

Grzegorz Osiński *

Wyższa Szkoła Kultury Społecznej i Medialnej w Toruniu

WIZUALIZACJA INFORMACJI. BADANIA STRUKTUR INFORMACJI W POSZUKIWANIU PRAWDY

*„Gdzie mądrość utracona w wiedzy?
Gdzie wiedza utracona pośród wiadomości?”*

(T.S. Eliot, *The Rock*)

Informacja, wiedza i nauka

Współczesny rozwój technologii komputerowych oraz związana z tym bardzo dynamiczna rozbudowa światowej sieci komputerowej była możliwa dzięki badaniom nad, wydawałoby się, bardzo abstrakcyjnymi zagadnieniami dotyczącymi odpowiedniego zdefiniowania pojęcia informacji. Powstały już odrębne dziedziny nauki, które zajmują się tą problematyką w krajach anglosaskich, znane pod nazwą *Computer Science*, a w Polsce jako *informatyka*. Ta ostatnia nazwa lepiej oddaje historyczne i metodologiczne obszary zainteresowań badawczych, chociaż jest potocznie kojarzona tylko z problematyką technologiczną. Technologia odegrała istotną rolę w rozwoju informatyki, ale nie najważniejszą. Bez ściśle określonej definicji informacji oraz bez odkrycia, że język matematyki może być wykorzystany w procesie tworzenia i przetwarzania komunikatów wyrażanych w językach naturalnych, komputery zapewne nigdy nie opuściłyby laboratoriów badawczych i pozostały domeną naukowców, odizolowane od problemów współczesnego świata.

Stało się jednak inaczej. Najważniejszą rzeczą, która pozwoliła wkroczyć obliczeniom numerycznym w strefę zarezerwowaną do tej pory tylko dla człowieka, obejmującą wyrażanie myśli za pomocą „słowa” w języku naturalnym, była praca George’a Boole’a, profesora z Queen’s College w Cork w dzisiejszej Irlandii. Ten genialny samouk wykazał w swoim wydanym w 1854 r. studium

* **Grzegorz Osiński** – doktor, informatyk, kognitywista, fizyk kwantowy. Wykładowca w Wyższej Szkole Kultury Społecznej i Medialnej w Toruniu. Naukowo zajmuje się analizą nieliniowych modeli chaotycznych w naukach medycznych i psychologii eksperymentalnej. W latach 2001-2002 stypendysta National Research Council w laboratoriach NASA/JPL w Pasadenie w Kalifornii. Autor licznych publikacji naukowych, współautor popularnych interdyscyplinarnych pozycji książkowych m.in. „Chaos w głowie. Mózg i życie”, Aspektpress 2012 oraz portalu wizualizacjainformacji.pl.

pt. „Badanie praw myślenia, na którym oparte są matematyczne teorie logiki i prawdopodobieństwa”¹, że prawa logiki Arystotelesa mogą być przedmiotem obliczeń. Opracował on podstawowe elementy struktury języka matematycznego, który pozwala określać, czy zdania tworzone w języku naturalnym są prawdziwe. Ta swoista „matematyzacja” języka naturalnego była przełomowa w ówczesnym czasie. Większość ludzi uważała, że matematyka to tylko „arytmetyczne liczenie” – bardzo przydatne w prowadzeniu interesów i inżynierii, ale zupełnie rozdzielne z prawdziwą humanistyczną nauką. Praca Boole’a wykazała, że zdania można również „liczyć” i im przyporządkowywać numeryczne wartości logiczne. Dzisiaj elementy algebry Boole’a są już nauczane w szkole, więc dla studentów wydają się czymś zupełnie naturalnym, wręcz aksjomatycznym. A przecież stanowiły one przez wiele lat przełomu XIX i XX wieku, obiekt burzliwej dyskusji w świecie naukowym. Zastanawiano się i pytano, czy wyniki obliczeń matematycznych mogą odnosić się bezpośrednio do rzeczywistości, czy też są one jedynie „wewnętrznym” elementem systemu wykorzystywanego do obliczeń. Upraszczając znacznie problem, stawiano pytanie, czy wyniki obliczeń teoretycznych mogą odnosić się bezpośrednio do obserwowanej rzeczywistości, czy nie są one tylko przybliżeniem, czyli stanowią niepełny obraz, zatem nie można uznać, że są prawdziwe. Tak stawiane pytania były spowodowane pewnym nieporozumieniem formalnym, polegającym na traktowaniu obliczeń matematycznych jako deterministycznego procesu mechanicznego. Natomiast system logiki Boole’a traktuje tak naprawdę o języku matematyki jako pełnoprawnym medium do opisu rzeczywistości, podobnie jak języki naturalne.

Procedury numeryczne, stosowane np. w symulacji zjawisk naturalnych, zostały uznane jako pełnoprawny trzeci paradygmat naukowy – oprócz metody teoretycznej i empirycznej – w latach siedemdziesiątych ubiegłego wieku². Zakończyło to właściwie spory proceduralne i przyczyniło się do rozpowszechnienia metod obliczeniowych we wszystkich dziedzinach nauki. W szczególności należy zwrócić uwagę na zastosowanie metod numerycznych w badaniach lingwistycznych, dotyczących nie tylko biernej analizy języka naturalnego³, ale przede wszystkim umożliwiających aktywne symulowanie struktur lingwistycznych i badanie procesów ich historycznego powstawania. Języki naturalne to przecież

¹ G. Boole, *An Investigation into The Laws of Thought on Which Are Founded The Mathematical Theories of Logic and Probabilities*, Cork 1853, s. 174.

² T. Hey, S. Tansley, K. Tolle, *The Fourth Paradigm. Data-Intensive Scientific Discovery*, Redmond, Washington 2009, s. 210.

³ J. Szymański, W. Duch, *Representation of Hypertext Documents Based on Terms, Links and Text Compressibility*. Proceedings of Neural Information Processing. Theory and Algorithms, 17th International Conference, ICONIP 2010, Sydney, Australia, November 21-25, 2010, Proceedings, Part I, Sydney 2010, s. 283.

żywe, dynamiczne struktury, które oprócz funkcji komunikacyjnych posiadają wiele istotnych własności⁴. Jednak dla nas najistotniejsze będzie teraz zrozumienie, w jaki sposób informacja zawarta w zdaniach może stać się fundamentem, na którym budowana jest wiedza. Chociaż dzisiaj ten problem nie jest widoczny tak bardzo jak przed rozpowszechnieniem technologii komputerowych, to nie tylko nie stracił on swojej ważności, lecz wręcz przeciwnie: stał się bardziej istotny, choć mniej zauważalny. Przyczyna tego jest prosta. Obecnie dostępność do źródeł informacji jest po prostu powszechna. Bez wychodzenia z domu możemy na naszym komputerze podłączonym do sieci internetowej znaleźć i przeczytać praktycznie każdy artykuł i książkę, które znajdują się we wszystkich bibliotekach świata. Informacja jest na wyciągnięcie ręki, a właściwie palca wędrującego po klawiaturze bądź ekranie tabletu. Wydaje nam się, że skoro mamy dostęp do informacji, to od razu jesteśmy w stanie posiąść wiedzę. To złudne wrażenie. Doskonale wiemy, że możemy przeczytać kilkadziesiąt stron instrukcji obsługi nowego telefonu komórkowego, który właśnie zakupiliśmy, a nadal nie potrafimy prawidłowo skonfigurować usługi poczty elektronicznej. Z pewnością lepiej zauważymy ten problem, kiedy zaczniemy czytać dzieła Platona. Czy po jednokrotnym przeczytaniu, czyli przyswojeniu informacji, posiadamy już wiedzę na temat filozofii Platona?

Proces tworzenia wiedzy jest dużo bardziej skomplikowany; oczywiście wymaga on najpierw dużej porcji dobrej jakościowo informacji – jest to warunek konieczny, ale niewystarczający. Niestety, dzisiaj studenci czasami sprawiają wrażenie, jakby nie zauważali tego problemu; wydaje im się, iż wystarczy odnaleźć właściwą informację w sieci i jest to jednoznaczne z posiadaniem wiedzy na jakiś temat. Czasami zapominają nawet, że odnalezioną informację trzeba „przeczytać ze zrozumieniem”, pozostają na etapie odnalezienia informacji. W dawnych czasach, kiedy informację traktowano jako „cenną zdobycz” i aby przeczytać rzadką książkę należało odbyć podróż do odległej biblioteki, sytuacja była zgoła odmienna. Student bądź naukowiec doceniał samą wartość informacji i rozpoczynał od razu „studia” nad odnalezionym tekstem.

Wiedza powstaje w długotrwałym i żmudnym procesie studiowania, i to za zwyczaj różnych źródeł i dostępnych informacji. Aby z przyswojonej informacji zbudować gmach wiedzy, potrzebujemy nakładu pracy i czasu, niekiedy nawet wielu lat. Dlatego tak bardzo dziwią coraz częściej pojawiające się w przestrzeni publicznej wypowiedzi różnych osób, które na pytanie dotyczące konkretnego zdarzenia odpowiadają: „Nie mam wiedzy na ten temat”. Powinno być oczywiste, że nie posiadamy wiedzy na jakiś temat – chyba że odbyliśmy na tym polu długotrwałe studia. W takiej sytuacji zazwyczaj chodzi o to, że nie mamy informacji na

⁴ W. Duch, P. Matykieicz, J. Pestian, *Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis*, „Neural Networks” 2008, 21 (10), s. 3.

jakiś temat lub po prostu nie wiemy. Politycy takie odpowiedzi traktują najczęściej jako swoiste alibi, kiedy pytanie dotyczy aktualnych wrażliwych tematów; wolą odpowiedzieć, że nie mają wiedzy, bo gdyby publicznie przyznali się, że posiadają informację, ale nic z nią nie uczynili, byłoby to dla nich stwierdzenie obciążające. Lecz kiedy mówią, że nie posiadają wiedzy, to w przyszłości, kiedy prawda zostanie ujawniona, zawsze mogą sprostować, iż informacje posiadali, ale wiedzy na ich podstawie sobie nie wyrobili. Opisany przykład pokazuje, jak poważny będziemy mieli problem, kiedy zaczniemy zastanawiać się nad właściwym wykorzystaniem metod wizualizacji informacji w celu nie tylko przekazania odpowiedniego zasobu informacji, ale przede wszystkim umożliwienia odbiorcom zdobywania wiedzy na jakiś temat.

