ANALIZA DANYCH

Wykład 3: Uczenie nienadzorowane



dr inż. Bartosz Czech

Katedra Genetyki · Uniwersytet Przyrodniczy we Wrocławiu

bartosz.czech@upwr.edu.pl

Plan i cel

Plan wykładu

1. Uczenie bez etykiet: zastosowania, ordynacja, dwa główne zadania
2. Redukcja wymiarów (ang. *dimensionality reduction*): PCA (wektory i wartości własne, ładunki, biplot), t-SNE i UMAP (intuicja, parametry, ograniczenia)
3. Miary odległości (ang. *distance measures*; znaczenie, dobór), ordynacja PCoA i NMDS, test PERMANOVA
4. Analiza skupień: k-średnich, metoda hierarchiczna, k-medoidów (enterotypy)
5. Liczba skupień i ocena wyniku bez etykiet
6. Typowe pułapki metod nienadzorowanych

Cel

Umieć **odkryć strukturę** w danych bez znajomości etykiet — zobaczyć, jak próbki układają się względem siebie, i wiedzieć, na ile temu obrazowi ufać.

Uczenie nienadzorowane — przypomnienie (W1)

Uczenie nienadzorowane (*unsupervised learning*)

Uczenie, w którym dostępna jest wyłącznie macierz cech X (ang. *feature matrix*), bez wektora celu **y** (ang. *target*). Celem jest opis struktury danych, a nie przewidywanie etykiety.

Nadzorowane (*supervised*)
dane: X oraz y
uczymy odwzorowania $X \to y$
predykcja, klasyfikacja

znamy odpowiedź,
oceniamy trafność

Nienadzorowane
dane: tylko X
szukamy struktury w X
skupienia, redukcja wymiarów

nie znamy odpowiedzi,
oceniamy spójność

> Ta sama tabela danych bywa punktem wyjścia dla obu podejść. Różnica leży w pytaniu, nie w danych.

Zastosowania analizy nienadzorowanej

Ordynacja (*ordination*)

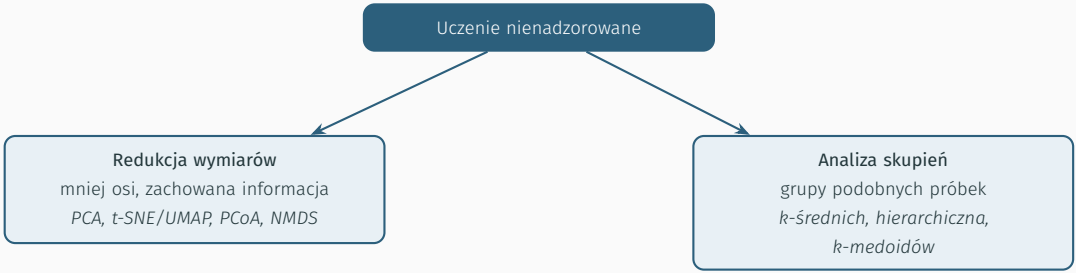
Rozmieszczenie próbek w kilku osiach tak, by odległości na wykresie odzwierciedlały ich niepodobieństwo (ang. *dissimilarity*); termin z ekologii, obejmuje m.in. PCA, PCoA i NMDS.

Analiza nienadzorowana to zwykle **pierwszy** krok po przygotowaniu danych:

- **Eksploracja** — zanim zbudujemy model, chcemy zobaczyć dane w całości.
- **Kontrola jakości** (ang. *quality control*) — próbki odstające (ang. *outliers*) i efekt partii (*batch effect*, W2) są często dobrze widoczne na wykresie ordynacji.
- **Generowanie hipotez** — czy próbki grupują się zgodnie z chorobą, czy raczej według ośrodka albo partii sekwencjonowania?
- **Redukcja** — zastąpienie setek cech kilkoma osiami, łatwiejszymi do wizualizacji i dalszej analizy.

> Jeśli próbki grupują się według **partii**, a nie według choroby, dane wymagają uwagi przed modelowaniem: korekty albo uwzględnienia partii w analizie (z zastrzeżeniami z W2).

Dwa zadania tego wykładu



Te zadania często łączymy: najpierw redukujemy wymiary, a dopiero w zredukowanej przestrzeni szukamy skupień. Tak zrobimy na danych demo.

Redukcja wymiarów

Uzasadnienie redukcji wymiarów

Dane mikrobiomowe mają setki taksonów przy kilkudziesięciu — kilkuset próbkach ($p \gg n$). To rodzi konkretne kłopoty:

- **Nie da się tego zobaczyć** — nie narysujemy wykresu w 300 wymiarach.
- **Cechy bywają skorelowane** — część taksonów niesie podobną informację.
- **Szum** — część zmienności to przypadek, nie biologia.
- **Odległości tracą kontrast** (przekleństwo wymiarowości, ang. *curse of dimensionality*, W1) — dlatego skupienia i ordynację liczy się zwykle w kilku pierwszych osiach.

Pomysł

Zastąpić setki skorelowanych cech kilkoma nowymi osiami, które obejmują najwięcej zmienności — i na nich oglądać oraz analizować dane.

Analiza głównych składowych — intuicja

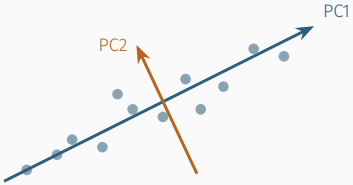
Analiza głównych składowych (*principal component analysis*, PCA)

Metoda redukcji wymiarów, która szuka **kierunków największej zmienności** w chmurze próbek i opisuje dane nowymi, wzajemnie prostopadłymi osiami.

- Pierwsza składowa (PC1) to kierunek, wzdłuż którego próbki są najbardziej rozrzucone.
- Druga składowa (PC2) to kolejny taki kierunek — **prostopadły** do pierwszego.
- I tak dalej: każda następna wyjaśnia najwięcej z **pozostałej** zmienności.

> Nowe osie to **kombinacje** oryginalnych cech, a nie pojedyncze taksony. W wielu danych dwie–trzy osie uchwytują główny obraz; w danych demo nie (PC1 + PC2 = 14%).

Składowe główne



- PC1 biegnie wzdłuż głównego rozciągnięcia chmury (najwięcej wariancji).
- PC2 jest do niej prostopadła i wyjaśnia resztę.
- Współrzędne próbki w nowym układzie to jej **wyniki składowych** (ang. *scores*).

Macierz kowariancji — punkt wyjścia PCA

Macierz kowariancji (*covariance matrix*)

Kwadratowa, symetryczna tabela o tylu wierszach i kolumnach, ile jest cech; na przekątnej stoją wariancje cech, poza przekątną — kowariancje każdej pary cech.

- Kowariancja dodatnia: dwa taksony rosną razem; ujemna: gdy jeden rośnie, drugi maleje; bliska zeru: brak związku liniowego.
- U nas tabela **262 × 262** (cała kohorta), liczona na danych CLR po wycentrowaniu każdej cechy.
- Suma wariancji z przekątnej to **całkowita wariancja** (ang. *total variance*) danych (u nas ok. 5520); PCA jej nie zmienia, tylko rozdziela ją między nowe osie.

Wektory własne i wartości własne

Wektor własny i wartość własna (*eigenvector*, *eigenvalue*)

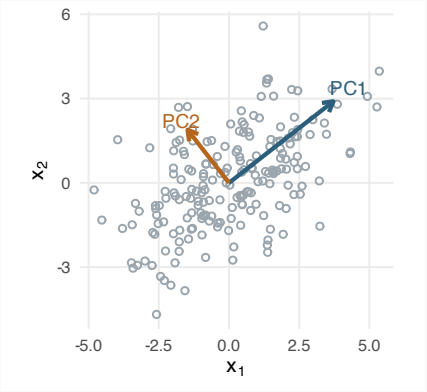
Wektory własne macierzy kowariancji to **kierunki nowych osi** (PC1, PC2, ...); wartość własna mówi, **ile zmienności** danych leży wzdłuż danej osi.

- PC1 to kierunek o największej wartości własnej, PC2 — o drugiej co do wielkości itd.
- Osie są **wzajemnie prostopadłe**, więc nie powtarzają tej samej informacji.
- Procent wariancji wyjaśnionej osi = jej wartość własna / suma wszystkich wartości własnych.

PCA krok po kroku

1. **Centrowanie** (ang. *centering*) każdej cechy; **skalowanie** (ang. *scaling*; podział przez SD) tylko gdy cechy mają różne jednostki lub zakresy — po CLR nie.
2. **Macierz kowariancji** cech — kto zmienia się razem z kim.
3. Przekształcenie macierzy: wyznacza się jej **wektory własne** (kierunki osi) i **wartości własne** (wariancja wzdłuż każdej osi).
4. Osie porządkuje się od największej wartości własnej: PC1, PC2, …
5. Każdą próbkę **rzutuje się** (ang. *projection*) na nowe osie — to jej wyniki składowych.

Rachunek wykonuje program (`prcomp`); nie liczy się go ręcznie.



Dwie cechy: strzałki to kierunki PC1 i PC2 (wektory własne), długość proporcjonalna do odchylenia standardowego wzdłuż osi.

