorzystany do oceny znaczenia poszczególnych cech



[3]

# Lasy losowe (2)

- Klasyczny algorytm konstrukcji lasu losowego wygląda następująco:

**Input:**  $D$  - zbiór uczący,  $p$  - liczba zmiennych niezależnych,  $T$  - liczba tworzonych drzew (wielkość zespołu),  $d < p$  - wymiar podprzestrzeni

**Result:**  $L$  - las losowy

**begin**

**for**  $t = 1$  **to**  $T$  **do**

    utwórz zestaw danych (ang. bootstrap sample)  $D_t$  poprzez wylosowanie  $\text{card}(D)$  rekordów ze zbioru  $D$  ze zwracaniem;

    zbuduj drzewo  $M_t$  na podstawie  $D_t$ , nie stosując przycinania;

    W każdym węźle wylosuj  $d$  spośród  $p$  zmiennych i rozważ podziały bazujące na tych  $d$  zmiennych

**end**

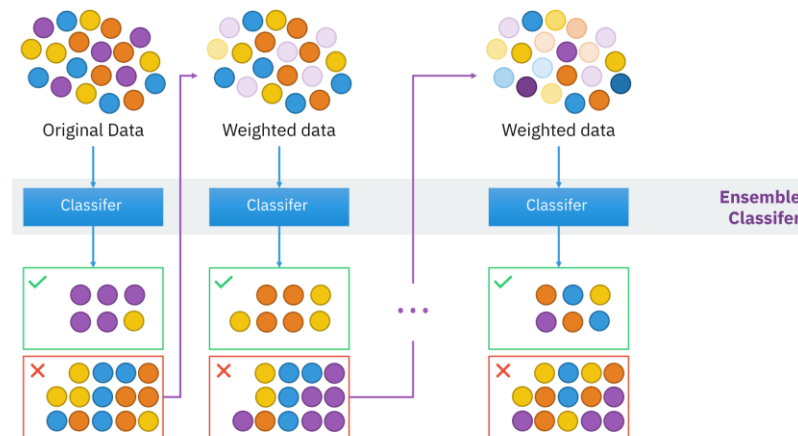
**return**  $\{M_t : t = 1, \dots, T\}$

**end**

[4, 5]