Kiedy mamy do czynienia z sytuacją, w której jako specjaliści w jakiejś dziedzinie posiadamy rzeczywistą i dobrze usystematyzowaną wiedzę na konkretny temat, to winniśmy zadać sobie pytanie dotyczące mądrości. Wszak cechą wszelkich dociekań naukowych winna być mądrość, najlepiej ogólna, obejmująca całe społeczeństwo, chociaż nie zawsze taki stan rzeczy odpowiadałby politykom. Czy mądrość polega więc na tym, że chwyta się istotę rzeczy? I czymże owa „istota” jest? Czy tylko konstruktem umysłu powstałym w wyniku oglądu i analizy, czy też realnym składnikiem otaczającej rzeczywistości? Jeśli analizowaną informację słowną bądź obraz wizualny potraktujemy jako zbiór składników, to czy jego istotą będzie znalezienie elementu najważniejszego, czy też odnalezienie struktury powiązań między nimi, która mówi więcej niż sam przekaz? Czasami nazywamy taki proces analizą kontekstu. Nie możemy podejmować dalszej dyskusji nad problematyką procesu powstawania wiedzy bez odwołania się do podstawowych definicji aksjologicznych. Wszak mądrość winna powstawać na drodze poszukiwania prawdy. W dziedzinie nauk technicznych podstawowym narzędziem stosowanym w opisie rzeczywistych obiektów jest stosowanie dobrze zdefiniowanych procedur logicznych, które na gruncie informatyki możemy identyfikować z pojęciem algorytmu. Takie podejście wymaga stosowania zarówno procedur numerycznych w odniesieniu do relacji pomiędzy obiektami, jak i logicznych w stosunku do definiowanych zdań.

Dochodzimy w tym momencie do podstawowego problemu dotyczącego możliwości obliczania każdego problemu w ramach dostępnego systemu języka matematyki. Są to zagadnienia szczególnie ważne dzisiaj, kiedy debata naukowa zaczyna wkraczać na tereny dotyczące połączenia i współdziałania bezpośredniego komputerów i człowieka. Chociaż zagadnienia dotyczące sztucznej inteligencji, symulacji procesów zachodzących w ludzkim mózgu, czy też coraz silniejszego obecnie nurtu transhumanizmu⁵ znacznie wykraczają poza ramy

⁵ Transhumanizm – ideologia postulująca użycie nauki do przezwyciężenia ludzkich ograniczeń i stworzenia idealnego „nadczołowieka”.

niniejszego artykułu, warto również na nie zwrócić uwagę. Ważnym problemem byłoby wykazanie, że język matematyki jest zupełny i spójny. Jednak w 1931 r. austriacki naukowiec Kurt Gödel⁶ udowodnił twierdzenie o niezupełności. Wskazuje on jednoznacznie, że w aksjomatycznej niesprzecznej teorii matematycznej, zawierającej pojęcie liczb naturalnych, da się sformułować takie zdanie, którego w ramach tej teorii nie da się ani udowodnić, ani obalić⁷.

Konsekwencją tego faktu jest stwierdzenie, że matematyki nie można w całości i ściśle zaksjomatyzować. Jest ona – podobnie jak język, którym się posługuje – systemem otwartym. Nie da się więc nigdy napisać takiego programu komputerowego, nawet gdybyśmy mieli do dyspozycji olbrzymie moce obliczeniowe, który mógłby jednoznacznie udowodnić wszystkie problemy matematyczne. Zatem to ograniczenie w sposób naturalny umożliwia z jednej strony twórcze korzystanie ze zdobyczy technologii komputerowych, z drugiej jednak pozostawia tylko w gestii człowieka rozstrzyganie podstawowych problemów natury moralnej i etycznej. Możemy zatem teraz spojrzeć na problem wiedzy z innej perspektywy.

Wiedza w pojęciu bardzo ogólnym kojarzy się nam nieodłącznie z nauką. Jednak współczesna nauka znacznie odbiega od uniwersalnych wzorców panujących od wieków w społeczeństwach europejskich. Postać naukowca – człowieka, który poświęcił swoje życie na odkrywanie prawdy o świecie, jest wypierana przez pojęcie „pracownika naukowego”, który koncentruje się na działalności w wąskiej, ściśle określonej aktywności badawczej, nierzadko bardzo odizolowanej od pozostałych dziedzin nauki.

Taki podział nauki na wąskie specjalizacje gwarantuje szybkie, spektakularne sukcesy w rozwiązywaniu szczegółowych problemów. Współcześnie taki sukces jest dość szybko przekładany na sukces komercyjny. Stąd zainteresowanie zarówno rządów, jak i instytucji finansowo-przemysłowych rozwijaniem i pogłębianiem takiego „podzielonego” świata nauki. Wąskie, zamknięte grupy naukowców łatwiej zarówno kontrolować, jak i motywować do konkretnych zadań badawczych. Czy jednak w tym systemie jest miejsce dla naukowców bezinteresownie poszukujących prawdy? Czy określenie i finansowanie pewnego rodzaju aktywności naukowej jako „badań podstawowych”⁸ – czyli z ekonomicznego punktu widzenia takich, które nie dają szybkich, bezpośrednich korzyści – wystarcza, aby nauka

⁶ Kurt Gödel (1906-1978) – austriacki logik i matematyk, twórca twierdzenia o niezupełności i niesprzeczności.

⁷ K. Gödel, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme. I*, „Monatshefte für Mathematik und Physik” 1931, nr 38, s. 174.

⁸ Badania podstawowe – oryginalne prace eksperymentalne lub teoretyczne podejmowane przede wszystkim w celu zdobycia nowej wiedzy o podstawach zjawisk i obserwowalnych faktów bez nastawienia na bezpośrednie zastosowanie praktyczne, Dz.U. z 2010 r., nr 96, poz. 615. Ustawa z dnia 30 IV 2010 r. o zasadach finansowania nauki, art. 2 pkt 3 lit. a.

mogła koncentrować się na zagadnieniach natury ogólnej, na obszarach dotąd niedostępnych i niepoznanych? Aby odpowiedzieć na tak postawione pytania, powinniśmy najpierw przyrzeć się nauce światowej w sposób całościowy. Aby to uczynić, musimy zastosować specjalistyczne komputerowe algorytmy wizualizacji dużych zbiorów danych. Proces ten będzie dokładnie opisany w części poświęconej metodom wizualizacji informacji.

Zagadnienia związane z rozpoznaniem struktur budowy współczesnej aktywności naukowej znajdują się obecnie w centrum zainteresowania wielu ośrodków akademickich, niekoniecznie tych dominujących w różnorodnych rankingach tworzonych przez specjalistyczne agencje i urzędy. Problemem jest nie tylko zebranie dokładnych danych dotyczących aktywności badawczej, ale przede wszystkim określenie, czym jest dzisiaj działalność naukowa. Z jednej strony mamy do dyspozycji rozbudowane i skomplikowane systemy oceny naukometrycznej, bazujące głównie na wartościach specyficznych indeksów (np. Impact Factor⁹ i Indeks Hirscha¹⁰) obliczonych według algorytmów odnoszących się do globalnych baz danych artykułów naukowych (np. Web of Science¹¹, Scopus¹²), z drugiej zaś strony – z krytyczną oceną takiego stanu rzeczy dokonywaną przez wybitnych naukowców praktyków, takich jak Nassim Taleb¹³. Ten absolwent paryskiej Sorbony i pracownik wielu prestiżowych uczelni wprowadził do przestrzeni publicznej nowe spojrzenie na główny nurt nauki światowej: „Radziecko-harwardzkie kręgi intelektualne nie rozumieją zjawiska antykruchości, przez co rzadko dostrzegają opisaną asymetrię i nigdy o niej nie uczą”¹⁴. Określenie to ukazuje rzeczywiste powiązania w sposobie myślenia pomiędzy naukowcami pracującymi na czołowych amerykańskich uczelniach „ligi bluszczowej”¹⁵ a przedstawicielami marksistowskiej nauki uprawianej na uniwersytetach w byłym ZSRR. To typowe

⁹ Impact Factor – wprowadzony przez koncern wydawniczy Thomson Reuters indeks cytowań publikacji naukowych określający prestiż i siłę oddziaływania czasopism naukowych.

¹⁰ Indeks Hirscha – współczynnik wprowadzony w 2005 r. mający w zamierzeniu wykazać wagę i znaczenie wszystkich prac naukowych danego autora.

¹¹ Web of Science – komercyjna naukowa baza danych koncernu Thomson Reuters określająca indeks cytowań artykułów zawartych w bazie.

¹² Scopus – komercyjna naukowa baza danych koncernu wydawniczego Elsevier.

¹³ Nassim Nicholas Taleb – profesor New York University Polytechnic School of Engineering oraz Uniwersytetu Oksfordzkiego. Specjalista w zakresie systemów dynamicznych, strategii ryzyka oraz heurystycznych modeli ekonomicznych.

¹⁴ N. Taleb, *Antykruchość. O rzeczach, którym służą wstrząsy*, Warszawa 2013, s. 24.

¹⁵ Liga bluszczowa (ang. *Ivy League*) – związek najbardziej prestiżowych uczelni amerykańskich: Brown University, Columbia University, Cornell University, Dartmouth College, Harvard University, Princeton University, University of Pennsylvania oraz Yale University.

połączenie na polu tzn. nowej lewicy¹⁶ przedstawicieli ideologii marksistowskiej jest widoczne obecnie w wielu dziedzinach aktywności społecznej. Dlaczego nauka miałaby być wyjątkiem? Pamiętając o tym, iż budżet „ligi bluszczowej” szacowany jest sumarycznie na 114 miliardów dolarów¹⁷, powinniśmy zrozumieć, dlaczego model nauki uprawiany na tych uczelniach stał się dominujący w świecie. Nassim Nicholas Taleb, jako autor tego określenia, wie najlepiej, jak to wygląda w praktyce, gdyż sam pracował przez wiele lat na owych czołowych uczelniach oraz uczestniczył w podejmowaniu decyzji dotyczących finansowania badań naukowych i rozwojowych, będąc m.in. konsultantem w agendach Międzynarodowego Funduszu Walutowego. Jego krytyka nie może pozostać niezauważona, a i dalsza dyskusja na temat kształtu współczesnej nauki powinna brać pod uwagę jego argumenty. Musimy również pogodzić się z faktem, że analiza problemów współczesnej nauki nie może się odbywać z zupełnym pominięciem aspektów ekonomicznych.

Teoria informacji

Już sama funkcja komunikacyjna języka przysporzyła naukowcom pracującym nad stworzeniem użytecznej teorii komunikacji wielu problemów. Najważniejszym był problem z ilościowym określeniem zawartości informacji w zdaniu wypowiedzianym w języku naturalnym. Wiemy przecież z życia codziennego, szczególnie słuchając przemówień polityków, że można powiedzieć wiele zdań, a nie przekazać żadnej istotnej informacji. W jaki sposób określić wartość użytecznej informacji zawartej w przekazie słownym? Tymi zagadnieniami zajmują się specjaliści z zakresu teorii informacji. Powszechnie uważa się, iż twórcą współczesnej teorii informacji jest amerykański matematyk i inżynier Claude Shannon¹⁸. Jako inżynier zajmował się on budową i zarządzaniem sieciami telefonicznymi; ponieważ był zafascynowany omówioną wcześniej koncepcją algebry Boole’a, zastosował jego koncepcje w praktyce, wprowadzając do użycia przekaźniki elektryczne. Te proste urządzenia można potraktować jako dwustanowe kontrolery, które mogą przyjmować dwie logiczne wartości: 1 – kiedy są włączone, oraz 0 – kiedy nie przewodzą prądu. Prace nad tymi zagadnieniami naprowadziły go na pomysł, aby za najmniejszą miarę informacji przyjąć właśnie

¹⁶ M. Kacprzak, *Odczytanie marksizmu przez Antonio Gramsciego i jego znaczenie dla współczesnej kultury*, „Fides, Ratio et Patria. Studia Toruńskie” 2014, nr 1.

