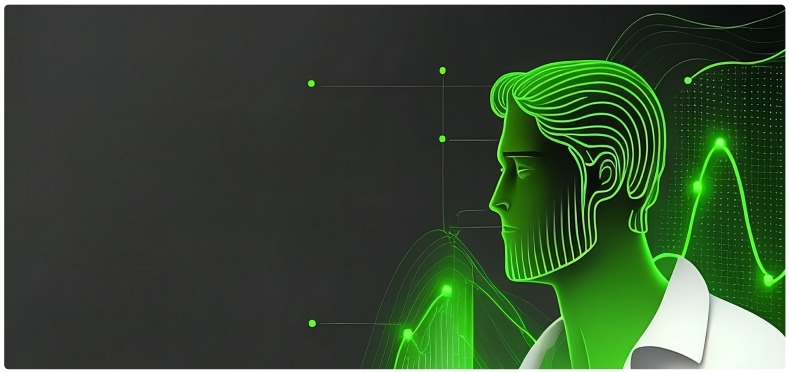
Czym Jest Danetyka?

14 kwietnia 2023 · 12 min



Na początek odrobina historii

Jeżeli choć odrobinę interesujecie się programowaniem i tematami pokrewnymi, a ostatnich 10 lat nie spędziliście na Erasmusie wśród Amiszów, to nie ma opcji, żebyście nie spotkali się do tej pory z angielskim terminem *Data Science* (pl. *nauka o danych*, czy też *danetyka/danologia* - do wyboru do koloru…). Jeżeli jednak jakimś cudem ominęły Was ostatnie lata życia w cywilizacji, to pozwólcie, że opowiem Wam krótką historyjkę.

Wcale nie tak dawno, bo w 1962 roku, ale również wcale nie tak blisko, bo na kampusie Uniwersytetu w Princeton, pomieszkiwał i popracowywał sobie za pieniądze amerykańskich podatników John Tukey. Był to jegomość o szczególnie wyróżniających się umiejętnościach we władaniu matematyką i statystyką, za co w pewnym momencie swojego życia został nawet uhonorowany Nagrodą Nobla. John Tukey wyróżniał się jednak jak na swoje czasy nie tylko umiejętnościami ścisłymi, ale również humanistycznymi, bo parał się, dość biegle zresztą, jasnowidztwem. Przewidział mianowicie, że w bliskiej przyszłości komputery znajdą swoje zastosowanie w rozwiązywaniu problemów statystycznych tak, jak do tej pory służyły w rozwiązywaniu problemów matematycznych. Ostatecznie, wieszczył, nastąpi połączenie metod matematycznych, metod statystycznych oraz metod komputerowych do nowej, zunifikowanej, interdyscyplinarnej dziedziny skupiającej się na analizie danych.

Swoje przemyślenia opisał dość szczegółowo w swoim artykule z 1962 roku zatytułowanym, notabene, *The Future of Data Analysis* [1], w którym najprawdopodobniej jako pierwszy użył terminu *Data Analysis* (pl. analiza danych) w kontekście przetwarzania, analizy i modelowania danych w interdyscyplinarnym podejściu łączącym metody matematyczne, statystyczne oraz informatyczne, korzystając z dobrodziejstw komputerów.

