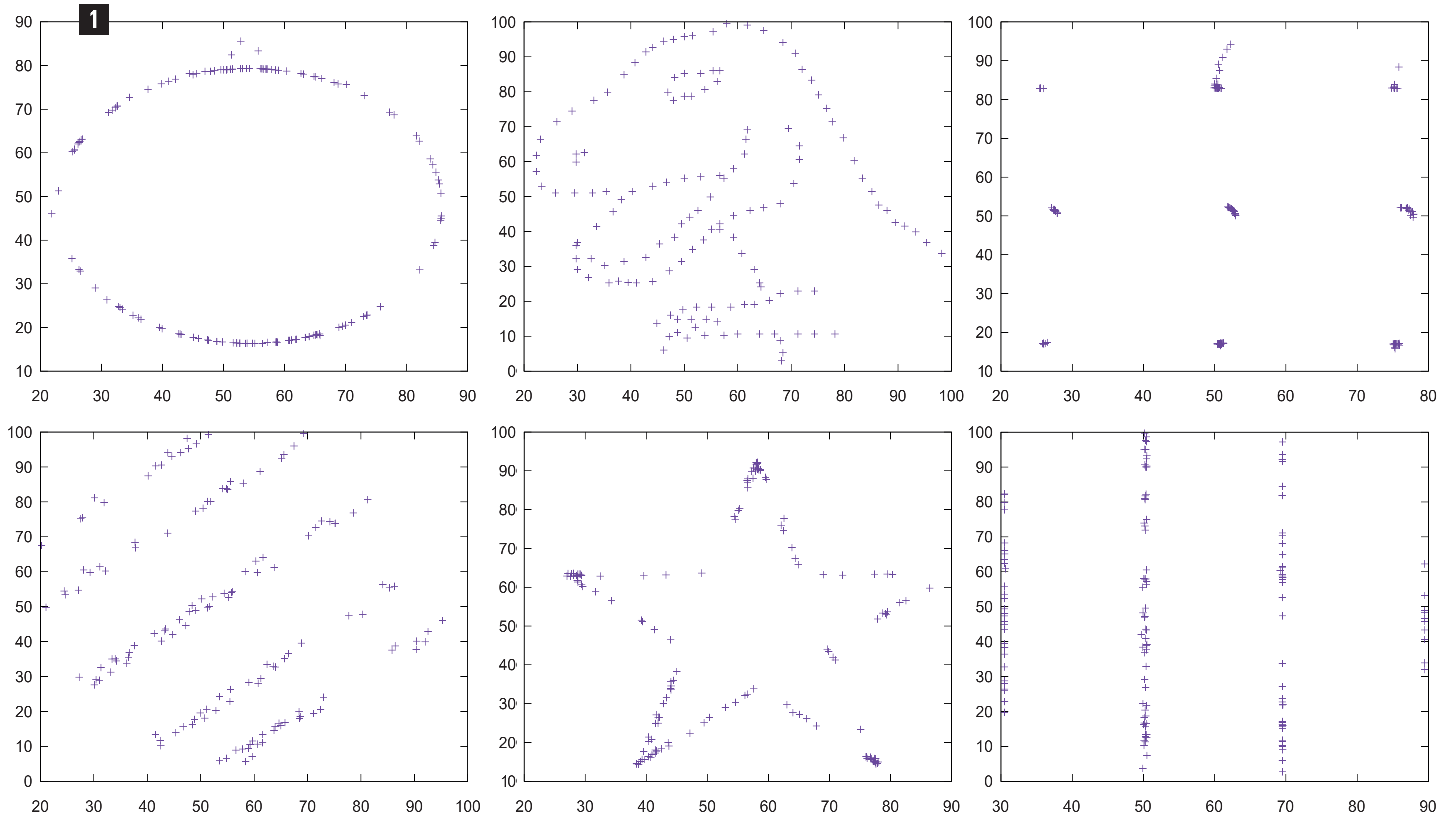
DOI: 10.24425/academiaPAN.2022.143980



OBRAZY FUNKCJI W PRAKTYCE

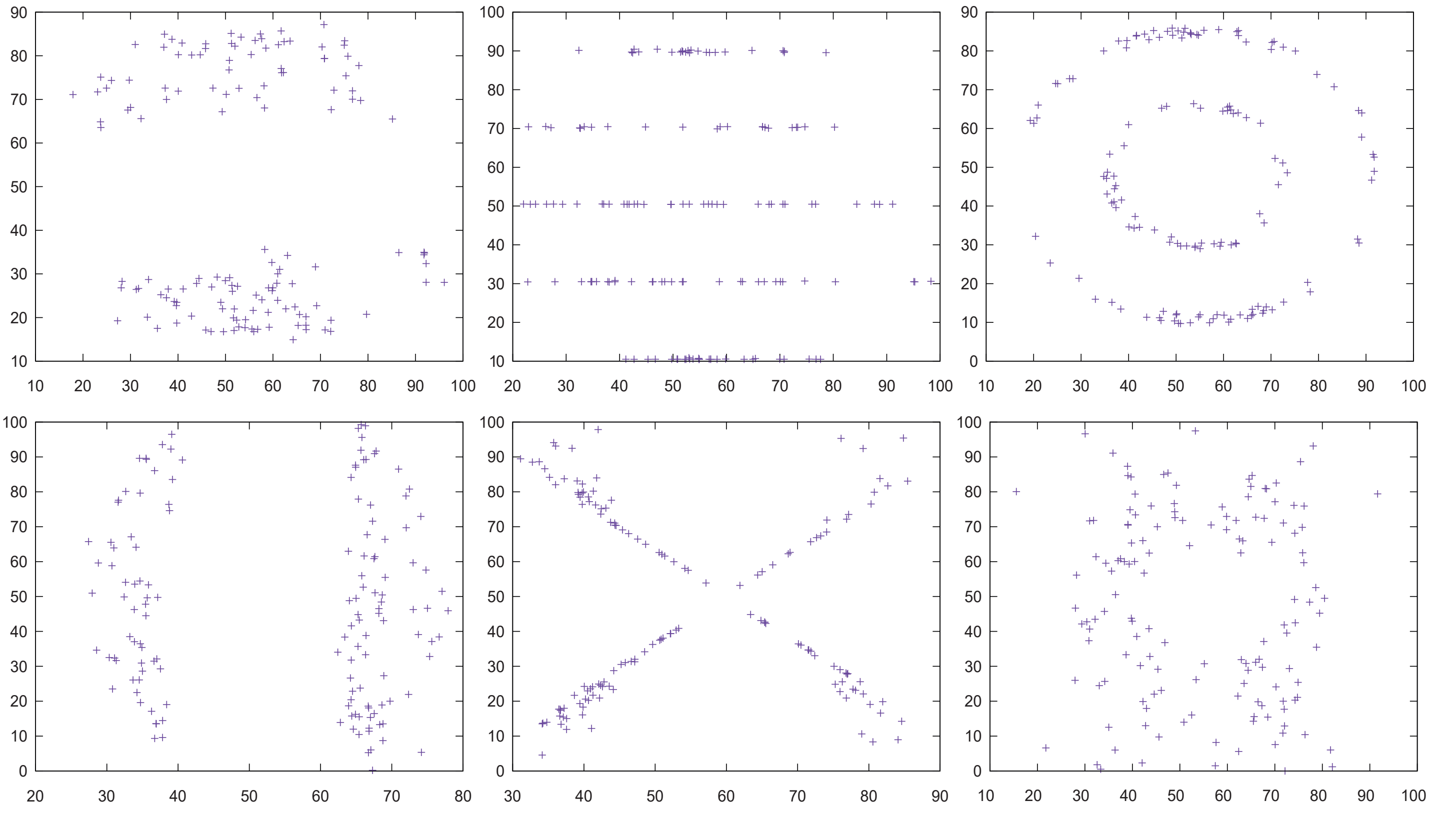
dr Paweł Dłotko
Ukończył studia magisterskie i doktoranckie z matematyki obliczeniowej na Uniwersytecie Jagiellońskim. Następnie pracował na Uniwersytecie Pensylwanii, w Inria Saclay Centre i na Uniwersytecie w Swansea. Obecnie jest dyrektorem Centrum Dioscuri w Topologicznej Analizie Danych mieszczącym się w Instytucie Matematycznym PAN.
pawel.dlotko@impan.pl