¹⁷ Całkowity majątek uczelni „ligi bluszczowej” określony w roku 2014 wynosi ok. 114 miliardów USD. W Princeton University, najbogatszym z nich, w przeliczeniu na jednego studenta przypada 2,7 miliona USD. Dla porównania całkowity budżet nauki w Polsce w roku 2012 to ok. 2,2 miliarda USD, <http://www.sprawynauki.edu.pl/> (dostęp: 2 XI 2015).

¹⁸ Claude Elwood Shannon (1916–2001) – amerykański matematyk i inżynier, profesor MIT.

ową minimalną wartość logiczną, którą nazwał *bitem*. Podstawy swojej teorii opisał w opublikowanej w 1948 r. pracy pt. „Matematyczna teoria komunikacji”¹⁹. Shannon podaje w niej nie tylko podstawowe definicje, ale również wyrażenia matematyczne pozwalające obliczać parametry kanałów przepływu informacji:

$$H = -\sum p_i \log_2 p_i ,$$

gdzie H jest miarą informacji, a p_i to prawdopodobieństwo każdej informacji składowej. Taka definicja informacji nie tylko pozwoliła na szybki rozwój technologii obliczeniowych, ale również umożliwiła rozpoczęcie prac teoretycznych nad analizą zasobów informacyjnych. Dodatkowo wprowadzenie swego kodowania binarnego²⁰ – czyli przyporządkowania literom alfabetu odpowiednich wartości numerycznych – pozwoliło na ilościowe określenie wielkości niepewności informacji, zwanej entropią²¹. W ten sposób otrzymaliśmy do ręki potężne narzędzie, które łącząc w sobie koncepcje z różnych dziedzin wiedzy, opisuje językiem matematyki konkretny sposób analizy strumienia informacji, który pierwotnie może być zapisany w dowolnym języku naturalnym. Shannon badał teksty zapisane w języku angielskim i stwierdził, że redundancja²² dla tego języka wynosi nawet 75 procent. Owa nadmiarowość wynika oczywiście z naturalnych cech języka; pozwala ona na unikanie błędów w czasie przekazu, ale również otwiera możliwości kodowania i kompresji²³ przy użyciu algorytmów numerycznych. W tym miejscu warto zaznaczyć, iż badania nad teorią informacji nie były domeną tylko naukowców z krajów anglosaskich. Co prawda zarówno pierwsze komputery (ENIAC, Colossus) oraz systemy przetwarzania informacji zostały zbudowane w USA i Wielkiej Brytanii, jednak ich powstanie byłoby niemożliwe bez udziału naukowców z innych krajów, również z Polski. O ile aktywność naukowa znakomitych polskich matematyków z przedwojennej lwowskiej szkoły matematycznej, takich jak Stefan Banach i Stanisław Ulam,

¹⁹ Praca ukazała się najpierw w formie artykułu pt. *A Mathematical Theory of Communication*, a w 1949 r. jako mała książeczka *The Mathematical Theory of Communication*, University Illinois Press 1949.

²⁰ ASCII (*American Standard Code for Information Interchange*) – siedmiobitowy kod przyporządkowujący liczby z zakresu 0–127 literom alfabetu angielskiego, cyfrom, znakom przestankowym i innym symbolom.

²¹ Entropia – pojęcie funkcji stanu używane zarówno w fizyce, jak i teorii informacji, gdzie jest definiowana jako średnia ilość informacji przypadająca na pojedynczą wiadomość ze źródła informacji.

²² Redundancja (łac. *redundantia* – powódź, nadmiar) to ilość informacji przekraczająca minimum wartości wymagane do rozwiązania problemu.

²³ Kompresja – polega na zmianie sposobu zapisu informacji tak, aby zmniejszyć redundancję.

została już dość dobrze opisana w literaturze²⁴, o tyle wkład Polaków w prace nad teorią informacji nadal jest niedoceniony. Mowa tutaj o Marianie Mazurze²⁵, który winien być uznany na równi z Claude'em Shannonem za współtwórcę podstawowych koncepcji teorii informacji. Podobnie jak Shannon, zajmował się on w 1937 r. automatyzacją połączeń telefonicznych na linii Warszawa–Katowice i również postanowił zastosować w tym celu automatyczne przekaźniki. Jednak w Polsce w 1939 r. rozpoczęła się wojna i Mazur musiał zaprzestać jakichkolwiek prac badawczych i technicznych. Skupił się jednak na pracach teoretycznych, bo tylko takie mógł prowadzić prywatnie w swoim mieszkaniu w okupowanej Warszawie. Tak powstała ponad 200-stronicowa praca dotycząca zaawansowanych teorii przesyłania informacji; niestety, w sierpniu 1944 r. manuskrypt spłonął w trakcie Powstania Warszawskiego. Nie było wtedy mowy o jakichkolwiek publikacjach, czy nawet tworzeniu kopii dzieła. Straciliśmy je zatem bezpowrotnie. Zaraz po wojnie Marian Mazur nie miał żadnej możliwości kontynuowania swojej pracy w zniszczonej Warszawie, a kiedy udało mu się w końcu w 1948 r. wrócić do aktywności zawodowej w Instytucie Elektrotechniki w Warszawie, to praca Shannona już się ukazała drukiem.

Metody big data

W czasach kiedy tworzone podwaliny teorii komunikacji, powstawały też pierwsze komputery, ale aplikacyjne wykorzystanie metod teoretycznych ograniczało się do przesyłania i analizy krótkich komunikatów, zazwyczaj pisanych w języku maszynowym²⁶. Szybki rozwój technologii koncentrował się nie tylko na miniaturyzacji elementów elektronicznych, ale również na usprawnieniu komunikacji pomiędzy komputerem i człowiekiem. Wykorzystywane jeszcze w latach siedemdziesiątych ubiegłego wieku karty perforowane – których historia powstania i wykorzystania jest również bardzo zajmująca, ale nie stanowi przedmiotu niniejszego opracowania²⁷ – były powoli wypierane przez profesjonalne urządzenia wejścia-wyjścia, takie jak klawiatura i monitor. Komputery były wykorzystywane głównie do trudnych obliczeń numerycznych, do sterowania skomplikowanymi układami elektromechanicznymi oraz przeszukiwania dużych baz danych. Nas

²⁴ M. Urbanek, *Genialni. Lwowska szkoła matematyczna*, Warszawa 2014.

²⁵ Marian Mazur (1909–1983) – naukowiec zajmujący się elektrotermią i cybernetyką, twórca polskiej szkoły cybernetycznej.

²⁶ Język maszynowy – zestaw rozkazów procesora zaimplementowany sprzętowo, w którym zapis programu wyrażony jest w postaci liczb binarnych stanowiących rozkazy oraz ich argumenty.

²⁷ E. Black, *IBM i holokaust. Strategiczny sojusz hitlerowskich Niemiec z amerykańską korporacją*, Warszawa 2006.

najbardziej będzie interesowało to ostatnie zagadnienie. Problem wyszukiwania informacji jest równie istotny jak problem jej przesyłania, czyli komunikacji. Z życia codziennego doskonale wiemy, że im więcej danych zgromadzimy, tym dłużej musimy przeszukiwać zasoby, aby odnaleźć interesujące nas informacje. Komputery znakomicie przyspieszały wyszukiwanie informacji pod warunkiem właściwego skonstruowania struktur i relacji w bazach danych. Obecnie mamy do dyspozycji całe klasy specjalistycznych algorytmów służących do wyszukiwania, sortowania i klasyfikacji. Dopóki komputery pracowały jako jednostki autonomiczne, dane były gromadzone lokalnie przy jednej maszynie. Jednak kiedy połączone je w globalną sieć informatyczną, możliwe stało się zbieranie informacji bardzo różnorodnej, z różnych miejsc, i to w ogromnych ilościach. Ilość gromadzonych informacji była tak ogromna, że wykorzystanie klasycznych algorytmów stało się już nieefektywne. Prezentowanie wyników za pomocą tabel i prostych wykresów w takich analizach nie pozwala już na wnioskowanie całościowe, a wyniki odnoszące się tylko do poszczególnych podgrup danych nie są reprezentatywne dla całości. Analiza takich różnorodnych, obszernych i wzajemnie niezależnych baz danych jest określana obecnie jako problematyka *big data*²⁸. Aktualnie definiujemy takie zbiory w czterowymiarowej przestrzeni konfiguracyjnej 4V, gdzie odpowiednie wymiary określone są jako:

- *volume* – ilość danych, obecnie liczonych już w petabajtach²⁹;
- *variety* – dane pochodzą z różnych, niezależnych i niespójnych ze sobą źródeł informacji;
- *velocity* – szybkość napływania nowych danych i ich analiza w czasie rzeczywistym;
- *value* – wyodrębnianie, selekcja z ogromnej masy informacji tylko tych najważniejszych.

Taka metoda analizy jest już czymś więcej niż tylko prostym przetwarzaniem informacji; pozwala na odkrywanie nowych jakościowo informacji na podstawie tych już istniejących, spełnia więc kryteria omawianego wcześniej procesu tworzenia wiedzy. Często mówimy dzisiaj, że analizy typu *big data* pozwalają na odkrywanie nowej wiedzy z dostępnych różnorodnych struktur danych³⁰. Dobrym i powszechnie znanym przykładem zastosowania tej technologii jest jej wykorzystanie w usprawnieniu zarządzania tak dużym miastem jak Nowy Jork. Burmistrz tego miasta Mike Bloomberg w 2012 r. powołał specjalną jednostkę

²⁸ *Big data* – pojęcie odnoszące się do zbiorów danych, które ze względu na swoją ogromną skalę nie mogą być analizowane za pomocą tradycyjnych narzędzi informatycznych.

²⁹ Petabajt – to 10^{15} bajtów w odniesieniu do dzisiejszej pojemności dysków komputerowych. 1 petabajt = 1 milion gigabajtów.