# Boosting

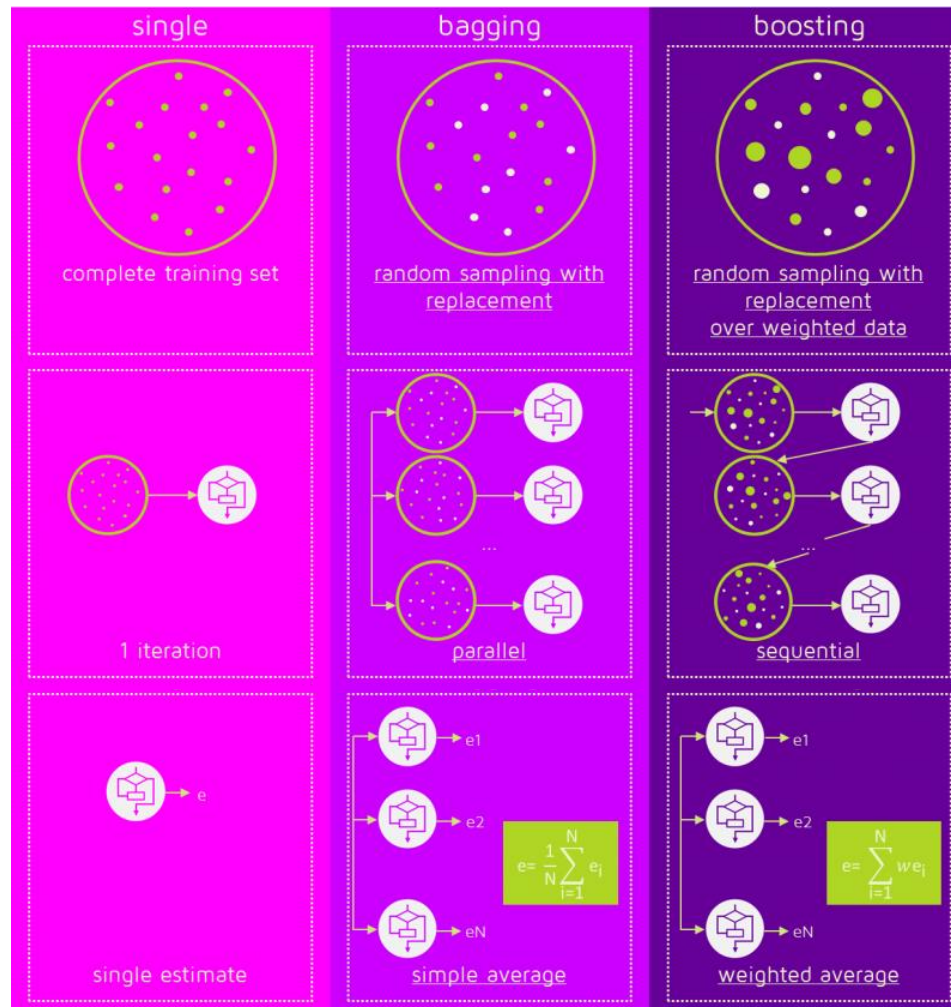
- Zbiory treningowe generowane są poprzez losowanie ze zwracaniem
- Nowy model budowany jest tak, aby lepiej poradzić sobie z instancjami, w których poprzedni model zawiódł
- Ostateczna predykcja podejmowana jest poprzez uśrednienie predykcji (lub głosowanie większościowe) bazowych klasyfikatorów
- Średnia ważona na podstawie jakości predykcji poszczególnych modeli



[6]



# Pojedynczy model/bagging/boosting



[7]



# Selekcja/ekstrakcja cech

- Redukując liczbę wejściowych cech, adresujemy problem wysokowymiarowych i rzadkich danych
  - *Random subspace*
    - podzbiór cech jest losowo wybierany i na ich podstawie konstruowane są klasyfikatory bazowe
    - predykcja z reguły przy użyciu ważonego głosowania, gdzie wagą pojedynczego modelu jest jego *accuracy*
  - *Input decimation*
    - cechy wybierane są w sposób nielosowy
    - algorytm redukuje korelację między błędami klasyfikatorów bazowych poprzez trenowanie ich na różnych podzbiorach cech wejściowych
    - rozszerzenia metody polegające na jednoczesnym sprawdzaniu *accuracy* i różnorodności, jak również zastosowaniu algorytmów ewolucyjnych



# Poziom modeli

- Modele niegeneratywne
  - Opierają się na zbiorze bazowych, dobrze nauczonych, klasyfikatorów/regresorów
  - Nie służą do tworzenia nowych bazowych modeli, ale starają się w odpowiedni sposób połączyć/wyselekcjonować już istniejące
- Modele generatywne
  - Tworzą nowe bazowe modele
  - Na nowopowstający model mają wpływ wcześniej utworzone modele
  - Nowopowstający model tworzony jest zgodnie z ogólną regułą maksymalizacji jakości własnej predykcji oraz różnorodności zespołu



# Łączenie klasyfikatorów (I)

- *Combiner* dostaje na wejściu wyjście z bazowych (pojedynczych) modeli, np. wybraną etykietę (klasę), ranking etykiet lub wektor prawdopodobieństw dla każdej klasy
- Zwraca ostateczną etykietę danej instancji
- Trenowalne i nietrenowalne *combinery*
- Metody
  - *Majority voting*/ważone *Majority voting*
  - Bazujące na regułach Bayesa
  - Wykorzystujące różne operatory algebraiczne (maksimum, minimum, mediana, średnia, itp.)