Centrowanie i skalowanie danych

Zarówno PCA, jak i odległości reagują na **jednostki** i **rozrzut** cech. Dlatego dane zwykle przygotowujemy:

- **Centrowanie** — odejmujemy średnią każdej cechy. Kowariancja mierzy zmienność wokół średniej, więc PCA zawsze pracuje na danych wycentrowanych; `prcomp` centruje domyślnie.
- **Skalowanie** — dzielimy przez odchylenie standardowe, gdy cechy są w **różnych** jednostkach; inaczej cecha o dużej skali zdominuje osie.
- **Dane kompozycyjne** (ang. *compositional data*; mikrobiom) — udziały sumują się do 100%: wzrost jednego taksonu obniża udziały pozostałych, co tworzy pozorne korelacje (ang. *spurious correlations*) i zniekształca kowariancję. **CLR** (ang. *centered log-ratio*, W2) zamienia udział na logarytm ilorazu do średniej geometrycznej próbki: zastępuje ograniczenie sumy ilorazami względem średniej geometrycznej; zera trzeba przy tym czymś zastąpić (pseudoliczność).

> Pominięcie tego kroku to częsty błąd: PCA „pokazuje” wtedy po prostu najobfitszy takson, a nie strukturę społeczności. Kolejność jest stała: najpierw CLR, potem PCA.

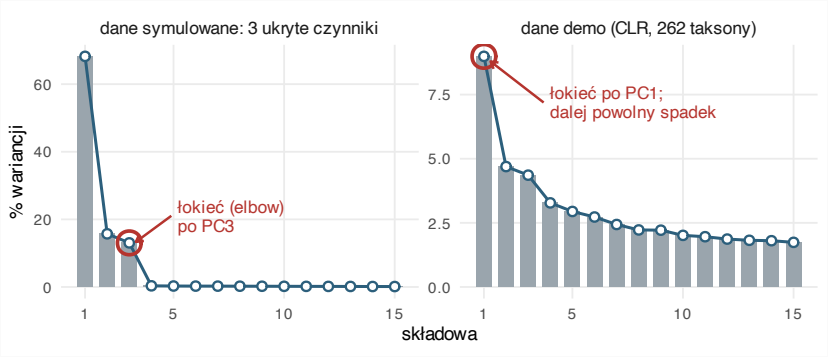
PCA w R i odczyt wydruku

```r
Xf <- X[, colMeans(X > 0) >= 0.10]        # filtr częstości występowania: 114 próbek × 262 taksony
Xclr <- clr(Xf)                           # log-ratio; do udziałów (%) dodana pseudoliczność 1e-6 (W2)
pca <- prcomp(Xclr, center = TRUE, scale. = FALSE)
summary(pca)$importance[, 1:5]
```

```
                            PC1     PC2     PC3     PC4     PC5
Standard deviation     22.2783 16.0959 15.5125 13.4530 12.7483
Proportion of Variance  0.0899  0.0469  0.0436  0.0328  0.0294
Cumulative Proportion   0.0899  0.1368  0.1804  0.2132  0.2427
```

- `Standard deviation` — odchylenie standardowe wyników osi = pierwiastek z wartości własnej (PC1: $22{,}28^2 \approx 496$).
- `Proportion of Variance` — wartość własna / całkowita wariancja ($496/5520 \approx 0{,}09$); `Cumulative Proportion` — suma narastająca.
- `pca$rotation` — ładunki (wektory własne); `pca$x` — wyniki składowych próbek.

Wariancja wyjaśniona i wykres osypiska



- **Wariancja wyjaśniona** (ang. *explained variance*) — jaki procent całkowitej zmienności niesie oś.
- **Wykres osypiska** (ang. *scree plot*) — udziały wariancji kolejnych osi, malejąco.
- **Łokieć** (ang. *elbow*) — miejsce, po którym słupki nagle się wypłaszczają: dalsze osie dokładają niewiele.

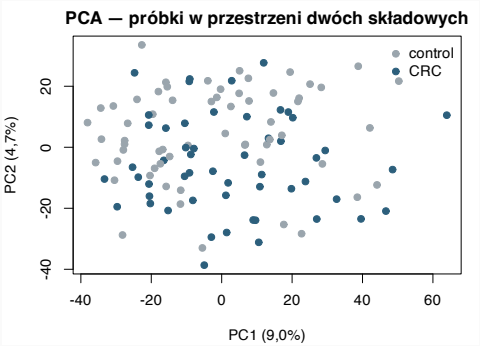
Liczba zachowanych składowych

Nie ma jednej reguły — stosujemy kilka przesłanek naraz:

- **Skumulowany procent wariancji** (ang. *cumulative explained variance*) — ile osi potrzeba, by uchwycić np. 70–80% zmienności. U nas 70% wymaga **37** osi, a 80% — **50**.
- **Łokieć na osypisku** — zatrzymujemy się tam, gdzie słupki się wypłaszczają.
- **Interpretowalność** — oś, której nie potrafimy powiązać z niczym sensownym, nie musi być warta zachowania.

> U nas PC1 i PC2 to łącznie zaledwie **14%** zmienności — zmienność jest rozłożona na wiele osi, co ostrzega przed czytaniem tylko dwóch pierwszych. W tych danych PCA służy głównie **wizualizacji i kontroli jakości**, a nie kompresji.

Wykres rozrzutu składowych



Każdy punkt to **jedna próbka** w układzie dwóch pierwszych składowych. Kolorem nanosimy znaną etykietę — **tylko po to, by spojrzeć**, bo PCA jej nie używa.

Pytamy: czy grupy się rozdzielają? Czy widać wartości odstające albo podejrzane skupiska?

Test t (ang. *t-test*) porównuje różnicę średnich z rozrzutem w grupach; **p-wartość** (ang. *p-value*) to prawdopodobieństwo co najmniej tak dużej różnicy, gdyby grupy w rzeczywistości się nie różniły.

Przeplatanie się grup nie znaczy, że sygnału nie ma — PCA wyjaśnia **największą** zmienność. U nas wzdłuż PC2 próbki CRC leżą średnio niżej (średnie: kontrola 4,8, CRC $-5{,}5$; test t Welcha na wynikach PC2: p $\approx$ 0,0005), choć chmury silnie się nakładają.

Ładunki — wkład cech w oś

Ładunki (*loadings*)

Wagi, z jakimi cechy wchodzą do kombinacji tworzącej składową główną (współrzędne wektora własnego). Wartość bezwzględna ładunku mierzy **wkład** cechy w oś.

- Znak całej składowej jest **umowny** — odwrócenie wszystkich znaków daje tę samą oś; interpretujemy znaki względne (które taksony po przeciwnych stronach).

```
# 4 taksony o największym wpływie na PC1 (co do wartości bezwzględnej)
ord <- order(abs(pca$rotation[, "PC1"]), decreasing = TRUE)[1:4]
data.frame(takson = sub("species:", "", rownames(pca$rotation)[ord]),
           ladunek = round(pca$rotation[ord, "PC1"], 3))
```

```
                        takson ladunek
          Enterocloster bolteae   0.175
              Eggerthella lenta   0.171
 Erysipelatoclostridium ramosum   0.163
         [Ruminococcus] torques  -0.161
```

PC1 a zmienne techniczne

Zanim zinterpretujemy oś biologicznie, sprawdzamy, czy nie odwzorowuje zmiennej **technicznej**. Prosty test: korelacja PC1 z **bogactwem** (ang. *richness*) próbki (liczbą wykrytych taksonów).

```
rich <- rowSums(Xf > 0)            # bogactwo = liczba obecnych taksonów
reads <- d$cd$number_reads         # d = load_demo(); głębokość
round(c(richness = cor(pca$x[, 1], rich),
        log10_reads = cor(pca$x[, 1], log10(reads))), 3)
```

```
   richness log10_reads
     -0.194       0.050
```

Bogactwo: $r^2 \approx 0{,}04$ — związek **słaby** (test korelacji Pearsona, $p \approx 0{,}039$); głębokość (ang. *sequencing depth*): $r^2 < 0{,}01$ — znikomy. PC1 nie jest odbiciem zmiennej technicznej.