³⁰ D.T. Larose, *Odkrywanie wiedzy z danych*, Warszawa 2013.

administracyjną zajmującą się tylko problematyką big data³¹. Na podstawie danych zebranych z różnych źródeł: z wydziału urbanistycznego, ze statystyk budynków miejskich, z wydziału ewidencji działalności gospodarczej, rozmieszczenia otwartych punktów dostępu wi-fi, ze statystyki przestępstw w poszczególnych regionach miasta, z liczby mieszkańców w danej dzielnicy i ich przynależności etnicznej, ze statystyk systemu edukacji publicznej oraz z danych urzędów podatkowych – dokonano specjalistycznej analizy big data. W jej efekcie udało się zidentyfikować m.in. punkty nielegalnej sprzedaży tytoniu. Dzięki temu policja zlikwidowała ogromną liczbę sklepów sprzedających nielegalnie wyroby tytoniowe. Oprócz tego usprawniono również system reagowania w sytuacjach kryzysowych, co zostało praktycznie sprawdzone podczas uderzenia huraganu Sandy³². Trzeba zauważyć, że zbiory danych wejściowych nie zawierały w sposób jawny informacji na temat spożycia i dostępności wyrobów tytoniowych, jednak specjalistyczna analiza doprowadziła do odkrycia nowej jakościowo informacji. To specyfika metody big data. Wykorzystał ją również sztab wyborczy Baracka Obamy podczas wyborów prezydenckich w 2012 r. Badanie preferencji wyborczych to obecnie domena właściwie całkowicie zdominowana przez technologie komputerowe; klasyczna kampania wyborcza staje się już tylko dodatkiem do prawdziwej wojny toczonej na serwerach dokonujących wymiany i analizy informacji. Tradycyjne sondaże przedwyborcze tracą swoją dokładność i proces ten będzie coraz bardziej widoczny, gdyż opierają się one tylko na badaniu jednej zmiennej, zazwyczaj w wyniku pomiaru bezpośredniego. Przecież doskonale wiemy, że ankietowany wcale nie musi być przekonany w momencie odpowiedzi, a jego motywacja i preferencje mogą zmienić się w bardzo krótkim czasie. Dlatego dzisiejsze modele symulacyjne analizujące preferencje wyborcze odróżniają tzw. twarde elektorat od osób niezdecydowanych. To tylko o tych ostatnich starają się politycy w kampaniach wyborczych i aby to robić skutecznie, muszą zdobyć jak najwięcej informacji na ich temat. Dlatego Obama³³ w trakcie swojej kampanii wyborczej zatrudnił ponad 1000 informatyków, specjalistów od analizy i wizualizacji danych, którym pomagało ponad 10 000 wolontariuszy zajmujących się zbieraniem różnorodnych danych, nawet tak egzotycznych, jak liczba sprzedanych hamburgerów w danej dzielnicy, czy też liczba rowerów znajdujących się na publicznych parkingach. Centrum komputerowe sztabu wyborczego przeprowadzało każdego dnia ponad 60 000 symulacji komputerowych i podejmowano decyzje na bieżąco na pod-

³¹ M. Kolanko, „Polityka oparta na Big Data”, <http://300polityka.pl/news/2013/03/03/polityka-oparta-na-big-data-jak-burmistrz-bloomberg-chce-zmienic-sposob-rzadzenia-nowym-jork> [dostęp: 14 VIII 2015].

³² K. Moisse, S. Lupkin, *Superstorm Sandy Tests Hospital Preparedness*. ABC News. Retrieved 3 XI 2012.

³³ www.barackobama.com [dostęp: 14 VIII 2015].

stawie wyników obliczeń. Na oficjalnej stronie internetowej prezydenta USA: www.barackobama.com, znajdziemy również jasne stwierdzenie, które warto przytoczyć w oryginale: *The objective of the campaign set out by Jim Messina was to "measure everything"*³⁴. Metoda zbierania nawet nieistotnych danych jest już coraz bardziej powszechna; umożliwia ją coraz częstsze używanie smartfonów, które tak naprawdę są bardzo zaawansowanymi komputerami, a ich właściciele praktycznie się z nimi nie rozstają. Mają one wbudowane bardzo liczne funkcje, a każdy użytkownik używa zazwyczaj tylko kilku z nich. I informacja o tym, z jakich funkcji użytkownik nie korzysta, jest dla operatora telefonicznego najcenniejsza, bo pozwala określić indywidualne cechy posiadacza aparatu i na podstawie analizy z innymi bazami danych, w których znajdują się dane klienta, można opracować indywidualną strategię marketingową. Podobne metody stosują banki i instytucje finansowe; korzysta z nich również policja i służby bezpieczeństwa. Nawet informacja o naszej bierności, czyli niekorzystaniu z różnych urządzeń i usług, ma swoją wartość. Dlatego tak ważna jest dbałość o ochronę danych osobowych. Bo jeśli dane o nas wykorzystują instytucje państwowe, to możemy mieć nadzieję, iż czynią to dla naszego dobra. Jednak to samo narzędzie i metody analizy w rękach np. terrorystów to olbrzymie niebezpieczeństwo. Ale to już tematyka wykraczająca poza ramy tego artykułu.

Wizualizacja informacji

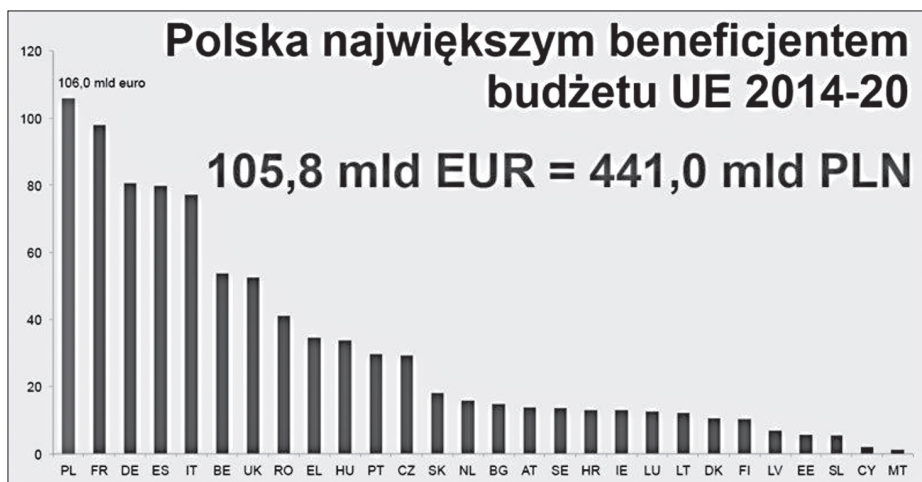
Jak już się przekonaaliśmy, prezentacja zbiorów danych dawno już przestała być prostą prezentacją wyników analiz, która redukuje i upraszcza przekaz informacji. Ten rodzaj komunikatu stosujemy raczej w celu osiągnięcia zamierzonego celu – manipulacji odbiorcą. Rozpatrzmy to na konkretnym przykładzie bardzo popularnego wykresu słupkowego. W lutym 2013 r. media podały niezwykle optymistyczną informację wypowiedzianą przez premiera rządu RP Donalda Tuska: *Dla mnie to jeden z najszcześniejszych dni w moim życiu. Nie sądzę, żebym mógł coś więcej dla Polski zrobić. Mamy budżet, w którym otrzymamy 72,9 mld euro na politykę spójności*³⁵.

Jednocześnie na stronach internetowych Kancelarii Premiera RP pojawił się szczegółowy raport dotyczący owego niebywałego sukcesu³⁶. Zaprezentowano w nim dużo diagramów wizualizacyjnych w postaci wykresów słupkowych. Pobieźna specjalistyczna analiza tych komunikatów od razu ukazuje całe spektrum klasycznych metod manipulacyjnych. Na rys. 1 przedstawiono sztandarowy ko-

³⁴ „Celem kampanii określonym przez Jima Messinę było »mierzyć wszystko«”.

³⁵ <http://www.forbes.pl/tusk-polska-najwiekszym-beneficjentem-budzetu-ue,artykuly,138696,1,1.html> [dostęp: 14 VIII 2015].

³⁶ https://www.premier.gov.pl/files/budzet_ue_prezentacja_0.pdf [dostęp: 7 XI 2015].

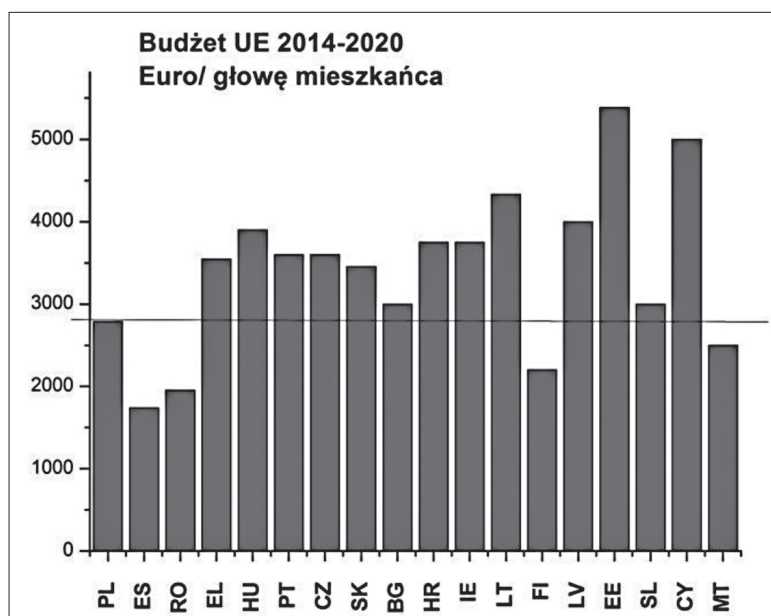


Rys. 1. Klasyczny przykład manipulacji; komunikat wizualny w postaci diagramu słupkowego pokazujący sukces negocjacyjny Polski w ustalaniu budżetu Unii Europejskiej na lata 2014–2015 [informacja publiczna Kancelarii Prezesa Rady Ministrów, <https://www.premier.gov.pl>]

munikat mający potwierdzać niewątpliwy sukces negocjacji. Należy zauważyć, iż kwoty finansowe odnoszą się do krajów członkowskich Unii Europejskiej i traktują je w sposób równy, bez jakiegokolwiek uwzględnienia ich indywidualnych charakterystyk. A przecież miliard euro dotacji dla małego kraju, np. Estonii, gdzie mieszka 1,3 miliona obywateli, ma dużo większe znaczenie niż ten sam miliard euro skierowany do dużego kraju, np. Rumunii zamieszkałej przez 20 milionów mieszkańców. Całkowita kwota dotacji nie jest więc żadnym miernikiem skuteczności dofinansowania. Należy podzielić kwotę całkowitą przez liczbę mieszkańców danego kraju i wtedy uzyskamy miarodajny parametr porównawczy. Oprócz tego na wykresie umieszczono wszystkie kraje członkowskie Unii Europejskiej, a przecież podstawą funduszu spójności jest wyrównanie szans krajów tzw. starej Europy w stosunku do nowych państw członkowskich.