# Łączenie klasyfikatorów (2)

- Metody (cd.)
  - Bazujące na teorii zbiorów rozmytych
  - *Meta-learning*
    - trenowalne *combinery*
    - output bazowych modeli staje się cechami *combinera*
- Jako wejście *combinera*, oprócz wyjść z pojedynczych modeli, mogą podane być również cechy wejściowe pojedynczych modeli



# Łączenie klasyfikatorów (3)

- „The combiner matters (a lot)” [8]
- „The trained combiner works better” [8]
- Train the combiner if you have *enough* data [9]



# Selekcja modeli (1)

- Selektor wybiera podzbiór zbioru bazowych modeli
- Metody
  - *Test and select*
    - podejście zachłanne; nowy klasyfikator jest dodawany do zespołu tylko wtedy, gdy zmniejsza błąd średniokwadratowy
    - dowolna metoda doboru kolejnego klasyfikatora do zespołu, analizowane były m.in. algorytmy ewolucyjne
  - *Cascading classifiers*
    - bazowe klasyfikatory są dodawane sekwencyjnie
    - jeśli pewność predykcji pierwszego klasyfikatora jest duża, jest uznawana za ostateczną
    - Wpp. decyzja jest przekazywana do kolejnego klasyfikatora, itd.
    - model kaskadowy znajduje szczególne zastosowanie w systemach czasu rzeczywistego, ponieważ decyzja powinna być szybka, podjęta maksymalnie przez kilka klasyfikatorów

# Selekcja modeli (2)

- Metody cd.
  - *Dynamic Classifier Selection*
    - kompetencje każdego klasyfikatora bazowego są oceniane na bieżąco
    - wybierane są te, które najbardziej nadają się do danego wejścia
    - funkcja oceniająca klasyfikatory priorytetyzuje te, które dobrze radziły sobie z przypadkami, które znajdują się blisko w przestrzeni cech z testowanym egzemplarzem
  - *Metody klasteringowe*
    - podejście dynamiczne może być w niektórych przypadkach zbyt kosztowne obliczeniowo
    - klastry tworzą modele bazowe, które dawały podobne rezultaty
    - z każdego klastra wybierany jest klasyfikator(y), co zapewnia odpowiednie zróżnicowanie zespołu



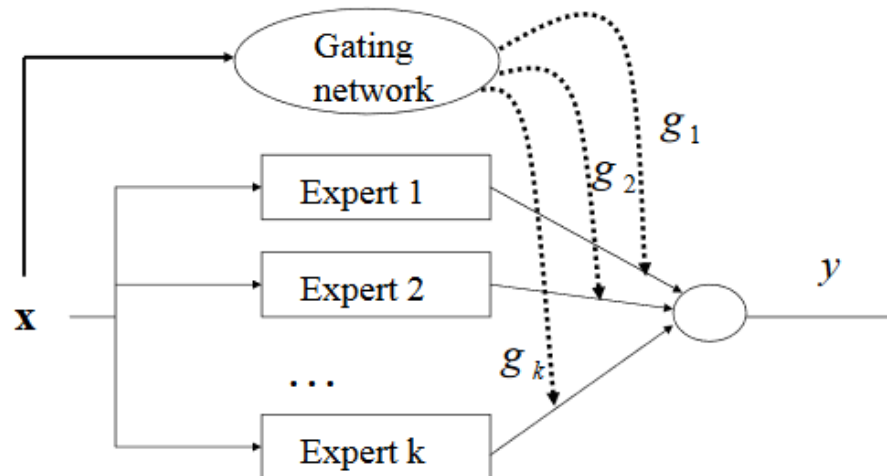
# Selekcja modeli (3)

- Metody cd.
  - *Orientation ordering*
    - algorytmy zachłanne, które preferują dołożyć do zespołu te modele, które skorygują dotychczasowe błędne predykcje zespołu
  - Procedury statystyczne
    - wybierają z bazy modele z wysokim *accuracy*, a ostateczna predykcja dokonywana jest przy użyciu *majority voting*
  - Bazujące na metodach selekcji cech
    - Np. *Forward selection/backward elimination* dodaje/usuwa klasyfikatory tak, aby minimalizować określoną funkcję celu