Wizualizacja funkcji matematycznych znalazła liczne zastosowania w życiu codziennym, od ekonomii po medycynę.

Paweł Dłotko

Instytut Matematyczny
Polska Akademia Nauk w Warszawie

Matematyka integruje i uogólnia liczne koncepcje obrazu, utożsamiając go z funkcją o określonych argumentach (dziedziną) oraz odpowiadającym im wartościom (przeciwdziedziną), nazywaną również – nie bez powodu – obrazem funkcji. Funkcje są wszechobecne i wykraczają daleko poza szkolną matematykę. Przykładem może być zdjęcie przechowywane w pamięci naszego telefonu. Jego dziedziną jest zbiór pikseli – malutkich kwadratów na ekranie telefonu, przeciwdziedziną zaś jest zbiór możliwych kolorów, typowo określany w informatyce za pomocą natężenia barwy czerwonej, zielonej i niebieskiej (paleta kolorów RGB). Każdy z pikseli świeci w unikatowej kombinacji tych kolorów, dając całość obrazu, czyli oglądane przez nas zdjęcie. Idealizując, możemy powiedzieć, że dziedziną zdjęcia jest



Datazaurus dozen – przykład chmur punktów o bardzo zbliżonych wartościach statystyki opisowej i skrajnie różnych kształtach

płaszczyzna (ekran telefonu), a obrazem przestrzeń o trzech wymiarach, gdzie wartości liczbowe przypisane kolejnym wymiarom określają natężenia trzech wymienionych barw. Dawne fotografie, zwane potocznie zdjęciami czarno-białymi, są łatwiejsze w interpretacji. Ich dziedziną może być ponownie zbiór pikseli, a obrazem – paleta odcieni szarości wyrażona w postaci wartości liczby rzeczywistej przedstawiającej nasycenie obrazu czarną barwą.

Chmury punktów

W analizie danych funkcje są zdefiniowane w przestrzeni wielowymiarowej. Ideę takiej przestrzeni można wyjaśnić na przykładzie pojedynczego punktu. Tradycyjnie używamy przestrzeni dwuwymiarowej, gdzie punkt umieszczamy w kartezjańskim układzie współrzędnych z osiami X i Y i opisujemy dwoma wartościami. Jeśli zaś chcemy przedstawić taki punkt w 12 wymiarach, to opisujemy go 12 wartościami, które określają jego poszczególne atrybuty, takie jak wysokość, szerokość, długość, współrzędne geograficzne i dowolne siedem innych cech, które chcemy mu przypisać.

W sumie zestaw przypisanych cech określa położenie punktu w 12-wymiarowej przestrzeni. Skończone zbiory punktów, zwane chmurami punktów, same w sobie mają pewien kształt, który często niesie istotną informację. Przykładem dwuwymiarowej chmury punktów jest Datasaurus Dozen (rys. 1).