Dlatego wielkość dofinansowania z tego funduszu dla krajów „starej Europy” ma być z założenia dużo niższa; te kraje korzystają z innych mechanizmów finansowych, niedostępnych dla wszystkich państw członkowskich. Jest to zabieg celowy, ma on wyrobić w odbiorcy przekonanie, iż najwyższy „słupek” na tle innych jest rzeczywiście dominujący. To efekt tła, mający na celu rozproszenie uwagi systemu percepcji wizualnej i wprowadzenie specyficznego dysonansu poznawczego³⁷. Kompletnie niezrozumiałe jest również umieszczenie na tym

³⁷ Dysonans poznawczy – w przypadku percepcji wizualnej stan aktywności psychomotorycznej pojawiający się u danej osoby wtedy, gdy jednocześnie występuje kilka elementów kognitywnie ze sobą niezgodnych.



Rys. 2. Komunikat graficzny z poprzedniego rysunku po usunięciu wizualnych elementów manipulacyjnych

wykresie Luksemburga; to kraj o specjalnym statusie, wybrany jako specyficzna enklawa ekonomiczno-społeczna, w której ulokowano największe instytucje finansowe z różnych krajów, aby uniknąć opodatkowania. To swoisty „raj finansowy”. Umieszczanie go na jednym wykresie z krajami o normalnej strukturze społeczno-finansowej jest błędem metodologicznym. Ale błędem celowym. Ma on pokazać iluzoryczną równość wszystkich podmiotów analizy. Jeśli znamy podstawy teoretyczne oraz metodologię wizualizacji informacji, możemy szybko wyeliminować elementy manipulacji z omawianego wykresu. Przeliczmy wartości dofinansowania na liczbę mieszkańców każdego państwa i usuniemy z wykresu kraje „starej Europy”. Poza tym zrezygnujemy również z elementu manipulacji kolorem i jednakowo zabarwimy wszystkie „słupki”. Wtedy uzyskamy zupełnie inny komunikat wizualny pokazany na rys. 2. Dla ustalenia punktu odniesienia wprowadzamy również poziomą linię, która wyraźnie ukazuje poziom finansowania dla Polski. I to jest prawdziwa reprezentacja wizualna informacji na temat dofinansowania. Ale na niej żadnego „sukcesu” negocyjacyjnego nie widać. Wręcz przeciwnie, okazuje się, że Polska otrzymała jedno z najniższych dofinansowań; jesteśmy na miejscu piątym od końca. Gorzej od nas wypadają tylko Rumunia, Estonia, Finlandia i Malta. Zważywszy na fakt, że dwa ostatnie kraje mają odmienną specyfikę, to tak naprawdę po brukselskich negocjacjach powinniśmy usłyszeć komunikat: „Kłęsa negocjacji – najmniej pieniędzy w nowym budżecie

UE otrzymały Polska i Rumunia”. Wykresy słupkowe są najbardziej podatną na manipulację formą prezentacji danych, dlatego profesjonalne serwisy informacyjne stosują je tylko w koniecznych przypadkach. Moje zalecenie jest bardziej restrykcyjne – nie używajmy ich wcale, gdyż istnieją już powszechnie dostępne metody profesjonalnej wizualizacji danych niepodatne na manipulację; wymagają one jedynie trochę większego nakładu pracy.

Aby uniknąć podobnych sytuacji, w poszukiwaniu ważnych zależności powinniśmy przekaz dostosować do aparatu percepcyjnego odbiorcy tak, aby on sam mógł dokonać twórczej analizy. Zamiast manipulować przekazem wizualnym przedstawmy jak najwięcej informacji na jednym obrazie i pozwólmy odbiorcy samodzielnie przeanalizować zagadnienie. Problemem jest wtedy sposób, w jaki mamy to uczynić, aby nie zaburzyć naturalnych zdolności i nie narzucać z góry ustalonej interpretacji. Aby tego dokonać, zastanówmy się najpierw, w jaki sposób człowiek odbiera informacje w postaci obrazu. Jest to najbardziej naturalny i najsprawniejszy element systemu percepcji człowieka. Zawartość informacyjna obrazu jest ogromna i jednocześnie jest on równolegle przetwarzany przez nasz mózg. Metody wizualizacyjne w postaci obrazu stosowane są obecnie we wszystkich prezentacjach ze szczególnym uwzględnieniem przekazu sieciowego oraz coraz liczniejszych urządzeń mobilnych. Jednak cały czas zbyt wielką rolę przypisuje się prezentacji danych w postaci tekstowej oraz w formie tabelarycznej. Jak się przekonałismy w przypadku metod big data, danych do prezentacji jest tak dużo, że trudno byłoby je zmieścić nawet na tysiącu dużych tabel zajmujących setki tysięcy stron. Również w omówionym poprzednio problemie manipulacji wykresem słupkowym prezentacja tabeli ze wszystkimi wartościami nie jest dobrym wyjściem. Takie prezentowanie danych byłoby zupełnie bezużyteczne, człowiek nie byłby w stanie nawet przeczytać takiej ilości danych, a co dopiero poddać ich analizie.

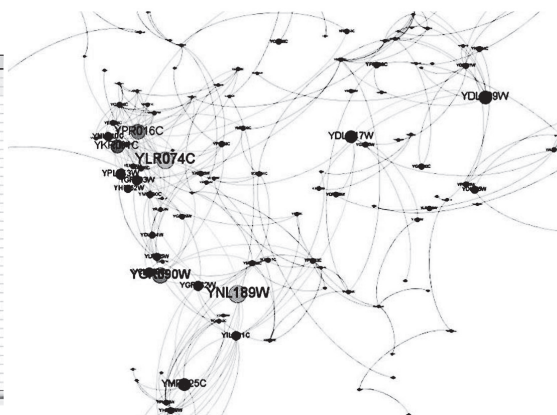
Zmieniając tematykę, przyjrzymy się teraz przykładowi z pola nauk biologicznych, dotyczącemu aktywności molekularnej drożdży. Są one wykorzystywane nie tylko w procesach stosowanych w przemyśle spożywczym, ale również przy produkcji leków oraz w przemyśle kosmetycznym. Stanowią więc doskonały przykład struktury big data, gdyż dane zbieramy niezależnie z różnych źródeł i możemy oczekiwać rezultatów w zupełnie innych dziedzinach wiedzy. Przeanalizujmy ten problem na dużym zbiorze danych zebranych w laboratorium bioinformatycznym Chińskiej Akademii Nauk w Pekinie³⁸. Badania dotyczą aktywności połączeń białkowych w strukturze drożdży. Zbiór obejmuje ponad 18 milionów elementów charakteryzujących sieć wzajemnych połączeń białkowych.

³⁸ D. Bu i in., *Topological structure analysis of the protein – protein interaction network in budding yeast*, „Nucleic Acids Research” 2003, vol. 31, No. 9.

W postaci tabeli, której fragment widzimy na rysunku 3, jest to struktura zupełnie nieczytelna; tak duża ilość elementów jest nawet trudna do przeglądania, nie wspominając o analizie. Aby zbadać taki zbiór danych, wykorzystamy darmowe otwarte oprogramowanie Gephi³⁹. Jest to jedno z najpopularniejszych narzędzi informatycznych stosowanych w metodach wizualizacyjnych. Po zastosowaniu algorytmów klasyfikacyjnych i wstępnej analizie klastrowej otrzymujemy wstępną wizualizację danych⁴⁰. Możemy zauważyć, że po oznaczeniu poszczególnych białek widać strukturę, która dla specjalisty jest typową „mapą aktywności”. Na jej podstawie można już podjąć decyzję o dalszych badaniach, analizach, a nawet eksperymentach mających na celu dokładniejsze zbadanie struktur aktywności białkowej i ich praktycznego wykorzystania.

Obraz stał się najistotniejszym wynikiem analizy. Metody wizualizacji to dzisiaj nie tylko narzędzie. Podobnie jak symulacje komputerowe stały się już one uznany paradygmatem badawczym. Traktowane jako sposób zdobywania wiedzy stają się metodą badawczą na równi z badaniami teoretycznymi, empirycznymi i numerycznymi⁴¹.

	AN	BN	CN	DN	EN	FN
Long Name	Net	File	XYZConv	Net	File 1	File 2
Comments						
1	1	2	4941	0.66332	3277.45488	-1363.90248
2	1	3	4942	-0.27131	-1340.82782	1327.40582
3	1	4	4943	-0.90565	-4727.87521	3066.41084
4	1	5	4944	-0.76229	-3768.73886	-1069.0487
5	1	6	4945	0.13277	604.54788	636.38777
6	1	7	4946	0.90576	6479.87558	3377.3883
7	1	8	4947	0.846	4186.13505	-698.02755
8	1	9	4948	0.00643	41.7982	-38.0016
9	1	10	4949	-0.80889	-4141.75228	3475.22649
10	1	11	4950	-0.91277	-4518.2251	-18.8663
11	1	12	4951	-0.14846	-738.87817	-424.43182
12	1	13	4952	0.75127	3726.26488	3375.94565
13	1	14	4953	0.96128	4761.22819	651.03882
14	1	15	4954	0.3075	1424.27073	1082.00128
15	1	16	4955	-0.65061	-3223.76455	2687.26889
16	1	17	4956	-0.9055	-4808.16263	1366.82185
17	1	18	4957	-0.41978	-2080.86548	-1374.03288
18	1	19	4958	0.53893	2662.09395	2632.02458
19	1	20	4959	0.99959	4958.96025	2023.06273
20	1	21	4960	0.54367	2698.58811	-1477.00021
21	1	22	4961	-0.4126	-2044.42711	2044.34754
22	1	23	4962	-0.98842	-4809.49897	2615.94364
23	1	24	4963	-0.65067	-3228.04516	1362.41854
24	1	25	4964	0.27982	1389.02888	1376.81032
25	1	26	4965	0.95904	4761.65287	3080.40531
26	1	27	4966	0.75653	3754.81908	-1097.5395
27	1	28	4967	-0.14154	-703.01822	678.72946
28	1	29	4968	-0.90047	-4518.26111	3379.91669
29	1	30	4969	-0.84124	-4180.13583	-644.79203
30	1	31	4970	4.220218-4	2.09744	9.91862
31	1	32	4971	0.8417	4184.08337	3490.66178
32	1	33	4972	0.90912	4520.15321	-60.01283
33	1	34	4973	0.1407	699.73002	-561.72463



Rys. 3. Przykład analizy big data zastosowany do badania struktury białkowej drożdży.
Opis w tekście

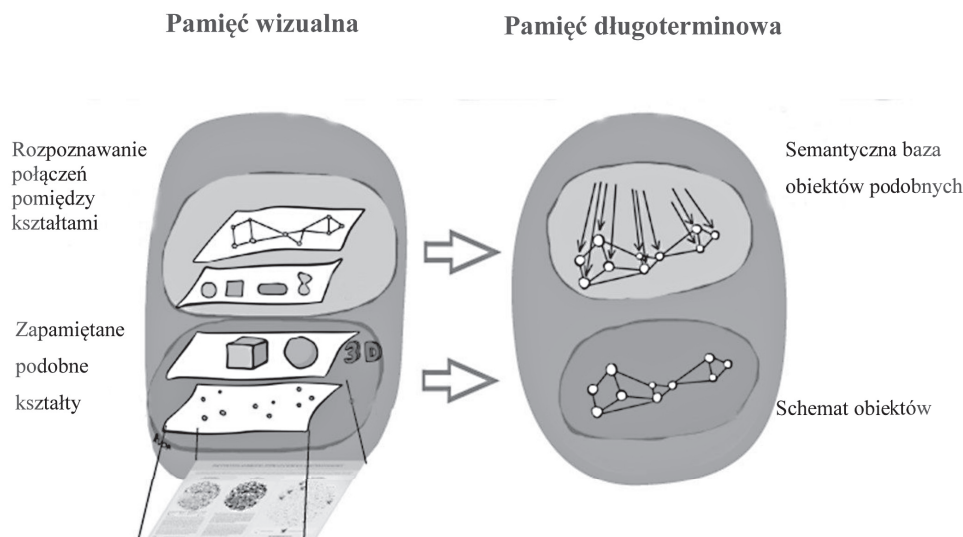
³⁹ Gephi.org – witryna otwartej platformy wizualizacyjnej stosowanej w procesach analizy big data.