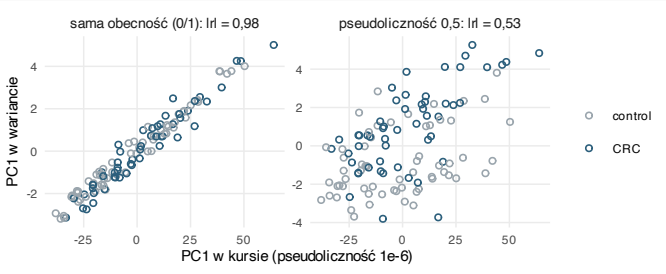
CLR i zera — przypomnienie (W2)

- **Dlaczego transformacja**: udziały sumują się do 100%, więc wzrost jednego taksonu obniża udziały pozostałych i tworzy pozorne korelacje; PCA na surowych udziałach pokazałaby głównie najliczniejsze taksony.
- **CLR**: logarytm udziału podzielonego przez średnią geometryczną udziałów w tej próbce — ile razy takson jest liczniejszy od „przeciętnego" w próbce (dodatnie: więcej, ujemne: mniej).
- **Zera**: logarytm zera nie istnieje, więc przed CLR do udziałów dodaje się małą liczbę — **pseudoliczność**.

	takson 1	takson 2	takson 3	takson 4
udział (%)	50	30	20	0
CLR, pseudoliczność 0,5	1,51	1,00	0,60	$-3{,}11$
CLR, pseudoliczność 10^{-6}	4,79	4,28	3,87	$-12{,}94$

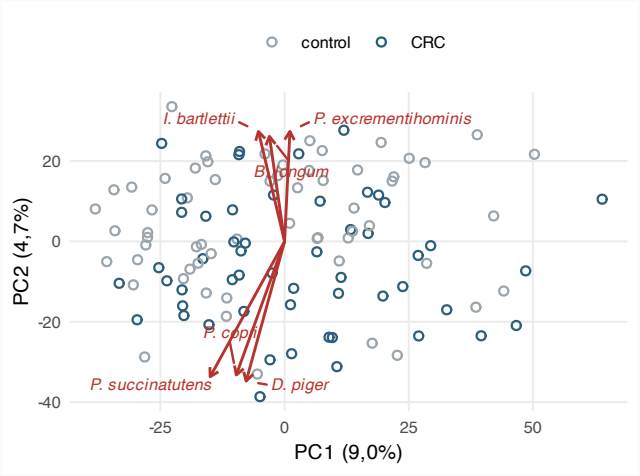
Różnice między taksonami obecnymi się nie zmieniają (1 a 2: 0,51 w obu wierszach); zero ląduje tym dalej, im mniejsza pseudoliczność.

Zera i pseudoliczność a PCA



- W danych demo **58,6%** wartości to zera; w kursie pseudoliczność wynosi 10^{-6} punktu procentowego.
- Przy tak małej pseudoliczności PC1 to prawie dokładnie oś **obecności/nieobecności** taksonów ($|r| = 0{,}98$ z PCA na danych 0/1), a nie gradient udziałów.
- Przy pseudoliczności 0,5 PC1 wygląda inaczej ($|r| = 0{,}53$): wynik PCA **zależy od sposobu obsługi zer**; wniosek z osi sprawdza się przy innej pseudoliczności (analiza wrażliwości, ang. *sensitivity analysis*).

Biplot — próbki i ładunki na jednym wykresie



- **Punkty**: próbki (wyniki na PC1 i PC2), kolor naniesiony po fakcie.
- **Strzałki**: taksony; kierunek pokazuje, w którą stronę takson „ciągnie" próbki, długość — jak silnie.
- Próbki leżące w kierunku strzałki mają w przybliżeniu **więcej** danego taksonu niż przeciętna próbka (tu dwie osie to tylko 14% wariancji).
- Pokazano po 3 taksony o skrajnych ładunkach na **PC2**, bo to wzdłuż PC2 różnią się grupy (PC1 — nie).

Odczyt osi PC2 w danych demo

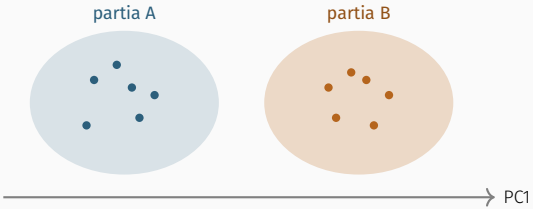
Dlaczego PC2: na tej osi grupy różnią się ($p \approx 0{,}0005$), na PC1 nie ($p \approx 0{,}22$; test t Welcha na wynikach składowej). Pytanie: z jakich taksonów składa się oś związana z chorobą i czy kierunek strzałek zgadza się z danymi?

Takson	ładunek PC2	wykryty w % próbek		średni CLR	
		control	CRC	control	CRC
Desulfovibrio piger	$-0{,}184$	26	45	$-1{,}8$	$0{,}9$
Phascolarctobacterium succinatutens	$-0{,}178$	41	34	$0{,}9$	$-0{,}3$
Prevotella copri	$-0{,}176$	46	53	$1{,}0$	$1{,}9$
Parasutterella excrementihominis	$0{,}145$	61	53	$2{,}3$	$1{,}3$
Intestinibacter bartlettii	$0{,}145$	69	53	$3{,}1$	$0{,}8$
Bifidobacterium longum	$0{,}138$	84	70	$6{,}3$	$4{,}1$

- Pięć z sześciu taksonów zachowuje się zgodnie z kierunkiem strzałki; *P. succinatutens* — odwrotnie, ale jego różnica między grupami jest nieistotna (test Wilcoxona, $p \approx 0{,}59$).
- Oś to **kombinacja** wielu taksonów: ładunek nie jest testem różnicy pojedynczej cechy.
- Zastosowanie: **lista kandydatów** do dalszej analizy i do modelu, nie gotowy wniosek.

Efekt partii w ordynacji

Przypomnienie (W2): efekt partii (*batch effect*) — systematyczna różnica w pomiarach między grupami próbek przetwarzanych w różnych warunkach technicznych.



Gdy próbki rozdzielają się na wykresie PCA wzdłuż partii, a nie choroby, oglądamy artefakt. Dlatego zawsze kolorujemy ordynację zmienną techniczną (partia, data, głębokość), zanim ogłosimy sygnał biologiczny.

Rzutowanie nowych próbek na istniejące osie

PCA to przepis: raz wyznaczone osie można **zastosować** do nowych próbek, nie przeliczając całości. W R służy do tego `predict`.

```
pca <- prcomp(Xtr)                    # osie uczymy TYLKO na treningu
nowe <- predict(pca, newdata = Xte)   # te same osie na nowych próbkach
```

- Nowe próbki trafiają w **tę samą** przestrzeń — można je porównać ze „starymi”.
- To wzorzec, który wróci w uczeniu nadzorowanym: wszystko, co „uczy się” z danych (osie, skala, filtr), wyznaczamy **wyłącznie** na treningu (ang. *training set*).

> Gdy PCA jest krokiem przed modelem predykcyjnym, liczenie go na **wszystkich** danych to **przeciek** (ang. *data leakage*): zbiór testowy (ang. *test set*) współtworzy osie, w których jest potem oceniany. W samej eksploracji (bez oceny modelu) PCA na całości jest poprawne — tak liczyliśmy dotąd.

Metody nieliniowe — t-SNE i UMAP

PCA wyznacza tylko kierunki **liniowe** — zakrzywionej struktury nie rozwinie. Metody nieliniowe układają próbki na płaszczyźnie tak, by podobne znalazły się blisko siebie.

t-SNE (*t-distributed stochastic neighbour embedding*)

Zamienia odległości w przestrzeni wejściowej na prawdopodobieństwa sąsiedztwa i szuka układu na płaszczyźnie, w którym te prawdopodobieństwa są możliwie podobne.

UMAP (*uniform manifold approximation and projection*)

Buduje graf k najbliższych sąsiadów (ang. *k-nearest neighbour graph*) próbek i szuka układu na płaszczyźnie, który najwierniej ten graf odtwarza.

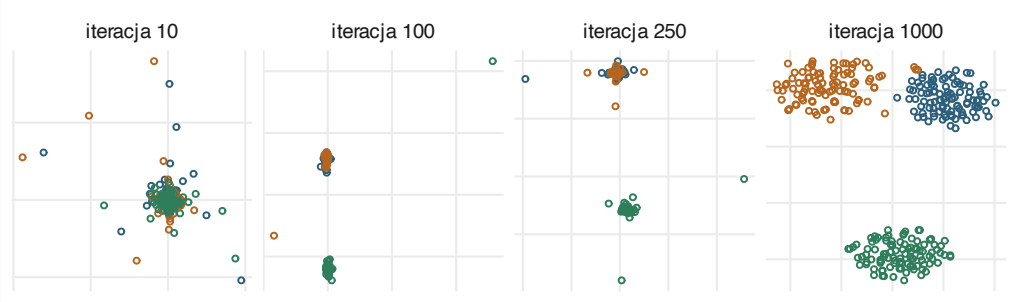
> To mapy **sąsiedztwa**, nie skali: wielkości skupisk ani odstępów między nimi nie czyta się jak z linijki.

t-SNE w trzech krokach — intuicja

1. **Kto jest czyim sąsiadem w danych?** Dla każdej próbki pozostałe dostają wagi: bliskie — duże, dalekie — prawie zero. Ilu sąsiadów liczy się naprawdę, ustala **perplexity**.
2. **Losowy start na płaszczyźnie.** Punkty rozrzucone przypadkowo; dla nich liczy się takie same wagi sąsiedztwa, ale według rozkładu **t** (ciężkie ogony, ang. *heavy tails*).
3. **Przesuwanie punktów w iteracjach** (domyślnie 1000 kroków). W każdym kroku punkty przyciągają się, gdy w danych są sąsiadami, a na mapie leżą daleko; odpychają, gdy jest odwrotnie. Kończy się, gdy wagi na mapie są możliwie podobne do wag w danych.