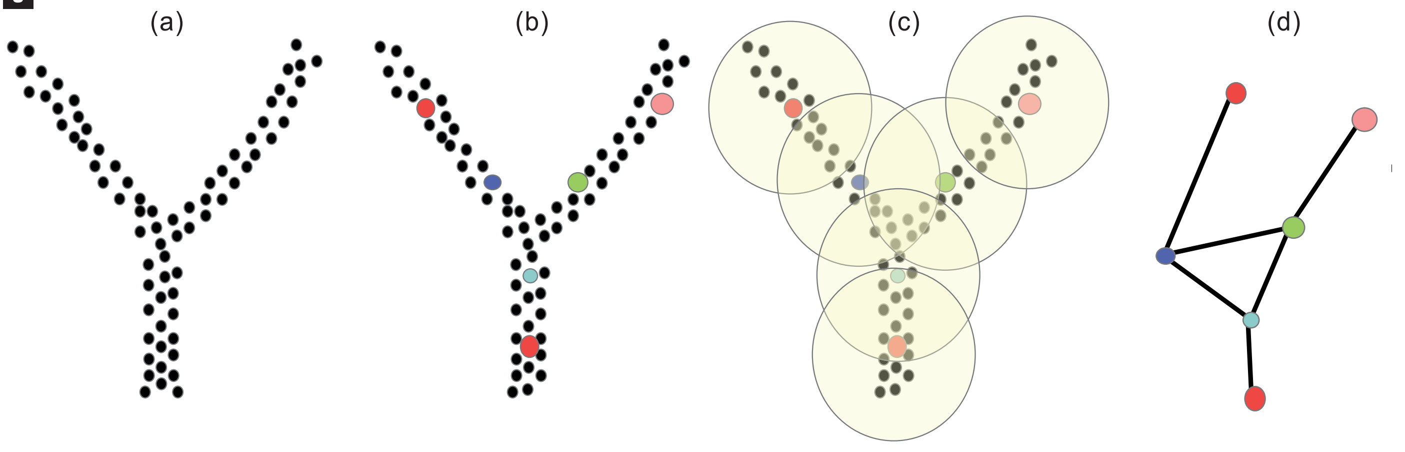
Powyższe 12 chmur punktów ma istotnie różne kształty. Jednak wartości statystyki opisowej, takie jak średnia, odchylenie standardowe czy korelacja dla X-owych i Y-kowych współrzędnych punktów, są bardzo zbliżone do siebie dla wszystkich przedstawionych chmur. Przykład przedstawiony na rys. 1 pokazuje, że w analizie danych bardzo istotna jest możliwość zobrazowania – wizualizacji danych. Tutaj każdy punkt chmury ma tylko dwie współrzędne, zatem cała chmura ma dwa wymiary. Możemy więc łatwo wygenerować jej obraz. Ale co zrobić, jeśli rozważana chmura ma dziesiątki, setki, a może nawet tysiące wymiarów? Rozważania można przeprowadzić na przykładzie figurki kota Maneki-neko (rys. 2). Wybranej figurce

2

PAWEŁ DŁOTKO

Kot Maneki-neko. Poznamy kształt zbioru jego zdjęć wykonanych pod różnymi kątami

3



Idea algorytmu BM dla przykładowej chmury punktów

wykonano serię czarno-białych zdjęć o wymiarach 128×128 pikseli, obchodząc obiekt dookoła, pod różnymi kątami i z tej samej odległości.

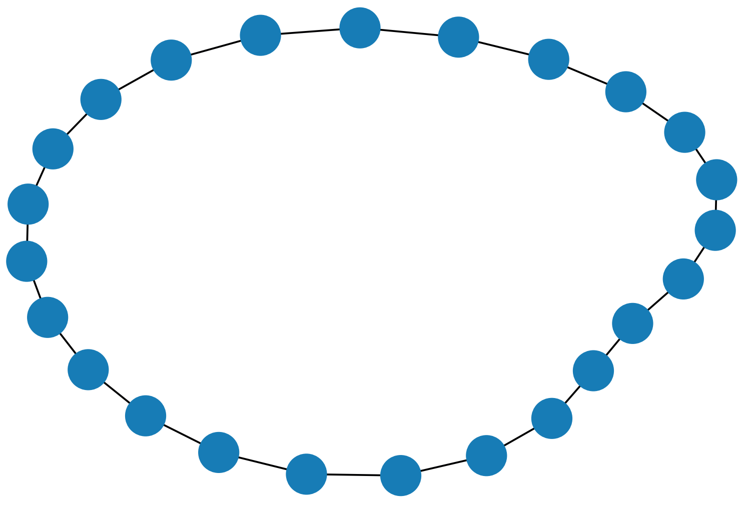
Piksele pojedynczego zdjęcia ponumerowane od 1 do 16 384 (128×128) tworzą punkt w 16 384-wymiarowej przestrzeni, tak że każde zdjęcie ma przypisany pojedynczy punkt. Kolejne współrzędne takiego punktu odpowiadają poziomowi jasności (skala szarości) pikseli danego zdjęcia. Zebranie wszystkich wykonanych zdjęć i przypisanie im odpowiadających punktów daje chmurę punktów. Kształt takiej chmury można zwizualizować za pomocą algorytmów mappera.