⁴⁰ V. Osińska, *Mapujemy naukę*, „Głos Uczelni” 2014, nr 8.

⁴¹ T. Hey, S. Tansley, K. Tolle, *The Fourth Paradigm. Data-Intensive Scientific Discovery*, Washington 2009, s. 18.

Zauważmy, że w wyniku wizualizacji otrzymujemy obrazy, które możemy określić jako mapy. Pojęcie mapy kojarzy się nam zazwyczaj z kartografią; to zazwyczaj pierwsze mapy, jakie poznaliśmy w naszym życiu, opisują świat na dwuwymiarowej płaszczyźnie przy użyciu specyficznego języka. Pozwalają one na płaskim kolorowym obrazie pokazać olbrzymią ilość informacji na temat świata, którego nie widzimy bezpośrednio, ale możemy go sobie wyobrazić – wytworzyć obraz świata z mapy w przestrzeni naszego umysłu. Podobnie w analizie informacji będziemy dążyli do tego w przestrzeni naszego umysłu, aby odtworzyć obraz zawarty w dużych zbiorach danych. Nie jest to zadanie proste; aby je wykonać, musimy choć przez chwilę zastanowić się, w jaki sposób nasz mózg analizuje obrazy, a skomplikowane mapy w szczególności. Ograniczymy się w naszej dyskusji do obrazów statycznych, gdyż proces percepcji filmów i animacji jest znacznie bardziej skomplikowany. Z kognitywnego punktu widzenia problem polega na zbyt dużym obciążeniu systemu percepcji olbrzymią ilością danych wizualnych z jednoczesnym wyłączeniem modułów odpowiedzialnych za analizę merytoryczną zawartości informacyjnej. Często zbyt duża ilość informacji podawana jednocześnie powoduje u odbiorcy bierność w próbach analizy i pozostanie na etapie biernego odbiorcy. Odbiór mapy powinien więc być całościowy, traktowany jako globalny rezonans pomiędzy prekoncepcjami w mózgu odbiorcy a reprezentowaną treścią graficzną⁴². Mówiąc prostym językiem, mapa ma zmuszać do myślenia. Jednak aby to było możliwe, najpierw trzeba nauczyć odbiorcę czytać mapy. A to już niewątpliwie problem z zakresu dydaktyki, który wymaga interdyscyplinarnego podejścia wielu specjalistów. Powinniśmy, wykorzystując dostępne środki komunikacji, dążyć do wykształcenia odbiorcy mądrego, rozumiejącego konieczność poniesienia kosztów zdobycia wiedzy. Inny problem to wypracowanie we współczesnym systemie szkolnictwa ścieżek edukacyjnych, które już w szkole będą wskazywały na płytkość komunikatów w formie prymitywnych reprezentacji graficznych (tabloidów) i zachęcały do wykorzystywania technik wizualnych zmuszających do myślenia i analizy. Na rysunku 4 przedstawiony jest schematyczny proces percepcji skomplikowanego obrazu w postaci mapy. Dla uproszczenia rozważań pominęmy analizę barw i wzajemnych interakcji kolorystycznych.

⁴² V. Osińska, G. Osiński, A.B. Kwiatkowska, *Visualization in Learning: Perception, Aesthetics and Pragmatism w: Maximizing Cognitive Learning through Knowledge Visualization*, New York 2015, s. 386.



Rys. 4. Schemat analizy informacji zawartej na mapie z punktu widzenia procesów kognitywnych [Veslava Osińska, za zgodą autora]

Kiedy patrzymy na mapę, nasz mózg najpierw dokonuje analizy podobnych kształtów, następnie rozpoznaje połączenia pomiędzy nimi i próbuje stworzyć obiekt całościowy, który jest porównywany z już istniejącymi – w pamięci długotrwałej patrzącego – obiektami. Jeśli struktura powiązań zostaje rozpoznana, w przestrzeni umysłu pojawia się reprezentacja wirtualna. Naturalna zdolność człowieka do „oglądania” sprawia, że w naszym mózgu pojawia się projekcja obiektu ze świata rzeczywistego, który najpierw spełnia rolę imitacji rzeczywistości. Dopiero potem nasz mózg porównuje, w skomplikowanym procesie wykorzystującym różne moduły, ową imitację z dostępnymi znanymi obiektami. Następnie pojawia się w przestrzeni naszego umysłu nowy byt rozpoznawany i uzewnętrzniany jako widziany obiekt. Każda teoria dotycząca percepcji wizualnej będzie niepełna bez uwzględnienia problemów wynikających z powstawania reprezentacji językowej, która opisuje w sposób semantyczny oglądany obraz⁴³.

W powyższym procesie kognitywnym bardzo ważną rolę odgrywa więc wiedza zgromadzona przez odbiorcę przez całe jego dotychczasowe życie, ale również duży wpływ na sprawność tego procesu mają preconcepcje, jakimi często posługujemy się w oglądzie świata, często w sposób automatyczny. Preconcepcje to nasze intuicyjne i stereotypowe wyobrażenia o procesach i obiektach, nad którymi nigdy nie zastanawialiśmy się dogłębnie sami, ale powstały one w spontaniczny sposób

⁴³ Więcej przykładów wizualizacji Czytelnik znajdzie na portalu: <http://wizualizacjainformacji.pl/> [dostęp: 7 XI 2015].

w czasie naszego życia⁴⁴. Często nie zdajemy sobie sprawy z ich istnienia, dopiero konkretny problem, przed którym stajemy, uruchamia te procesy w naszym mózgu w pierwszej kolejności. Kiedy w galerii sztuki stajemy przed wybranym obrazem najpierw w naturalny sposób zastanawiamy się, co nam przypomina to, co widzimy, a dopiero potem zaczynamy analizować treść przekazu. Obraz zawsze niesie w sobie ładunek informacyjny i emocjonalny. Podstawowy błąd w wielu metodach analitycznych polega na pomijaniu lub umniejszaniu znaczenia tego drugiego. A przecież to właśnie ładunek emocjonalny uruchamia te znaczenia prekonceptyjne, które dla odbiorcy mają kluczowe znaczenie – wspomnienia z dzieciństwa, czy też osobiste przeżycia mają większe znaczenie niż opis semantyczny, który często towarzyszy prezentowanemu obrazowi.

W czasie prezentacji wystaw *Places & Spaces*⁴⁵ często słyszę od odbiorców zarzut niezrozumiałości prezentowanych map nauki. Lecz kiedy pytam te osoby, czego konkretnie nie rozumieją, co jest dla nich niejasne, zazwyczaj uzyskuję odpowiedź: „Ale co te obrazy mają przedstawiać? Jaką tezę udowadniają?”. Przykład ten pokazuje, jakie jest nastawienie do przekazu wizualnego większości odbiorców: oczekują oni dydaktycznego „przekazu z tezą”; obraz stanowi dla nich jakby dokończenie procesu analizy, a nie jego początek. Jednak oglądając płótna w galeriach malarstwa ci sami ludzie nie zadaliby takich pytań; rozumieją oni doskonale, że percepcja sztuki wymaga czasami bardzo starannego umocowania w kontekście: w epoce historycznej, w której autor tworzył dzieło, w konkretnych wydarzeniach tamtego czasu, które mogły mieć na niego wpływ, w stanie emocjonalnym artysty i wielu innych czynnikach. Obraz więc nie narzuca interpretacji. On ten proces inicjuje. Dlaczego więc mapy wiedzy czasami nie powodują owej inicjacji, lecz raczej pierwotne „odrzućcie”?

Oprócz przyczyn natury prekonceptyjnej z pewnością estetyka przekazu graficznego staje się niezwykle ważnym elementem, który musi być dostosowany do indywidualnego odbiorcy. Szczególnie istotna staje się ona podczas czytania map – różna kolorystyka i odpowiedni dobór tekstur może być dla jednego odbiorcy bardzo atrakcyjny, ale dla innego wręcz odrzucający. Trudno dobrać uniwersalną szatę graficzną, ale technologia prezentacji map pozwala nam na szerokie zastosowanie różnych technik, nawet ich łączenie. Ten swoisty synkretyzm stosowany we właściwym zakresie pozwoli zapewne odnaleźć złoty środek we właściwym doborze odpowiedniej kompozycji graficznej dla prezentowanej mapy⁴⁶.

⁴⁴ G. Osiński, V. Osińska, *Wizualizacja informacji w nauczaniu przedmiotów przyrodniczych*, „Biuletyn Polskiego Stowarzyszenia Nauczycieli Przedmiotów Przyrodniczych” 2014, t. 42, nr 4.

⁴⁵ <http://scimaps.org/> – portal internetowy poświęcony wizualizacji nauki [dostęp: 7 XI 2015].

⁴⁶ S. Zeki, *Blaski i cienie pracy mózgu. O miłości, sztuce i pogoni za szczęściem*, Warszawa 2015.

Jednak nie istnieje prosta recepta na właściwy dobór kompozycji graficznych. Stosowanie prostych zasad wynikających z teorii Gestalt⁴⁷ jest oczywiście konieczne, ale absolutnie niewystarczające. Trudno przypuszczać, aby proces całościowej percepcji obrazu był oparty li tylko na zwykłej analizie kształtów, linii, położenia i konfiguracji poszczególnych kształtów. Niedoceniana rola układu limbicznego⁴⁸ w procesie percepcji obrazu odbija się często na skuteczności przekazu. Cały czas jest dla nas wielką tajemnicą olbrzymi ładunek emocjonalny zgromadzony w naszych mózgach od wczesnego dzieciństwa, który ma decydujący wpływ na postrzeganie przez nas otaczającej rzeczywistości. Niektórych kształtów i kolorów po prostu nie lubimy, różne kształty budzą w naszym mózgu uczucia odrzucania, czy nawet strachu⁴⁹. Z kolei pozytywne skojarzenia są przyswajane łatwiej i powodują uruchomienie tych procesów poznawczych, które mogą uruchomić procesy analityczne.