# Mixture of experts

- Przestrzeni cech dzielona jest na jednolite regiony
- Zespół bazowych klasyfikatorów jest nadzorowany przez tzw. *gating network*
- *Gating network* decyduje, który klasyfikator (lub w jakim stopniu) ma być użyty do którego regionu
- W procesie nauki uczone są zarówno modele bazowe jak i *gating network*

$g_1, g_2, \dots, g_k$  - gating functions



# Różnorodność (1)

## Diversity

Table 2 [5] lists definitions of 76 binary similarity and distance measures used over the last century where  $S$  and  $D$  are similarity and distance measures, respectively.

Table 2 Definitions of Measures for binary data

$S_{JACCARD}$	$\frac{a}{a+b+c}$	(1)
$S_{SIM}$	$\frac{2a}{2a+b+c}$	(2)
$S_{SOBEL}$	$\frac{2a}{2a+b+c}$	(3)
$S_{SOBEL}$	$\frac{2a}{2a+b+c}$	(4)
$S_{SMALL}$	$\frac{2a}{(a+b)+(a+c)}$	(5)
$S_{SILVERMAN}$	$\frac{a}{a+2b+2c}$	
$S_{SILVERMAN}$	$\frac{a+d}{a+b+c+d}$	
$S_{SILVERMAN}$	$\frac{2(a+d)}{2a+b+c+2d}$	
$S_{SILVERMAN}$	$\frac{a+d}{a+2(b+c)+d}$	
$S_{SILVERMAN}$	$\frac{a+0.5d}{a+b+c+d}$	
$S_{SILVERMAN}$	$\frac{a+d}{a+0.5(b+c)+d}$	
$S_{SILVERMAN}$	$a$	
$S_{SILVERMAN}$	$a+d$	
$S_{SILVERMAN}$	$\frac{a}{a+b+c+d}$	
$D_{HAMMING}$	$b+c$	
$D_{EUCLID}$	$\sqrt{b+c}$	
$D_{SQUARED EUCLID}$	$\sqrt{(b+c)^2}$	
$D_{COSINE}$	$(b+c)^2$	
$D_{MANHATTAN}$	$b+c$	
$D_{MANHATTAN}$	$\frac{b+c}{a+b+c+d}$	(20)
$D_{MANHATTAN}$	$b+c$	(21)
$D_{MANHATTAN}$	$(b+c)^2$	(22)

$D_{MAN}$	$\frac{(b+c)}{4(a+b+c+d)}$	(23)
$D_{MAN}$	$\frac{(b+c)^2}{(a+b+c+d)^2}$	(24)
$D_{MAN}$	$\frac{n(b+c)-(b+c)^2}{(a+b+c+d)^2}$	(25)
$D_{MAN}$	$\frac{4bc}{(a+b+c+d)^2}$	(26)
$D_{MAN}$	$\frac{b+c}{(2a+b+c)}$	(27)
$D_{MAN}$	$\frac{b+c}{(2a+b+c)}$	(28)
$D_{MAN}$	$2\sqrt{1-\frac{a}{\sqrt{(a+b)(a+c)}}$	(29)
$D_{MAN}$	$2\sqrt{1-\frac{a}{\sqrt{(a+b)(a+c)}}$	(30)
$D_{MAN}$	$\sqrt{\frac{a}{(a+b)(a+c)}}$	(31)