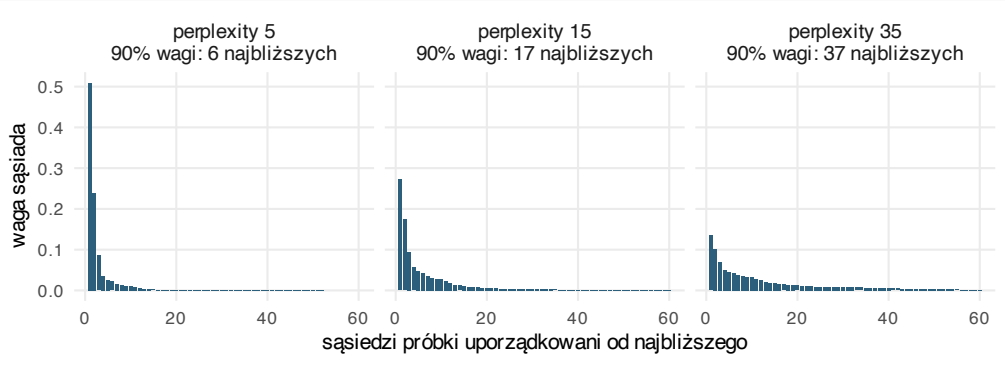
> Algorytm dba wyłącznie o **sąsiadów**: kto obok kogo. Odległości między odległymi skupiskami nie są przez niego pilnowane.

t-SNE krok po kroku



- Dane symulowane: trzy skupiska (A, B, C). Iteracja 10: losowy start skupiony wokół zera; 100 i 250: faza **wczesnego wzmocnienia** — punkty ściśnięte; C już oddzielone, A i B jeszcze razem; 1000: układ końcowy.
- **Wczesne wzmocnienie** (*early exaggeration*): przez pierwsze ok. 250 iteracji wagi w danych są sztucznie zwiększone, żeby skupiska zdążyły się rozdzielić.
- Za mało iteracji = mapa niedokończona; liczbę kroków ustala `max_iter`.

Perplexity — ilu sąsiadów „widzi” próbka



- Wagi sąsiadów jednej próbki danych demo: przy perplexity 5 prawie całą wagę ma 6 najbliższych, przy 35 — 37.
- Perplexity $\approx$ **efektywna liczba sąsiadów** (ang. *effective number of neighbours*); typowo 5–50, $< n/3$ (tu ≤ 37).
- Mała: mapa pokazuje drobną strukturę lokalną (i szum); duża: szerszy kontekst, mniej rozdrobnione skupiska.

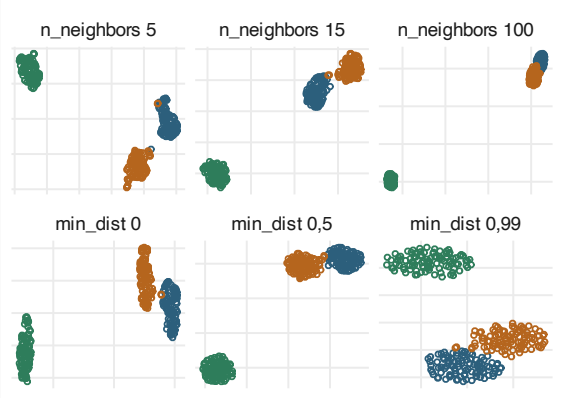
t-SNE — założenia

- **Podobieństwo = bliskość na wejściu**: próbki bliskie w danych (np. w 10 pierwszych PC, zwykle odległość euklidesowa) uznaje się za sąsiadów; mapa jest tak dobra jak przestrzeń wejściowa.
- **Każda próbka ma własne otoczenie**: w gęstym obszarze wąskie, w rzadkim szerokie, dobrane tak, by liczba sąsiadów była zbliżona do perplexity; stąd wyrównane gęstości skupisk.
- **Na mapie podobieństwo maleje wolniej** (rozkład t): na płaszczyźnie brakuje miejsca dla wszystkich sąsiadów (problem stłoczenia, ang. *crowding problem*), więc skupiska się rozsuwają.
- **Liczą się sąsiedzi, nie odległości**: rozdzielenie prawdziwych sąsiadów jest mocno karane, zbliżenie dalekich punktów — słabo.
- **Losowy start i przesuwanie krok po kroku**: wynik zależy od ziarna i liczby iteracji.

UMAP — intuicja i parametry

1. Dla każdej próbki znajduje się jej k **najbliższych sąsiadów** (`n_neighbors`) i łączy krawędziami: powstaje **graf**.
2. Punkty układa się na płaszczyźnie: połączone krawędzią się **przyciągają**, niepołączone **odpychają**.

- `n_neighbors`: jak szeroko patrzeć — mała wartość: szczegóły lokalne; duża: układ ogólny.
- `min_dist`: jak ciasno wolno upakować punkty — mała: zwarte skupiska; duża: rozlane.

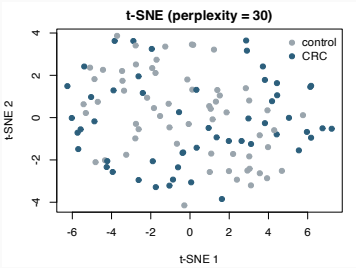


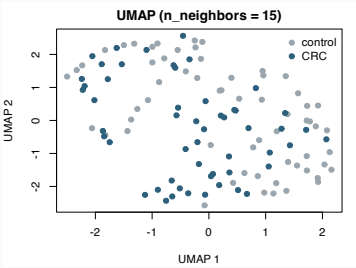
Te same trzy skupiska symulowane; górny rząd: `min_dist` 0,1, dolny: `n_neighbors` 15.

UMAP — założenia

- **Ważne są tylko bliskie odległości**: między bliskimi próbkami odległość jest wiarygodna, między dalekimi — nie.
- **Wystarczy k sąsiadów**, by opisać strukturę danych: dane zastępuje się grafem połączeń (`n_neighbors`); każda próbka ma co najmniej jednego sąsiada.
- **Start z układu grafu**, nie losowy; wyuczony model można zastosować do nowych próbek.

t-SNE i UMAP na danych demo

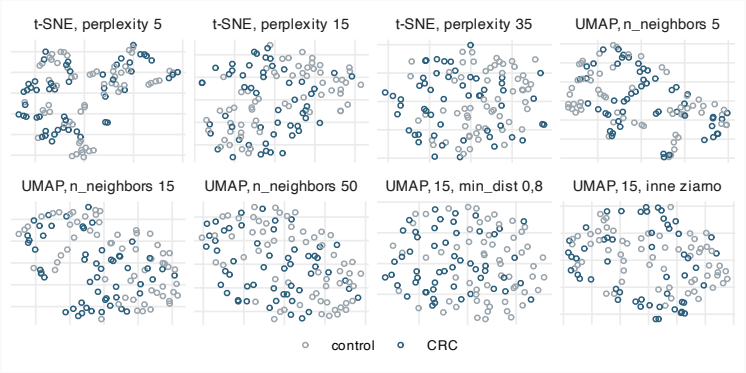




Wejście obu metod: 10 pierwszych składowych PCA (36% wariancji); t-SNE: `perplexity = 30` (w przybliżeniu efektywna liczba sąsiadów), UMAP: `n_neighbors = 15`. Kolor to etykieta naniesiona po fakcie.

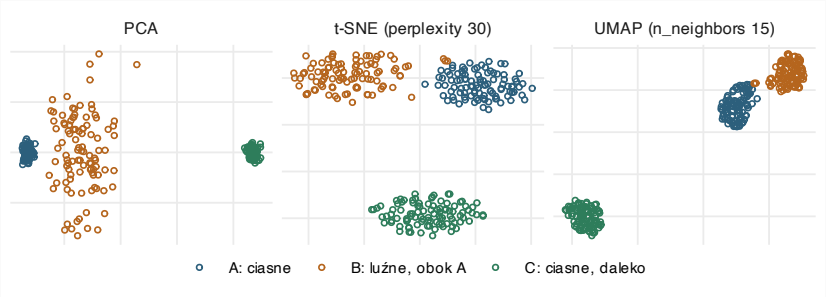
> Obie mapy: lokalne zagęszczenia, ale chore i zdrowe próbki **przeplatają się**. To przesłanka, że sygnał choroby jest słaby albo rozproszony — nie dowód, że go nie ma; obie metody liczono na tym samym wejściu, więc nie są niezależnymi potwierdzeniami.

t-SNE i UMAP — wpływ parametrów na danych demo



- Kształt i liczba „wysp” zależą od parametru i ziarna — pojedyncza mapa nie dowodzi istnienia skupisk.
- We wszystkich wariantach kolory się przenikają: to wniosek **stabilny**.
- t-SNE i UMAP lepiej niż PC1–PC2 zachowują bliskich sąsiadów, gorzej — odległości między odległymi próbkami.

Czego mapy t-SNE i UMAP nie zachowują — symulacja



Trzy skupiska po 100 punktów w 10 wymiarach: A ciasne, B 6 razy bardziej rozproszone obok A, C ciasne i daleko.

	dane	PCA	t-SNE	UMAP
rozrzut B / rozrzut A	6,3	6,8	1,2	1,0
odległość A–C / odległość A–B	4,2	4,2	2,4	3,0

t-SNE i UMAP — parametry i ostrożność

Co steruje wynikiem

- t-SNE: **perplexity** — jak dużego „sąsiedztwa” pilnować (typowo 5–50).
- UMAP: **n_neighbors** — podobna rola; `min_dist` — jak ciasno upakować punkty.
- ziarno losowe — inny start, inny obraz.

Reguły praktyczne

- uruchamiać przy kilku wartościach parametru i kilku ziarnach,
- liczyć na wejściu zredukowanym (PCA), nie na surowych cechach,
- traktować jako **wizualizację**, nie dowód skupień.