Rysunek 3 przedstawia ideę jednego z nich, algorytmu Ball Mapper, dalej zwanego BM. Przedstawiona jest ona na przykładzie dwuwymiarowej chmury punktów w kształcie litery Y (rys. 3a).

Chmury punktów można przedstawić w formie abstrakcyjnych grafów. W pierwszym kroku algorytmu BM z chmury punktów jest wyodrębniany jej równomiernie rozłożony podzbiór zaznaczony kolorowymi kropkami (rys. 3b). Te kolorowe kropki stanowią centra kul zawierających w swoim obrębie wszystkie punkty z zadanej chmury (rys. 3c). Mając kolekcję kul, budujemy graf (rys. 3d). W tej konstrukcji ważne jest to, że wierzchołki nie mają lokalizacji w przestrzeni – są one bytem abstrakcyjnym. W analizowanym przypadku wierzchołki odpowiadają kulom przedstawionym na rys. 3c, krawędzie reprezentują części wspólne kul. Przedstawiona konstrukcja może być zastosowana do chmury punktów o dowolnych wymiarach. W przypadku zdjęć figurki kota Maneki-neko po przejściu całego procesu uzyskujemy poniższy rysunek przedstawiający graf będący zamkniętym cyklem – ścieżką, po której możemy w kółko przeskakiwać z wierzchołka na kolejny wierzchołek (rys. 4).

Wymiar i kształt wybranego grafu przedstawia naturę naszych danych. Każde z wykonanych zdjęć stanowi punkt w 16 384-wymiarowej przestrzeni. Każde dwa kolejne zdjęcia są do siebie podobne, co sprawia, że odpowiadające im punkty są bliskie. Kiedy jednak zaczynamy od pierwszego zdjęcia, każde kolejne zaczyna się coraz bardziej od niego różnić, a po wykonaniu prawie pełnego obrotu, ponownie staje się do niego podobne. Tak samo dzieje się z odpowiadającymi zdjęciom punktami. Najpierw oddalają się one od punktu początkowego, by ostatecznie powrócić do niego z innej strony. W ten sposób w wysoko wymiarowej przestrzeni wracamy do punktu wyjścia po przejściu okręgu.

4



Obraz punktów odpowiadających zdjęciom kota Maneki-neko, zawartych w 16 384-wymiarowej przestrzeni

Zastosowania funkcji

Rozważmy sytuację, gdy każdy punkt chmury ma dodatkowo przypisaną wartość pewnej funkcji. Średnia wartość takiej funkcji dla punktów z kuli może zostać zwizualizowana za pomocą skali barw na wierzchołkach grafu. Dobrym przykładem tego procesu są banknoty, które tworzą zbiór danych składających się z czterech cech (cechy są dziedziną funkcji). Każdy z banknotów może być prawdziwy (wartość funkcji 0) albo fałszywy (wartość funkcji 1), co reprezentuje obraz funkcji. Przekształcenie tego zbioru danych za pomocą algorytmu BM do postaci grafu daje nam zbiór w kształcie litery Y (rys. 5).