Jednocześnie podstawowe założenia teorii Gestalt, a zwłaszcza prawa określane mianem *laws of grouping*⁵⁰ są przedstawiane jako fundamentalne zasady dotyczące percepcji wizualnej złożonych obiektów. Sukces aplikacyjny tych metod, w szczególności w projektowaniu interfejsów graficznych i w systemach *computer vision*, jest niezaprzeczalny, jednak w procesach analitycznych bardziej zależy nam na uruchomieniu wyższych procesów poznawczych za pomocą mechanizmów percepcji wizualnej. Musimy więc starać się zrozumieć, w jaki sposób powstają w mózgu nie tylko nowe wrażenia, ale również nowe idee. Procesy te są niewątpliwie dynamiczne, a często spotykane określenie *shining* oznacza, iż w końcowym etapie, w czasie powstawania nowej idei stają się one niezwykle szybkie. Neurodynamiczne koncepcje wynikające z eksperymentów psychologicznych i neurobrazowania pokazują ścisły związek pomiędzy mechanizmem percepcji wizualnej a aktywnością korelatów neuronalnych w wielu różnych obszarach mózgu – nie tylko w obszarach kory odpowiedzialnych za aktywność wzrokową. Procesy te charakteryzują się niezwykle dynamiką i dokładne zbadanie ich będzie zapewne wymagało w przyszłości użycia nowych metod eksperymentalnych wypracowanych przy jednoczesnym wykorzystaniu inżynierii genetycznej, nanotechnologii i optyki fluorescencyjnej. Obecnie bardzo szybko rozwija się dziedzina badania

⁴⁷ Gestalt – klasyczna teoria percepcji wizualnej.

⁴⁸ Układ limbiczny – układ struktur ludzkiego mózgu biorący udział w regulacji zachowań emocjonalnych.

⁴⁹ M. Klimarczyk, R. Mazur, G. Osiński, *Neuropsychological and psychiatric assessment of the clinical symptoms dynamics preceding brainstem's insufficiency in the course of stroke*, „Interdisciplinary Problems of Stroke” 2008, nr 10 (1), s. 21–26.

⁵⁰ *Laws of grouping* – zbiór zasad psychologii; po raz pierwszy zaproponowany przez psychologów reprezentujących podejście Gestalt w celu uwzględnienia obserwacji, że ludzie najpierw naturalnie postrzegają obiekty, a dopiero potem ich organizacje.

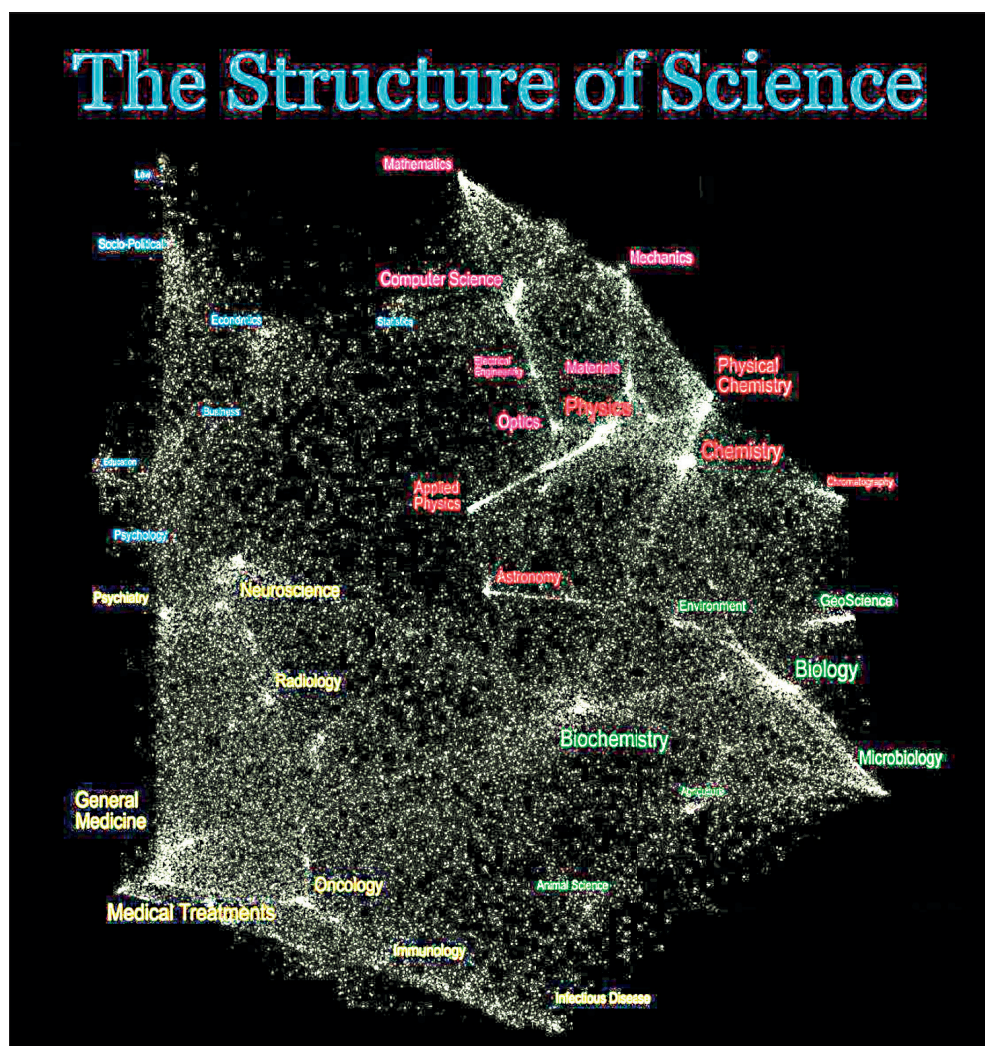
stanów dynamicznych korelatów neuronalnych. Odczucia wizualne mogą więc być odbierane przez ludzi na zasadzie natychmiastowego rezonansu – pomiędzy wzorcem wizualnym a istniejącą w naszym mózgu strukturą atraktorową.

Mapy nauki

Wizualizacja to tworzenie obrazów dla łatwiejszego zrozumienia form językowych, a nie dla ich zastąpienia, jak w przypadku tabloidyacji przekazu. Języki naturalne posługujące się słowem są nad wyraz skomplikowane i trudne; olbrzymie kłopoty sprawiają przekłady, a postępująca specjalizacja w badaniach naukowych nie tylko nie tworzy metajęzyków, ale posługuje się językami hermetycznymi, dostępnymi tylko dla nielicznych specjalistów. Zamknięcie form słownych w takich małych grupach powoduje, iż rozwijają się one tylko w obrębie szczegółowych dziedzin, a nie dają wglądu w naukę jako całość. Trudno jest nauczyć się wszystkich tych języków specjalistycznych (prawniczego, fizycznego, inżynierskiego, lekarskiego), a w poszukiwaniu prawdy trzeba często współpracować i łączyć idee i pomysły powstałe w różnych grupach naukowców. Wtedy idealnym, niejako uniwersalnym językiem będzie obraz – aczkolwiek zbudowany w skomplikowanych procesie, nierzadko algorytmicznym, to jednak ukazujący jakiś przekrój większej całości niż tylko specjalizacja naukowa. Ten obraz ma być przyczynkiem do podjęcia studiów, do rozwoju, a przede wszystkim dyskusji, która winna odbywać się za pomocą słowa. Ale słowa niejako umocowanego w uniwersalnej strukturze graficznej. Konstruowane obrazy nie mogą być abstrakcją, ale muszą być próbą odzwierciedlenia rzeczywistości przeprowadzoną na podstawie dostępnych informacji z różnych pól badawczych. Archipelagi wiedzy to dobry przykład takiego procesu – mapy nauki winny być początkiem długiej drogi, jaką winni przejść wszyscy naukowcy niezależnie od szczegółowej dyscypliny, jaką się zajmują na co dzień. Przyjrzyjmy się zastosowaniu metod wizualizacyjnych w badaniu struktur współczesnej nauki. Dobrym przykładem będzie mapa nauki. Jest ona stałym elementem wystawy *Places & Spaces*, na którą powoływałem się już w niniejszej pracy. Nowoczesne metody wizualizacji zostały zastosowane do szczegółowego zbadania struktury współczesnej nauki. Zespół kierowany przez Kevina Boyacka ze SciTech Strategies podjął się tego zadania w 2002 r. i w ciągu trzech lat poddał szczegółowej analizie ponad 730 000 prac naukowych, które ukazały się w 7300 renomowanych czasopismach naukowych.

To olbrzymi zbiór danych, zważywszy że publikacje często są tworzone przez zespoły naukowców, odwołując się do różnych dziedzin szczegółowych i powstają w różnym czasie. Poza tym to tylko wycinek aktywności naukowej obejmujący opisaną wcześniej „naukę utylitarną”. Jednak dla specjalisty z zakresu nauk komputerowych już taka analiza to duże wyzwanie metodologiczne: mamy tutaj do

czynienia z klasycznym przykładem analizy big data, i musimy stworzyć wydajny algorytm wizualizacyjny, który dokona odpowiednich przekształceń wielowymiarowej bazy danych tak, aby była możliwa prezentacja wyniku na powierzchni dwuwymiarowej mapy.