> Stabilność obrazu przy różnych parametrach i ziarnach jest ważniejsza niż jeden „ładny” rzut. Wniosek powtarzalny to wniosek wiarygodny.

PCA, t-SNE i UMAP — porównanie

	PCA	t-SNE	UMAP
typ	liniowa (obrót + rzut)	nieliniowa, oparta na sąsiedztwie	nieliniowa, oparta na grafie sąsiadów
co zachowuje	wariancję i odległości globalne	lokalne sąsiedztwo; układ grup zależy od startu	lokalne sąsiedztwo; układ grup zależy od startu
osie	interpretowalne: ładunki, % wariancji	bez znaczenia	bez znaczenia
powtarzalność	deterministyczna	zależy od ziarna i perplexity	zależy od ziarna, `n_neighbors`, `min_dist`
nowe próbki	`predict()`	brak w `Rtsne`	`umap_transform()`
skala danych	szybka; tysiące cech	wolna przy dużym n	szybsza od t-SNE
zastosowanie	kontrola jakości, efekt partii, wejście do map i modeli	wizualizacja skupisk	wizualizacja; wejście do skupień (ostrożnie)
nie wolno	—	czytać rozmiaru i odstępów skupisk	czytać rozmiaru i odstępów skupisk

t-SNE czy UMAP

Kryterium	t-SNE	UMAP
liczba próbek	do kilkudziesięciu tysięcy; wolniejszy	setki tysięcy; szybszy
nowe próbki	w `Rtsne` trzeba liczyć mapę od nowa	`umap_transform()` na zapisanym modelu
powtarzalność	losowy start; mapy różnią się między ziarnami	start z grafu; mniejsza zmienność
wygląd mapy	skupiska równomiernie rozłożone	zwarte skupiska z odstępami (`min_dist`)
zwyczaj dziedziny	cytometria (viSNE)	scRNA-seq (Seurat, Scanpy)

- **UMAP**: duże zbiory, potrzeba umieszczania nowych próbek na istniejącej mapie, analiza jednokomórkowa.
- **t-SNE**: zbiory średniej wielkości, gdy liczy się czytelny obraz lokalnych grup.
- Przy tym samym punkcie startowym obie dają podobne mapy; w obu sprawdza się kilka ziaren i parametrów.

Kiedy którą metodę stosować

- **Zawsze najpierw PCA**: wariancja wyjaśniona, próbki odstające, efekt partii; 10–50 składowych jako wejście dalszych metod.
- **scRNA-seq** (Seurat, Scanpy): normalizacja $\longrightarrow$ PCA (ok. 30–50 PC) $\longrightarrow$ graf k sąsiadów $\longrightarrow$ skupienia na grafie (Louvain, Leiden) $\longrightarrow$ UMAP **tylko do rysunku**.
- **Cytometria** (dziesiątki tysięcy komórek, kilkadziesiąt markerów): t-SNE (viSNE) lub UMAP do przeglądu populacji.
- **Mikrobiom, dziesiątki–setki próbek**: PCA po CLR albo PCoA/NMDS na wybranej mierze + PERMANOVA (omówione dalej); t-SNE i UMAP niewiele dodają.
- **Wnioski** (różnice grup, istnienie skupień) wyciąga się z testów i danych wejściowych, nie z położenia na mapie.

Rola miary odległości

Zarówno ordynacja, jak i analiza skupień sprowadzają się do jednego pytania: **jak bardzo dwie próbki się różnią?** Odpowiada na nie **miara odległości** (lub niepodobieństwa).

- W ekologii ordynację prowadzi się w dwóch krokach: macierz odległości między próbkami (np. **Bray–Curtis**), potem rozmieszczenie próbek w kilku osiach (PCoA — za chwilę).
- Zmiana miary zmienia, które próbki uznamy za „blisko” — a więc i wynik skupień oraz kształt ordynacji.
- Dla danych mikrobiomowych liczy się, czy patrzymy na **liczebności**, czy tylko na **obecność** taksonów, oraz czy uwzględniamy **pokrewieństwo** gatunków.

> Sześć miar: euklidesowa po CLR (Aitchisona), Bray–Curtis, Jaccard — policzone na danych demo; UniFrac (wymaga drzewa filogenetycznego), Jensena–Shannona i korelacyjna — w tabeli porównawczej.

Trzy miary w R

```r
# euklidesowa po CLR (liczebności w skali log-ratio)
Deu <- as.matrix(dist(Xclr))
# Bray-Curtis — różnica składów na surowych udziałach
Dbc <- as.matrix(vegan::vegdist(Xf, method = "bray"))
# Jaccard — tylko obecność/nieobecność taksonów
Dja <- as.matrix(vegan::vegdist(Xf > 0, method = "jaccard", binary = TRUE))
```

Każda z nich zwraca symetryczną macierz $n \times n$ z zerami na przekątnej. Poniżej róg 4×4 dla pierwszych czterech próbek.

Euklidesowa po CLR

Odległość Aitchisona: zwykła odległość euklidesowa („w linii prostej'') liczona na danych po CLR. Porównuje **względne** proporcje taksonów, więc pasuje do danych kompozycyjnych.

```
       S1    S2     S3     S4
S1   0.00 97.30 102.10 101.73
S2  97.30  0.00  93.60  86.96
S3 102.10 93.60   0.00  96.11
S4 101.73 86.96  96.11   0.00
```

- Skala zależy od liczby cech — liczby **same w sobie** niewiele mówią, porównujemy je **między parami**.
- Tu najbliżej siebie leżą S2 i S4 (87,0), najdalej S1 i S3 (102,1).

Bray–Curtis

Klasyczna miara ekologiczna: jaka **część** składu różni dwie społeczności. Na udziałach to 1 minus wspólna część składu; od 0 (identyczne) do 1 (żadnego wspólnego taksonu).

```
      S1    S2    S3    S4
S1 0.000 0.816 0.647 0.706
S2 0.816 0.000 0.660 0.749
S3 0.647 0.660 0.000 0.679
S4 0.706 0.749 0.679 0.000

zakres w całej kohorcie: [0.334, 0.999]   mediana: 0.725
```

> Mediana odległości w kohorcie to ok. **0,73** — typowa para próbek dzieli niewiele ponad jedną czwartą wspólnej liczebności; typowe dla rzadkich (ang. *sparse*) danych mikrobiomu.

Jaccard — tylko obecność

Ignoruje liczebności; pyta, jaka **część** taksonów obecnych w którejkolwiek z dwóch próbek nie jest wspólna. Od 0 (te same taksony) do 1 (żadnego wspólnego).

```
      S1    S2    S3    S4
S1 0.000 0.608 0.556 0.630
S2 0.608 0.000 0.601 0.527
S3 0.556 0.601 0.000 0.576
S4 0.630 0.527 0.576 0.000
```

> Te same próbki, inny obraz niż Bray–Curtis: para może dzielić wiele gatunków (mały Jaccard), a mimo to różnić się ich **udziałami** (duży Bray–Curtis).

UniFrac — odległość z filogenezą

UniFrac uwzględnia **pokrewieństwo** taksonów: na drzewie filogenetycznym sprawdza, jaka część długości gałęzi prowadzi tylko do taksonów jednej z próbek.

- **Nieważona** (ang. *unweighted*) — tylko obecność/nieobecność gałęzi (jak Jaccard, lecz z filogenezą).
- **Ważona** (ang. *weighted*) — dodatkowo bierze liczebności (jak Bray–Curtis, lecz z filogenezą).
- Dwie różne, lecz blisko spokrewnione bakterie liczą się jako **mała** różnica, a nie jako dwa całkiem obce taksony.

> Wymaga drzewa filogenetycznego dopasowanego do taksonów. Bywa czulsza na sygnał biologiczny, bo pokrewne gatunki często, choć nie zawsze, pełnią podobne funkcje.

Miary odległości — przykład liczbowy

	takson 1	takson 2	takson 3	takson 4
próbka A (%)	50	30	20	0
próbka B (%)	20	40	30	10

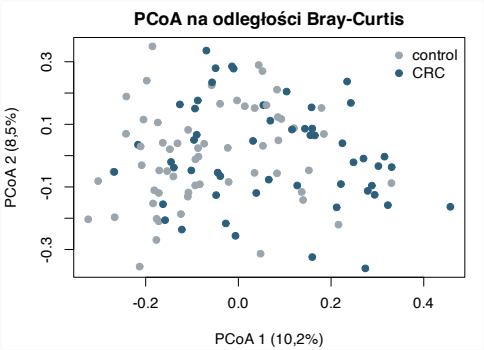
Miara	wartość	jak czytać
Bray–Curtis	0,30	próbki dzielą 70% składu
Jaccard	0,25	3 z 4 obecnych taksonów są wspólne
Jensena–Shannona	0,28	na skali 0–0,83: rozkłady umiarkowanie podobne
korelacyjna $(1 - r)$	0,63	kształt profili słabo zgodny ($r = 0{,}37$)
Aitchisona (CLR)	2,89	zależy od wartości wstawionej za zero (ang. *pseudocount*): od 1,88 do 4,19

Aitchison zależy od pseudoliczności dla zer (takson 4 w A); Bray–Curtis i Jaccard — nie.