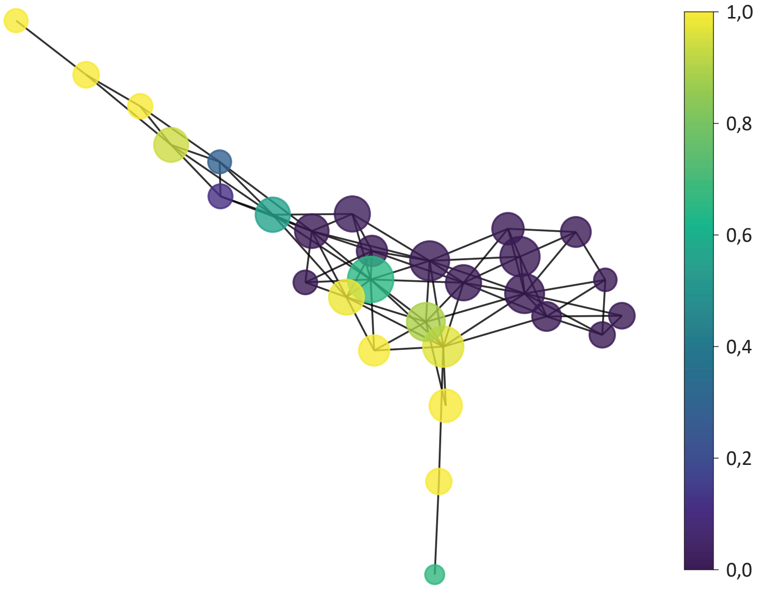
Wprowadzony na grafie kolor niesie informacje o banknotach: fioletowy – prawdziwe, żółty – fałszywe. Widoczne na grafie dwa żółte ramiona sugerują, że mamy do czynienia z dwoma różnymi typami fałszywych banknotów. Fakt ten można zaobserwować dzięki kształtowi przedstawionej chmury, gdzie wierzchołki zaznaczone na żółto nie są z sobą bezpośrednio powiązane krawędziami i znajdują się w przestrzeni oddalonej od siebie.

Dane wielowymiarowe są bardzo powszechne, a umiejętność ich wizualizacji kluczowa w wielu dziedzinach życia. Rozważmy na koniec klasyczny zbiór danych Holenderskiego Instytutu Badań nad Rakiem opisujący aktywność genów w grupie pacjentek z rakiem piersi. Wysokowymiarowe punkty naszego zbioru opisują aktywność tysiąca wybranych genów dla każdej z pacjentek. Naszym celem jest prognoza rozwoju choroby oraz zaproponowanie skutecznej terapii celowanej. Klasyczny algorytm mappera (opracowany w Stanfordzie w 2007 roku) tworzy obraz aktywności tysiąca genów z funkcją będącą czasem przeżycia pacjentek.

Podobnie jak dla banknotów uzyskana dziedzina ma kształt litery Y. Interesujące są dwie odnogi po lewej stronie grafu (rys. 6). Odnoga idąca ku dołowi odpowiada nowotworom *triple-negative* o złych prognozach i niskim czasie przeżycia. Odnoga idąca ku górze ma znacznie lepsze własności: przy jej końcu znajdziemy, oznaczoną na czerwono, wcześniej nieznaną podgrupę pacjentek o pełnej przeżywalności. Prowadząc dalszą analizę, możemy wywnioskować, że ta podgrupa jest charakteryzowana przez nowotwory z receptorem estrogenu, wysokim poziomem ekspresji genu c-MYB oraz niskim poziomem ekspresji genów „nieswoistej odpowiedzi odpornościowej". Kombinacja tych trzech parametrów jest obecnie używana, również w Polsce, do planowania terapii. Wykrycie tej podgrupy pacjentek było możliwe dzięki technikom tworzenia obrazów wielowymiarowych danych.