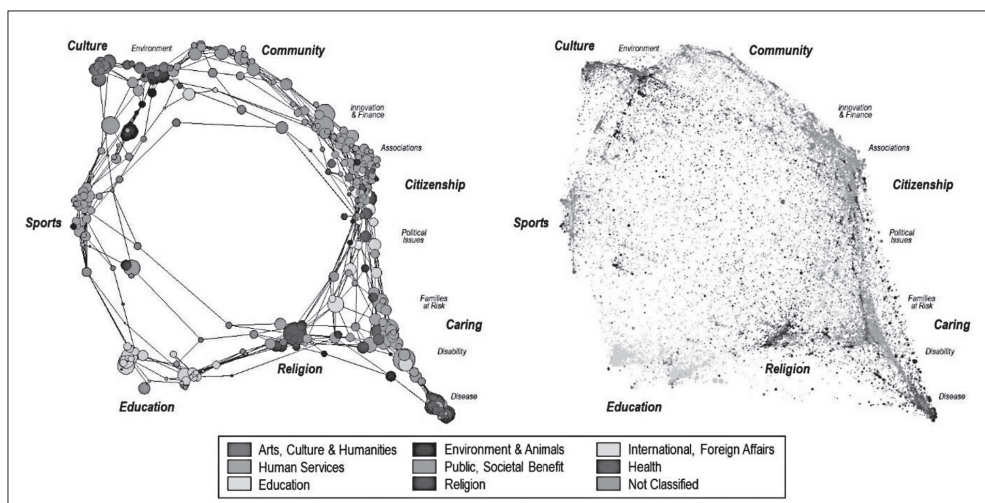


Rys. 5. Mapa nauki [dzięki uprzejmości Kevina Boyacka⁵¹]

Na mapie widzimy skupiska aktywności szczegółowych dziedzin nauki. Już pierwszy wgląd ukazuje nam ogólną strukturę składającą się z różnych trójkątnych struktur zanurzonych w czerni otaczającego tła. Poszczególne dziedziny wyłaniają się jak wyspy archipelagu z głębi oceanu. W lewym górnym rogu zgrupowały się

⁵¹ http://www.scimaps.org/detailMap/index/the_structure_of_sci_59 [dostęp: 14 VIII 2015].

nauki prawne i socjopolityczne⁵²; poniżej widzimy ekonomię oraz zgrupowane wokół *neuroscience* – psychologię i psychiatrię. Prawa strona mapy to archipeląg matematyki, fizyki i chemii. Pośrodku widzimy też dużą i szeroką płaszczyznę nauk komputerowych. Jednak najsilniejszy łącznik pomiędzy lewą i prawą stroną mapy to fizyka stosowana na północy oraz biochemia na południu. Dolna część mapy to domena nauk medycznych z lewej oraz biologii z prawej. Im jaśniejsze struktury poszczególnych obszarów, tym większa aktywność naukowa. Widzimy więc, które dziedziny obecnie rozwijają się najszybciej, a które przeżywają swoisty regres. Poza tym z zaprezentowanej mapy możemy odczytać, na styku których nauk prowadzi się obecnie największą aktywność badawczą. Na temat tej mapy wydano już wiele publikacji. Jest ona również cytowana i analizowana w licznych publikacjach książkowych. Nie zapominajmy jednak, że jest tylko jeden obraz. Dla kogoś, kto nie zajmuje się problematyką naukową, może to być tylko jedna z wielu ilustracji w książce, choć tak naprawdę pokazuje ona olbrzymią ilość informacji, którą można interpretować i odczytywać w zależności od potrzeb i aktualnie wykonywanych zadań badawczych.



Rys. 6. Mapa altruizmu [dzięki uprzejmości Kevina Boyacka, 2015]⁵³

Najnowszym dziełem autora mapy nauki jest opublikowana w 2015 r. mapa ukazująca związki pomiędzy aktywnością naukową a dobroczynnością. Poszukiwanie tak odległych związków nie powinno już Czytelnika dziwić. W celu wizualizacji Richard Klavans i Kevin Boyack zebrali dane tekstowe ze 125 000 stron

⁵² Na mapie użyto nazewnictwa zgodnego z angielską klasyfikacją nauk; w Polsce nie znamy np. odpowiednika dziedziny *neuroscience*, natomiast nauki społeczne i socjologia są traktowane jako odrębne dziedziny.

⁵³ http://scimaps.org/mapdetail/exploring_the_relati_180 [dostęp: 7 XI 2015].

internetowych organizacji non-profit działających na terenie USA. Drugie źródło danych to 43 miliony publikacji indeksowanych w bazach nauki Scopus. Następnie na podstawie modelu matematycznego obliczono wzajemne zależności pomiędzy poszczególnymi rodzajami aktywności i sporządzono wizualizacje. Z mapy możemy odczytać, że największą aktywność dobroczynną wykazują organizacje religijne. Dokładne dane w przypadku USA możemy znaleźć w oryginalnej pracy autorów⁵⁴, a wynoszą one, w zależności od wyznania: chrześcijaństwo – 82 procent, islam – 3,2 procent, hinduiści – 0,8 procent, buddyzm – 1,3 procent. Jest to pierwsza mapa działalności dobroczynnej, jaka powstała. Autor zaznacza, że ogromnym szokiem dla niego był fakt, że o pomocy biednym i wykluczonym tak dużo się mówi, organizuje się tak wiele spektakularnych akcji z udziałem światowych celebrytów, a do dzisiaj żaden naukowiec nie zadał sobie trudu, aby stworzyć mapę pokazującą, kto tak naprawdę jest zaangażowany w działania dobroczynne w skali globalnej.

Dodatkowo chciałbym zwrócić uwagę, że autor tego opracowania Kevin Boyack to naukowiec formatu cytowanego już Nassima Taleba. To zdecydowany krytyk obecnej komercjalizacji i ideologizacji współczesnej nauki. Opisywałem już problemy finansowania nauki, gdzie większość środków jest przeznaczona do wybranych, zideologizowanych i zazwyczaj skrajnie lewicowych uniwersytetów. Uprawiana tam nauka jest wspierana przez potężne korporacje międzynarodowe i służy raczej wspieraniu ich dominacji niż poszukiwaniu prawdy o otaczającym nas świecie. W tym roku udało mi się osobiście poznać Kevina Boyacka i krótko przedyskutować problemy związane z finansowaniem współpracy naukowców i podjąć temat wniosków, jakie można wysnuć na przyszłość.

Co czeka naukę w najbliższym czasie? Czy zostanie ona zdominowana przez ideologię transhumanizmu, komercjalizację i zupełnie oderwana od uniwersalnych korzeni historycznych, czy też nastąpi przeformułowanie paradygmatów badawczych, co ma już miejsce i co może zmienić niekorzystną strukturę zarządzania nauką? Najbliższa przyszłość pokaże, jakie zmiany nastąpią, a mapy nauki z pewnością pomogą rozpoznać krytyczne punkty i zaproponować korzystne zmiany.

Słowa kluczowe: wizualizacja, big data, informacja, percepcja.

⁵⁴ R. Klavans, K.W. Boyack, *Exploring the Relationships Between a Map of Altruism and a Map of Science*, „Bulletin of the American Society for Information Science and technology” 2015, t. 41(2), s. 30-33.

Summary

Information Visualization. The research of information structures in the search for truth

Visualization is creating images for easier understanding of the various forms of language. Visualization is no substitute forms of language, but creatively complement it. Natural languages creating the words by semantic compounds. They are extremely complicated and difficult, serious problems make translations between different languages and progressive specialization in scientific research produces hermetic language forms. Closing of the forms of the word such small groups causes that they develop only in the specific areas and do not give insight into studying, as a whole. The original academic universalism disappears turning into a scientific - financial corporation. In the search for truth, it is often necessary to collaborate, connect ideas and concepts arising in different environments. Then it seems that the best, the universal language will be - image - built in the complex algorithmic processes. Visualization in graphic form is the beginning of a deeper study, development and above all the discussion that must take place by means of the word. But the words somehow authorized in the universal structure of the graph.

Visualizations, however, cannot be abstract and must be a real attempt to reflect reality carried out on the basis of available information from various research fields. This visualization will not only expand the horizons of knowledge, but above all being able to show us a comprehensive grasp of describing the world, from which emerges a timeless order and truth, that which is unattainable with limited, closed-minded.

Keywords: *visualization, big data, information, perception.*

Bibliografia

Opracowania

- Black Edwin, *IBM i holokaust. Strategiczny sojusz hitlerowskich Niemiec z amerykańską korporacją*, Warszawa 2006.
- Boole George, *An Investigation into The Laws of Thought on Which Are Founded The Mathematical Theories of Logic and Probabilities*, Cork 1853.
- Bu D. i in., *Topological structure analysis of the protein - protein interaction network in budding yeast*, „Nucleic Acids Research” 2003, Vol 31, No. 9.
- Duch Włodzisław, Paweł Matykwicz, John Pesian, *Neurolinguistic Approach to Natural Language Processing with Applications to Medical Text Analysis*, „Neural Networks” 2008, 21(10).
- Godle Kurt, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, „Monatshefte für Mathematik und Physik” 1931, nr 38.
- Hey Tony, Tansley Stewart, Tolle Kristine, *The Fourth Paradigm. Data-Intensive Scientific Discovery*, Redmond, Washington 2009.

- Kacprzak Monika, *Odczytanie marksizmu przez Antonio Gramsciego i jego znaczenie dla współczesnej kultury*, „Fides, Ratio et Patria. Studia Toruńskie” 2014, nr 1.
- Klavans Richard, Boyack Kevin W., *Exploring the Relationships Between a Map of Altruism and a Map of Science*, „Bulletin of the American Society for Information Science and technology” 2015, t. 41(2).
- Klimarczyk Maciej, Mazur Roman, Osiński Grzegorz, *Neuropsychological and psychiatric assessment of the clinical symptoms dynamics preceding brainstem's insufficiency in the course of stroke*, „Interdyscyplinary Problems of Stroke” 2008, nr 10 (1).
- Larose Daniel T., *Odkrywanie wiedzy z danych*, Warszawa 2013.
- Moisse Katie, Sydney Lupkin, *Superstorm Sandy Tests Hospital Preparedness*. ABC News 2012.
- Osińska Veslava, Osiński Grzegorz, Kwiatkowska Anna B., *Visualization in Learning: Perception, Aesthetics and Pragmatism*, w: *Maximizing Cognitive Learning through Knowledge Visualization*, New York 2015.
- Osińska Veslava, *Mapujemy naukę*, „Głos Uczelni” 2014, nr 8.
- Osiński Grzegorz, Osińska Veslava, *Wizualizacja informacji w nauczaniu przedmiotów przyrodniczych*, „Biuletyn Polskiego Stowarzyszenia Nauczycieli Przedmiotów Przyrodniczych” 2014, t. 42, nr 4.
- Szymański Julian, Duch Włodzisław, *Representation of Hypertext Documents Based on Terms, Links and Text Compressibility. Proceedings of Neural Information Processing. Theory and Algorithms*, 17th International Conference, ICONIP 2010, Sydney, Australia, November 21-25, 2010, Proceedings, Part I, Sydney 2010.
- Taleb Nassim Nicholas, *Antykruchość. O rzeczach, którym służą wstrząsy*, Warszawa 2013.
- Urbanek Mariusz, *Genialni. Lwowska szkoła matematyczna*, Warszawa 2014.
- Zeki Semir, *Blaski i cienie pracy mózgu. O miłości, sztuce i pogoni za szczęściem*, Warszawa 2015.

Internet:

- <http://scimaps.org>.
- http://scimaps.org/mapdetail/exploring_the_relati_180.
- <http://wizualizacjainformacji.pl>.
- <http://www.forbes.pl/tusk-polska-najwiekszym-beneficjentem-budzetu-ue,artykuly,138696,1,1.html>.
- http://www.scimaps.org/detailMap/index/the_structure_of_sci_59.
- https://www.premier.gov.pl/files/budzet_ue_prezentacja_0.pdf.
- www.barackobama.com.