PCoA — ordynacja dowolnej odległości

Analiza współrzędnych głównych (*principal coordinates analysis*, PCoA)

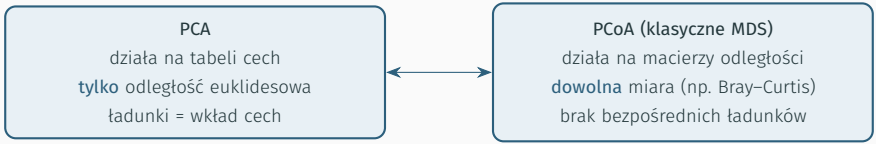
Inaczej klasyczne skalowanie wielowymiarowe (*classical MDS*): szuka współrzędnych próbek, przy których zwykłe odległości między nimi jak najwierniej odtwarzają zadaną macierz odległości — dowolnej miary, nie tylko euklidesowej.



PCoA na odległości Bray–Curtis. Pierwsze dwie osie: 10,2% i 8,5% (względem sumy dodatnich wartości własnych).

Chmury chorych i zdrowych silnie się nakładają, ale średni wynik na osi 1 różni się między CRC a kontrolą (test t Welcha na wynikach osi: $p \approx 1{,}5 \cdot 10^{-5}$, wobec $p \approx 5 \cdot 10^{-4}$ dla PC2 po CLR); oś 2 grup nie różnicuje.

PCA i PCoA



PCoA na odległości euklidesowej daje te same współrzędne co PCA (z dokładnością do znaku). Gdy miara pasująca do danych nie jest euklidesowa (Bray–Curtis, UniFrac) — sięga się po PCoA.

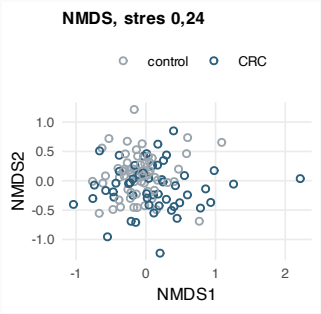
NMDS — ordynacja kolejności odległości

Niemetryczne skalowanie wielowymiarowe (*non-metric multidimensional scaling*, NMDS)

Szuka takiego układu próbek w zadanej liczbie wymiarów, w którym kolejność odległości na mapie jak najwierniej odtwarza kolejność odległości w macierzy wejściowej.

- Liczy się tylko to, **która** para jest bliżej, a nie o ile — dlatego metoda dobrze znosi miary takie jak Bray–Curtis.
- Liczbę wymiarów podaje się z góry; algorytm iteracyjny startuje losowo.
- Stres (*stress*) mierzy niezgodność kolejności: 0 — zgodność idealna.
- Osie NMDS **nie** mają procentu wariancji ani ładunków.

NMDS na danych demo



Stres na odległości Bray–Curtis:

- 2 wymiary: **0,24**
- 3 wymiary: **0,18**

Orientacyjnie (Clarke 1993): poniżej 0,1 — dobra mapa; do 0,2 — użyteczna; powyżej 0,3 — prawie dowolna.

Stres 0,24: mapa 2D jest tylko **zgrubna**. Różnice średnich na osiach (test t Welcha: $p \approx 0{,}03$ i $0{,}006$) są tylko opisowe (orientacja osi umowna); test różnicy składu to PERMANOVA.

Wybór miary odległości

Miara	Na co patrzy	Kiedy i do czego	Uwaga
Aitchisona (euklidesowa po CLR)	logarytmy ilorazów udziałów	dane kompozycyjne; spójna z PCA po CLR	zależy od pseudoliczności dla zer
Bray–Curtis	bezwzględne różnice udziałów	skład ilościowy; standard w NMDS, PCoA, PERMANOVA	zdominowana przez liczne taksony
Jaccard	obecność/nieobecność	skład jakościowy, rzadkie taksony	wrażliwa na głębokość sekwencjonowania
UniFrac (nieważona / ważona)	filogeneza (+ liczebności)	16S z drzewem; pokrewieństwo gatunków	nieważona — rzadkie linie, ważona — dominujące
Jensena–Shannona	rozkłady udziałów	profile udziałów; enterotypy (PAM)	zera bez pseudoliczności
korelacyjna ($1 - r$)	kształt profilu	grupowanie cech (taksonów, genów), mapy ciepła	na udziałach pozorna — po CLR

Wybór miary uzasadnia pytanie biologiczne; wrażliwość wyniku **sprawdza się** na dwóch miarach. U nas sygnał choroby: po CLR — PC2, w Bray–Curtis — pierwsza oś PCoA.

PERMANOVA — czy skład grup się różni

PERMANOVA (*permutational multivariate analysis of variance*)

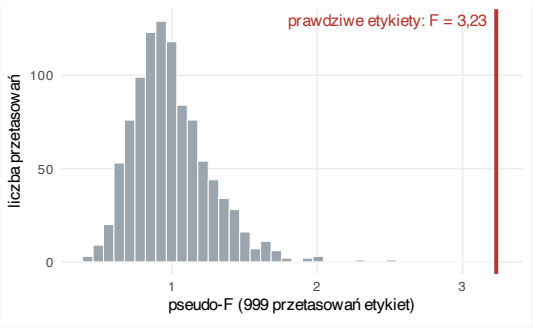
Permutacyjna wielowymiarowa analiza wariancji: test, czy grupy różnią się składem, liczony wprost z macierzy odległości między próbkami.

- **Co porównuje**: przeciętne odległości **między** próbkami z różnych grup i **wewnątrz** grup; ich stosunek to statystyka **pseudo-F** (jak F w analizie wariancji). Duże F — grupy wyraźnie się różnią.
- **Skąd p**: z permutacji (następny slajd), bo odległości nie mają rozkładu normalnego.
- R^2: część zmienności składu przypisana grupie (wielkość efektu, ang. *effect size*).
- Działa z dowolną miarą (Bray–Curtis, Jaccard, UniFrac, euklidesowa po CLR).

> Ordynacja **pokazuje**, PERMANOVA **testuje**. W pracach o mikrobiomie obie zwykle występują razem.

Test permutacyjny — idea

1. Liczy się statystykę (pseudo-F) dla **prawdziwych** etykiet.
2. Etykiety „chory/zdrowy” **tasuje się losowo** między próbkami (permutacja, ang. *permutation*): związek etykiety ze składem znika.
3. Statystykę liczy się ponownie; powtarza się to np. 999 razy.
4. p = odsetek przetasowań, które dały wynik co najmniej tak duży jak prawdziwy.



Dane demo: żadne z 999 przetasowań nie dało F tak dużego jak prawdziwe (3,23); $p = 1/1000 = 0{,}001$.

PERMANOVA na danych demo

```r
library(vegan)
D <- vegdist(Xf, method = "bray")          # Bray–Curtis na udziałach
adonis2(D ~ grupa, permutations = 999)
```

```
          Df SumOfSqs      R2      F Pr(>F)
Model      1   0.8574 0.02805 3.2319  0.001
Residual 112  29.7138 0.97195
Total    113  30.5712 1.00000
```

- $p = 0{,}001$ (kolumna `Pr(>F)`) — najmniejsza możliwa wartość przy 999 permutacjach: różnica składu jest wyraźna.
- $R^2 = 0{,}028$ — choroba wyjaśnia ok. 3% zmienności składu; reszta to różnice między osobami.

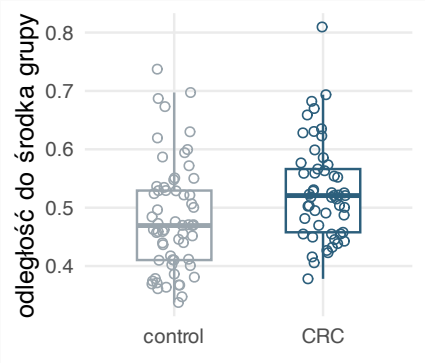
PERMANOVA z czynnikami zakłócającymi (*confounders*)

```
# by = "margin": efekt każdego czynnika po uwzględnieniu pozostałych
adonis2(D ~ wiek + plec + BMI + grupa, data = meta,
        permutations = 999, by = "margin")
```

```
          Df SumOfSqs      R2      F Pr(>F)
wiek       1   0.2309 0.00792 0.8824  0.655
plec       1   0.2581 0.00885 0.9863  0.463
BMI        1   0.3408 0.01168 1.3021  0.129
grupa      1   0.7703 0.02641 2.9434  0.001
Residual 105  27.4788 0.94207
```

Po uwzględnieniu wieku, płci i BMI (110 próbek z kompletnymi danymi) efekt choroby pozostaje ($R^2 = 0{,}026$, $p = 0{,}001$); żaden z trzech czynników nie wiąże się istotnie ze składem.

Rozrzut grup — betadisper



PERMANOVA reaguje na **przesunięcie środka** grupy, ale także na różny **rozrzut** (ang. *dispersion*) próbek wokół środka.

`betadisper()` + `permutest()`: średnia odległość do środka grupy

- kontrola **0,48**, CRC **0,53**; test permutacyjny $p = 0{,}005$
- próbki CRC są **bardziej zróżnicowane** między sobą

Wynik raportuje się razem: skład różni się ($R^2 \approx 3\%$), a część tej różnicy może wynikać z większego rozrzutu w grupie CRC. Przy niemal równych grupach (61 i 53) PERMANOVA jest na to względnie odporna.