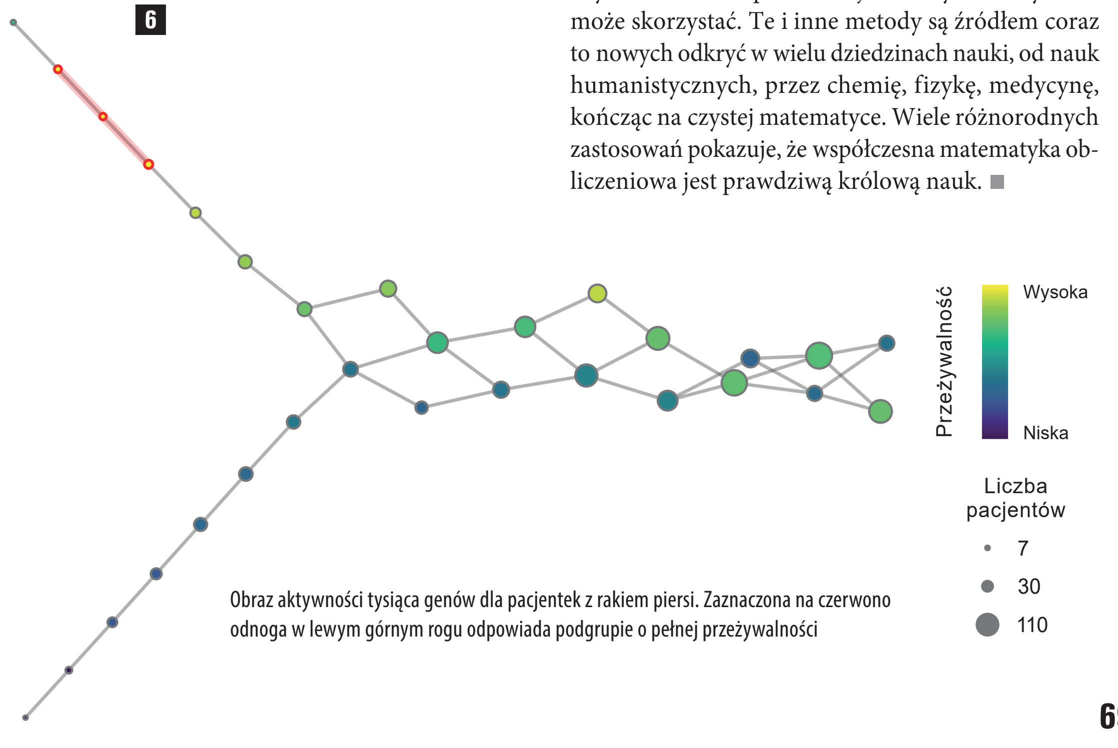
Przykłady wysokowymiarowych danych i ich analizy można mnożyć. Mimo że metody współczesnego nauczania maszynowego i sztucznej inteligencji dają możliwości ślepej analizy danych, umiejętność tworzenia i bezpośredniej analizy ich obrazów daje dodatkowy poziom zrozumienia rozważanego problemu i pozwala naukowcom obiektywnie wytłumaczyć obserwowane zjawiska. Metody matematyczne nie mogą obejść się bez efektywnych i przyjaznych dla użytkowników implementacji, z których każdy z nas może skorzystać. Te i inne metody są źródłem coraz to nowych odkryć w wielu dziedzinach nauki, od nauk humanistycznych, przez chemię, fizykę, medycynę, kończąc na czystej matematyce. Wiele różnorodnych zastosowań pokazuje, że współczesna matematyka obliczeniowa jest prawdziwą królową nauk. ■

5



Czterowymiarowa przestrzeń z charakterystyk prawdziwych (fiolet) oraz fałszywych (żółć) banknotów

6



Obraz aktywności tysiąca genów dla pacjentek z rakiem piersi. Zaznaczona na czerwono odnoga w lewym górnym rogu odpowiada podgrupie o pełnej przeżywalności

Chcesz wiedzieć więcej?

Tomala L., *Matematycy przychodzą na pomoc medykom, inżynierom, socjologom*, https://naukawpolsce.pl/aktualnosci/news%2C87009%2Cmatematycy-przychodza-na-pomoc-medykom-inzynierom-socjologom.html.