$S_{SILVERMAN}$	$\frac{a}{\sqrt{(a+b)(a+c)}}$	(47)
$S_{SILVERMAN}$	$\frac{na-(a+b)(a+c)}{n(a+b+c+d)-(a+b)(a+c)}$	(48)
$S_{SILVERMAN}$	$\frac{a}{(a+b)(a+c)}$	(49)
$S_{SILVERMAN}$	$\frac{a+d}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	(50)
$S_{SILVERMAN}$	$\chi^2$ where $\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$	(51)
$S_{SILVERMAN}$	$\frac{a}{a+2}$	(52)
$S_{SILVERMAN}$	$\frac{a-d-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	(53)
$S_{SILVERMAN}$	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	(54)
$S_{SILVERMAN}$	$\frac{a}{\sqrt{(a+b)(a+c)}}$	(55)

$S_{SILVERMAN}$	$\frac{a-d^2}{2a}$	(70)
$S_{SILVERMAN}$	$\frac{\sqrt{ad+a}}{\sqrt{ad+a+b+c}}$	(71)
$S_{SILVERMAN}$	$\frac{\sqrt{ad+a-(b+c)}}{\sqrt{ad+a+b+c}}$	(72)
$S_{SILVERMAN}$	$\frac{ab+bc}{ab+2bc+cd}$	(73)
$S_{SILVERMAN}$	$\frac{a(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)}$	(74)
$S_{SILVERMAN}$	$\frac{a}{(a+b)}$	(75)
$S_{SILVERMAN}$	$\frac{a}{(c+d)}$	(76)

$$S_{EYRAUD} = \frac{n^2(na - (a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)} \quad (74)$$

$$S_{TARANTULA} = \frac{a}{(a+b)} = \frac{a(c+d)}{c(a+b)}$$

$$S_{AMPLE} = \frac{\frac{a}{(a+b)}}{\frac{a}{(c+d)}} = \frac{a(c+d)}{c(a+b)} \quad (76)$$

SEVENTY SIX !!!

The inclusion or exclusion of negative matches,  $d$  in the binary similarity measures have been an ongoing issue [1, 12, 15, 16, 17, 18, 26, 27]. The Sokal & Michener, the Roger & Tanimoto, the Faith, the Ochiai II, the Cole, the Gower, Pearson I, and the Stiles etc. are included in the negative match inclusive measures. The Jaccard, the Tanimoto, the Dice & Sorenson, the Kulczynski I, the Ochiai I, the Kulczynski II, the Sokal & Michener etc. are included in the negative match exclusive measures. The Sokal & Michener, the Roger & Tanimoto, the Faith, the Ochiai II, the Cole, the Gower, Pearson I, and the Stiles etc. are included in the negative match inclusive measures. The Jaccard, the Tanimoto, the Dice & Sorenson, the Kulczynski I, the Ochiai I, the Kulczynski II, the Sokal & Michener etc. are included in the negative match exclusive measures.