Analiza skupień

Analiza skupień (*clustering*)

Podział próbek na grupy podobnych, bez znajomości jakiejkolwiek etykiety.

- Próbki w jednym skupieniu mają być do siebie podobne,
- a do próbek z innych skupień — możliwie niepodobne.
- „Podobne” zależy od przyjętej miary odległości.

> Skupienia to **propozycja** podziału wygenerowana z danych — nie etykiety z góry. Czy pokrywają się z chorobą, trzeba dopiero sprawdzić.

Rodziny metod skupień

Metody różnią się tym, co uznają za „skupienie” — i dlatego dają różne wyniki na tych samych danych:

Rodzina	Co grupuje	Przykład
centroidalne (ang. *centroid-based*)	próbki wokół środka	k-średnich, k-medoidów
hierarchiczne (ang. *hierarchical*)	zagnieżdżone łączenia	`hclust`, Ward

> W danych mikrobiomowych najczęściej stosuje się te dwie rodziny; enterotypy wyznaczono metodą k-medoidów.

Metoda k-średnich — intuicja

Metoda k-średnich (*k-means*)

Algorytm dzielący próbki na z góry zadaną liczbę k skupień wokół środków (centroidów), które są średnimi przypisanych próbek.

1. Wybór k i rozmieszczenie k środków (centroidów).
2. Przypisanie każdej próbki do najbliższego środka.
3. Przeliczenie środków jako średnich przypisanych próbek.
4. Powtarzanie kroków 2–3 do ustabilizowania się przypisań.

Liczbę k **podajemy z góry** — metoda jej nie odkrywa. Wynik zależy też od losowego startu, dlatego uruchamiamy ją wielokrotnie (`nstart`).

k-średnich w R

Skupienia wyznaczamy na 10 pierwszych składowych PCA (36% wariancji; resztę pomijamy z założenia, że to głównie szum).

```
scores <- pca$x[, 1:10]                # pierwsze 10 składowych
set.seed(1)
km <- kmeans(scores, centers = 2, nstart = 25)
km$size                                  # liczebności skupień
table(skupienie = km$cluster, grupa = y)  # skupienie a rzeczywista grupa
```

```
[1] 64 50
         grupa
skupienie control CRC
        1      36  28
        2      25  25
```

Oba skupienia mają podobne proporcje chorych i zdrowych — podział **nie** biegnie wzdłuż choroby.

Czynniki wpływające na wynik k-średnich

Metoda jest szybka, ale ma dwa czułe punkty:

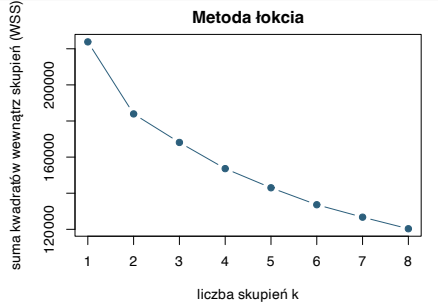
- **Start** — losowe położenie środków może dać różne skupienia przy każdym uruchomieniu. Stąd `nstart = 25`: bierzemy najlepszy z wielu startów.
- **Kształt skupień** — k-średnich zakłada skupienia **kuliste** i podobnej wielkości; wydłużone lub zagnieżdżone struktury rozcina błędnie.

> k-średnich przypisze **każdą** próbkę do któregoś skupienia — także obserwacje odstające. Nie ma kategorii „szum”, choćby punkt leżał z dala od wszystkich środków.

Liczba skupień: metoda łokcia

Skoro k podajemy ręcznie, potrzebujemy przesłanki, ile go przyjąć.

- **Metoda łokcia** (ang. *elbow*) — rysujemy WSS dla kolejnych k: sumę kwadratów odległości próbek od środków ich skupień.
- Spada z k; szukamy **załamania**, po którym poprawa jest już niewielka.
- **Sylwetka** (ang. *silhouette*) — mierzy, jak dobrze próbka pasuje do swojego skupienia względem sąsiedniego.



Żadne kryterium nie wskaże „prawdziwego” k — podpowiadają sensowny zakres, decyzja należy do analityka i kontekstu biologicznego.

Szerokość sylwetki

Szerokość sylwetki (*silhouette width*)

Jak dobrze próbka pasuje do swojego skupienia w porównaniu z najbliższym innym: blisko 1 — dobrze, ok. 0 — leży na granicy, ujemna — raczej w złym skupieniu.

Średnia sylwetka wszystkich próbek (k-średnich na PC1–PC10):

k	sylwetka
2	0.161
3	0.125
4	0.120
5	0.124
6	0.128

- Najwyższa wartość przy $k = 2$ — to ją wybralibyśmy.
- Ale **0,16** to **mało**: poniżej ok. 0,25 skupienia są słabo oddzielone (Kaufman i Rousseeuw).
- Dla $k = 3$–6 wartości są prawie równe — żadne k się nie wyróżnia.

Niska sylwetka przy każdym k oznacza, że w danych nie ma wyraźnych, oddzielonych grup.

Metoda hierarchiczna — intuicja

Hierarchiczna analiza skupień (*hierarchical clustering*)

Metoda budująca drzewo zagnieżdżonych skupień przez kolejne łączenie najbliższych grup; nie wymaga podania k z góry.

- Start: każda próbka to osobne skupienie.
- W każdym kroku łączymy dwa najbliższe skupienia.
- Powtarzamy, aż zostanie jedno — zapisując kolejność łączeń.

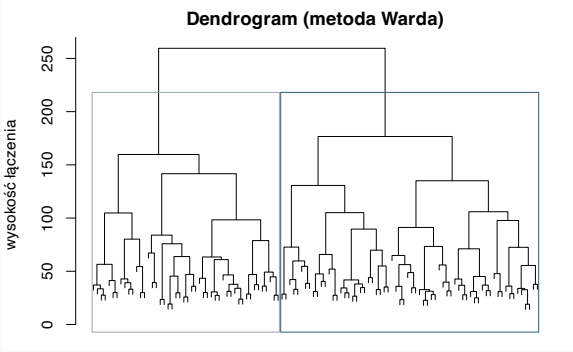
Efekt to **dendrogram** — drzewo, które można „przeciąć” na dowolnej wysokości i dopiero wtedy zdecydować o liczbie skupień.

hclust w R i dendrogram

```r
d  <- dist(scores)
hc <- hclust(d,
       method = "ward.D2")

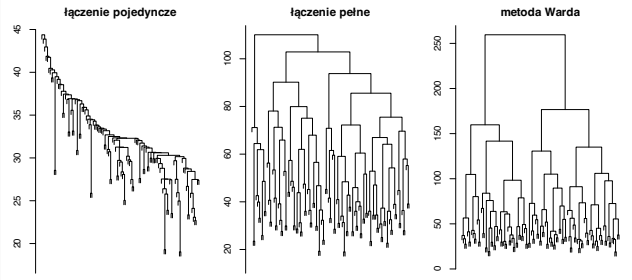
table(
  skupienie = cutree(hc, k = 2),
  grupa     = y)
```

```
          grupa
skupienie control CRC
        1      32  34
        2      29  19
```



Trzy metody łączenia obok siebie

Wynik zależy od **metody łączenia** (ang. *linkage*) skupień. Ten sam zbiór, ta sama odległość:



- **Pojedyncze** (ang. *single*) — łączy po najbliższej parze; skłonne do długich „łańcuchów” i jednego dominującego skupienia.
- **Pełne** (ang. *complete*) — po najdalszej parze; zwarte, zbliżone rozmiarem skupienia.
- **Warda** (ang. *Ward*) — minimalizuje przyrost wariancji; zwykle najbardziej zbalansowane drzewo.

Dendrogram

- **Wysokość łączenia** — im wyżej łączą się gałęzie, tym mniej podobne skupienia.
- **Cięcie** poziomą linią wyznacza liczbę skupień: niżej — więcej i drobniejszych, wyżej — mniej i większych.
- Kolejność liści na osi **nie** jest informatywna — to samo drzewo można narysować na wiele sposobów.

> Dendrogram pokazuje strukturę **na wielu poziomach** naraz — stąd jego przewaga, gdy nie wiemy z góry, ilu skupień szukać.

Metoda k-medoidów (*partitioning around medoids*, PAM)

Odmiana metody k-średnich, w której środkiem skupienia jest medoid — rzeczywista próbka o najmniejszej sumie odległości do pozostałych członków skupienia.

- Wymaga tylko macierzy odległości, więc działa z dowolną miarą (Bray–Curtis, Jensen–Shannon), a nie tylko z euklidesową.
- Mniej wrażliwa na wartości odstające niż k-średnich: medoid nie „przesuwa się” w stronę skrajnej próbki.
- Medoid jest prawdziwą próbką, więc można go opisać jak typowego przedstawiciela skupienia.

Enterotypy — klasyczny przykład z mikrobiomu

Enterotyp (*enterotype*)

Grupa osób o podobnym ogólnym składzie mikrobiomu jelitowego, wyróżniona analizą skupień i nazwana od rodzaju bakterii, który w niej dominuje.