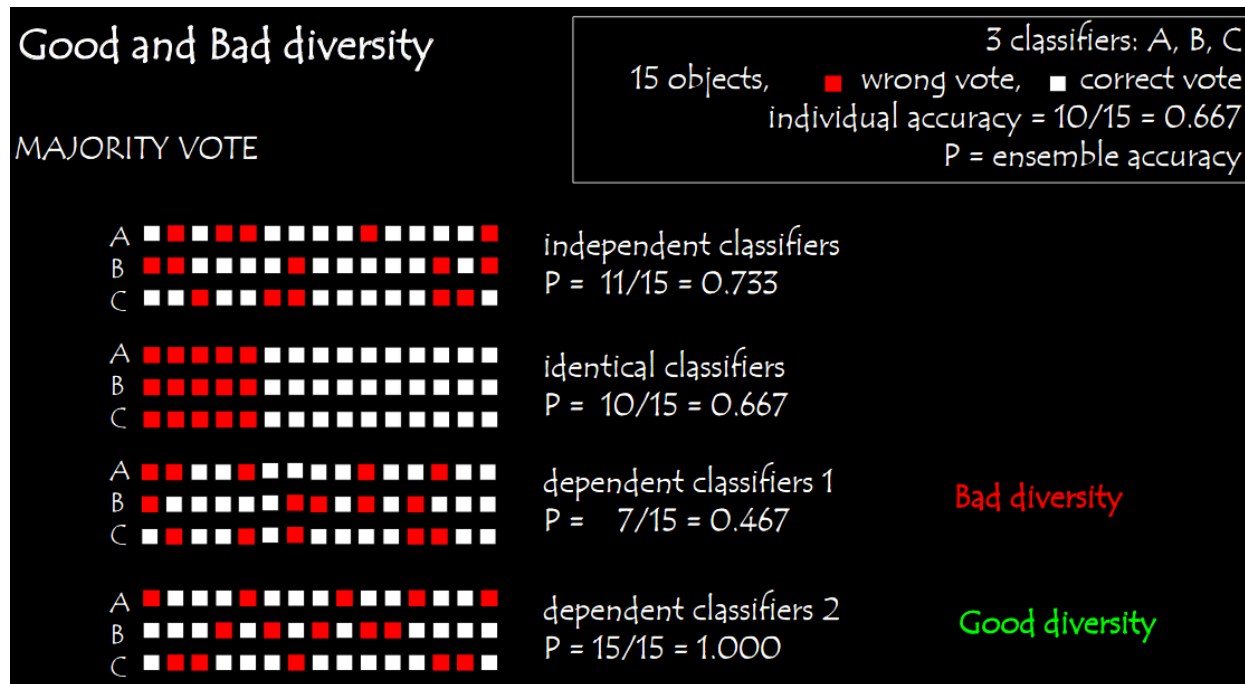
In cases where the two binary states are not equally important, the positive matches are usually more significant than the negative matches [1, 6, 10, 26]. Faith included the negative match but only gave the half credits while giving the full credits for the positive matches in eqn (10) [1]. In [4], different weights for positive and negative matches were studied. Weighted similarity measures such as weighted hamming distance or  $d_{wH}$  [4] are not covered in this paper though.

Historically, all the binary measures observed above had a meaningful performance in their respective fields. The binary similarity coefficients proposed by Peirce, Yule, and Pearson in 1900s contribute to the evolution of the various correlation based binary similarity measures. The Jaccard coefficient proposed at 1901 is still widely used in the various fields such as ecology and biology. The discussion of inclusion or exclusion of negative matches was actively arisen by Sokal & Sneath in during 1960s and by Goodman & Kruskal in 1970s. In Figure 1, the measures are arranged in historical order.

[1, 9]

# Różnorodność (2)

- Nie ma monotonicznej zależności pomiędzy różnorodnością modeli w zespole, a jego *accuracy*
  - A może istnieje taka miara?



[1]



# „Last step: ensemble it”

Best Kagglers

\* Wystarczy losowość na poziomie modelu oraz majority/soft-voting



# Źródła (1)

1. [http://www.icpram.org/Documents/Previous\\_Invited\\_Speakers/2016/ICPRAM2016\\_Kuncheva.pdf](http://www.icpram.org/Documents/Previous_Invited_Speakers/2016/ICPRAM2016_Kuncheva.pdf)
2. [https://en.wikipedia.org/wiki/Bootstrap\\_aggregating](https://en.wikipedia.org/wiki/Bootstrap_aggregating)
3. <https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d>
4. Learning, M. (2012). The Art and Science of Algorithms that Make Sense of Data.
5. M. Grzenda, slajdy z wykładów przedmiotu *Metody Data Science*
6. [https://en.wikipedia.org/wiki/Boosting\\_\(machine\\_learning\)](https://en.wikipedia.org/wiki/Boosting_(machine_learning))
7. <https://quantdare.com/what-is-the-difference-between-bagging-and-boosting/>

## Źródła (2)

8. Duin, R. P. (2002). The combining classifier: to train or not to train?. In Object recognition supported by user interaction for service robots (Vol. 2, pp. 765-770). IEEE.
9. Choi, S. S., Cha, S. H., & Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1), 43-48.
10. <https://people.cs.pitt.edu/~milos/courses/cs2750-Spring04/lectures/class22.pdf>