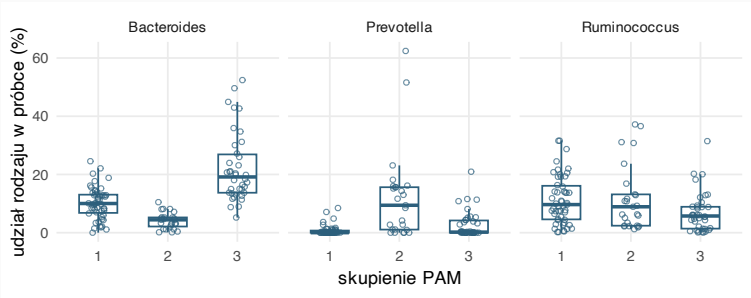
Na 33 metagenomach osób z sześciu krajów zespół Arumugam (2011) metodą k-medoidów wyróżnił **trzy** typy społeczności — **enterotypy** — potwierdzone w tej samej pracy na dwóch większych zbiorach (85 i 154 osoby):

- zdominowany przez *Bacteroides*,
- zdominowany przez *Prevotella*,
- zdominowany przez *Ruminococcus*.

To wzorcowy przykład uczenia nienadzorowanego: struktura wyłoniła się **z danych**, bez żadnej etykiety z góry; późniejsze badania wiązały ją z długotrwałą dietą (Wu i in., 2011).

Dalsze prace pokazały, że granice enterotypów bywają **płynne** — to raczej gradient niż trzy ostre pudełka. Idealna ilustracja zasady: skupienie to hipoteza, którą trzeba zweryfikować, nie fakt.

Enterotypy w danych demo (PAM)



Jak u Arumugam i in.: udziały **rodzajów**, odległość Jensena–Shannona, PAM, $k = 3$ (51, 25 i 38 próbek). Średni udział *Bacteroides*: 10, 4 i 22%; *Prevotella*: 1, 12 i 3%.

> Średnia sylwetka dla $k = 2$–6 wynosi tylko 0,05–0,07: trzy „typy” są widoczne w średnich, ale granice między nimi są **płynne**. Skupienia nie pokrywają się z chorobą (ARI 0,02).

Porównanie metod skupień

	k-średnich	k-medoidów (PAM)	hierarchiczna
liczba skupień	podana z góry	podana z góry	wybierana po fakcie (cięcie drzewa)
odległość	euklidesowa	dowolna	dowolna
środek skupienia	średnia (punkt sztuczny)	medoid (prawdziwa próbka)	brak
wynik	jeden podział	podział + medoidy	całe drzewo
powtarzalność	zależy od startu	stała	stała

Reguła praktyczna: hierarchiczna do eksploracji, k-średnich do szybkiego podziału, PAM gdy miara nie jest euklidesowa (np. Bray–Curtis).

Analiza skupień taksonów

Dotąd grupowaliśmy **próbki**. Tę samą macierz można obrócić i grupować **taksony** — po wzorcu ich liczebności w próbkach.

- Taksony, które **współwystępują**, tworzą grupy współwystępowania (ang. *co-abundance groups*).
- Interpretujemy je jako możliwe **gildie** (ang. *guilds*) funkcjonalne lub człony sieci zależności ekologicznych.
- Technika ta sama (odległość + skupienia), zmienia się tylko to, **co** jest wierszem, a co kolumną.

> Połączenie obu kierunków daje **mapę ciepła z dendrogramami** na obu osiach — klasyczny obraz, na którym naraz widać grupy próbek i grupy taksonów.

Ocena i pułapki

Ocena skupień bez etykiet

Bez prawdziwych etykiet nie policzymy dokładności. Oceniamy **spójność** podziału:

- **Szerokość sylwetki** — bliska 1: próbki dobrze dopasowane; bliska 0: leżą na granicy; ujemna: prawdopodobnie w złym skupieniu.
- **Stabilność** — czy podział się utrzymuje po drobnej zmianie danych lub parametrów.
- **Zgodność z etykietą** — porównanie ze znaną etykietą, **jeśli** jest, traktowane jako sprawdzian po fakcie.
- **Ocena względna** — porównanie wariantów (różne k, miary, metody).

> Dobra szerokość sylwetki świadczy o spójności geometrycznej, **nie** o znaczeniu biologicznym. To dwie różne rzeczy.

Zgodność skupień z etykietą — skorygowany indeks Randa

Gdy etykieta istnieje (tu: chory/zdrowy), można po fakcie sprawdzić zgodność. **Indeks Randa**: odsetek par próbek, co do których oba podziały się zgadzają (razem albo osobno). **Skorygowany** (ARI, ang. *adjusted Rand index*) odejmuje zgodność oczekiwaną przy losowym podziale: 1 — idealna, 0 — poziom przypadku, wartości ujemne — poniżej przypadku.

```
# skupienia k-średnich a rzeczywista etykieta choroby
mclust::adjustedRandIndex(km$cluster, y)
```

```
[1] -0.004
```

> ARI ≈ 0: podział na skupienia pokrywa się z chorobą **nie lepiej niż losowo**. Najsilniejsza struktura w tej przestrzeni (10 składowych po CLR) jest innej natury niż choroba.

Stabilność — test wiarygodności

Jeśli podział odzwierciedla realną strukturę, powinien **przetrwać** drobne zaburzenie danych.

- **Bootstrap** — losowanie próbek ze zwracaniem (ang. *sampling with replacement*; próba tej samej liczności, szczegóły w W6); **podpróbkowanie** (ang. *subsampling*) — losowa część próbek bez zwracania. Po każdym losowaniu powtarzamy analizę skupień i sprawdzamy, jak często te same próbki są razem.
- **Perturbacja parametrów** — zmieniamy k, miarę, ziarno losowe.
- Skupienie **powtarzalne** jest wiarygodniejsze niż takie, które rozpada się po najmniejszej zmianie.

> Reguła praktyczna: zanim nazwiemy skupienie „podtypem choroby”, pokażmy, że jest stabilne — inaczej opisujemy przypadkowy ślad tej jednej próbki danych.

Skupienie jako hipoteza

Metody skupień **zawsze** zwrócą jakiś podział — nawet dla danych czysto losowych. Istnienie skupień w wyniku nie dowodzi, że istnieją w rzeczywistości.

- k-średnich poproszone o $k = 3$ zwróci trzy grupy, **czy mają sens, czy nie**.
- Podział trzeba **zweryfikować**: spójnością, stabilnością, wiedzą dziedzinową.
- Skupienie to **punkt wyjścia** do hipotezy, a nie jej potwierdzenie.

Typowe pułapki analizy nienadzorowanej

- **Brak skalowania** — cechy w różnych jednostkach zaburzają odległości.
- **Ślepa interpretacja osi** — PC1 bywa artefaktem technicznym.
- **Nadinterpretacja t-SNE/UMAP** — odczyt odległości jak z linijki.
- **Wybór k pod z góry przyjętą tezę** — zamiast z danych i kontekstu.
- **Mylenie spójności ze znaczeniem** — ładny podział to nie dowód biologii.

> Metody nienadzorowane są bardzo użyteczne w **eksploracji** — pod warunkiem, że ich wyniki traktujemy jako pytania, nie odpowiedzi.

Podsumowanie

Do zapamiętania

- Uczenie nienadzorowane opisuje **strukturę** danych bez etykiet.
- **PCA** zastępuje skorelowane cechy kilkoma osiami największej zmienności (wektory własne macierzy kowariancji); czytamy wariancję wyjaśnioną, wykres rozrzutu (ang. *scatter plot*) i ładunki.
- **Biplot** i ładunki wskazują taksony-kandydatów, nie wynik testu; t-SNE i UMAP to mapy **sąsiedztwa** — bez odczytu rozmiaru i odstępów skupisk.
- Obraz zależy od **miary odległości**; ordynacja (PCoA, NMDS ze stresem) **pokazuje**, **PERMANOVA** testuje: $R^2 \approx 3\%$, $p = 0{,}001$, przy większym rozrzucie w CRC.
- **Analiza skupień** proponuje grupy: k-średnich, hierarchiczna, k-medoidów; sylwetka i enterotypy wskazują tu strukturę **gradientową**.
- Każdy wynik to **hipoteza** do weryfikacji, nie gotowa prawda.

Literatura

- James, Witten, Hastie, Tibshirani. *An Introduction to Statistical Learning* — rozdz. 12 „Unsupervised Learning” (PCA, k-średnich, metoda hierarchiczna). Wydanie R oraz Python (ISLP), dostęp bezpłatny: statlearning.com.
- Boehmke, Greenwell. *Hands-On Machine Learning with R* — rozdz. o PCA i analizie skupień: bradleyboehmke.github.io/HOML.
- VanderPlas. *Python Data Science Handbook* — PCA, k-średnich: jakevdp.github.io/PythonDataScienceHandbook.
- Legendre, Legendre. *Numerical Ecology*, wyd. 3 (2012) — miary odległości, PCoA, NMDS.
- Anderson (2001). A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26:32–46 — PERMANOVA.

Na następny wykład

Wykład 4: Uczenie nadzorowane

- dane do modelu: podział i przekształcenia liczone na treningu; granica decyzyjna (ang. *decision boundary*)
- trzy proste modele: k najbliższych sąsiadów, regresja logistyczna, naiwny Bayes
- walidacja krzyżowa (ang. *cross-validation*) i wybór hiperparametru
- regresja wartości ciągłej; pełny potok (ang. *pipeline*) bez przecieku

Dziękuję za uwagę

dr inż. Bartosz Czech · bartosz.czech@upwr.edu.pl