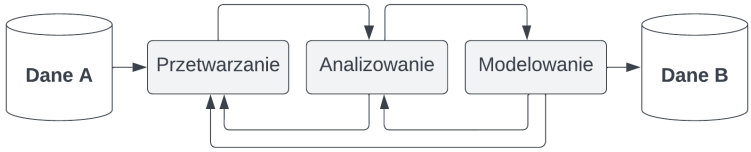
Co ważne, zwracał on również uwagę, że analiza danych powinna się skupiać przede wszystkim na zrozumieniu danych oraz zjawisk, które te dane opisują, a nie po prostu na dopasowywaniu istniejących modeli i teorii do danych w celu ich wyjaśnienia. Ta zmiana dotychczasowego paradygmatu otworzyła drzwi dla eksploracyjnej analizy danych (ang. *exploratory data analysis - EDA*), której głównym celem jest zrozumienie struktury danych i występujących w nich trendów oraz zależności, korzystając nie tylko z suchych liczb zwracanych przez statystyki opisowe, ale przede wszystkim korzystając z dobrodziejstw graficznej reprezentacji danych, czyli metod ich wizualizacji. Tukey zwracał również uwagę na konieczność traktowania EDA, jako procesu iteracyjnego, w ramach którego testujemy założenia szczątkowe i w zależności od potrzeb dostosowujemy pod kątem uzyskanych wyników pozostałe kroki przewidziane w ramach analizy danych. No dobrze, ale czym właściwie ta analiza danych jest?

Cóż, lata sobie mijały, aż w 1974 roku Peter Naur opublikował książkę zatytułowaną *Concise Survey of Computer Methods* [2], w której przedstawił definicję przetwarzania i analizy danych pozostającą aktualną do dnia dzisiejszego. Definicję tę można sparafrazować do następującej formy:

> Przetwarzanie i analiza danych, to zbiór procesów które na podstawie zbioru danych A tworzą zbiór danych B zapewniający przy pomyślnych wiatrach nowe informacje względem zbioru A. Nowy zbiór danych B powinien pozwalać na realizację bardziej efektywnych działań, niż zbiór A.

Dziś należy dodać jeszcze do przetwarzania i analizy danych modelowanie danych; istniało ono oczywiście wcześniej, ale ogólnie rzecz biorąc nie w takiej formie, jak istnieje i jest rozumiane obecnie; generalizuję, ale nie chcę teraz wchodzić za bardzo w szczegóły. Niemniej, na każdym z wymienionych etapów może być konieczne cofnięcie się do etapu poprzedniego, tj. po przetworzeniu danych i rozpoczęciu pierwszej tury analiz możemy dojść do wniosku, że powinniśmy przetworzyć dane raz jeszcze, tym razem inaczej, po czym powtórzyć analizy itd. Procesy przetwarzania, analizy i modelowania danych trwają dopóty, dopóki nie uzyskamy satysfakcjonujących nas wniosków/efektów.

Wszystko można przedstawić wizualnie w sposób następujący:



No dobrze, a co wchodzi w skład każdego z wymienionych procesów oraz jakie pełnią one role? Najlepiej będzie jeśli po prostu przyjrzymy się w sposób bardziej drobiazgowy każdemu z nich. Jednak zanim to zrobimy, to szybka odpowiedź na pytanie o to, po co w ogóle to robimy. Czyli po co “danetykujemy”?

Dane a informacje

Zacznijmy od szybkiego wytłumaczenia czym są dane oraz czym są informacje. Dane, to suche fakty, których odpowiednie przetworzenie i analiza mogą zapewnić informacje, czyli użyteczne wnioski z punktu widzenia interesariusza (ang. *stakeholder*), czyli osoby, która dzięki tym wnioskom będzie mogła usprawnić wybrany aspekt działania np. swojej firmy. Dla przykładu, danymi mogą być wiek, wykształcenie oraz dochód klientów banku, natomiast informacją może być częstość zaciągania kredytów gotówkowych w zależności od wymienionych “parametrów” klientów.

> W skrócie, dane, to suche fakty, natomiast informacje, to użyteczne wnioski.

Oczywiście, nic nie stoi na przeszkodzie, aby informacje stały się danymi, na podstawie których zostaną wyciągnięte inne wnioski, czyli powstaną nowe informacje i tak w koło Macieju.

A dlaczego? A komu to potrzebne?

No dobrze, danetykujemy, żeby dane przekuwać w informacje. A w jakich branżach i w jakich celach? Dla przykładu:

- Energetyka (np. analiza przepustowości sieci przesyłowych dla wielu źródeł energii różnego typu dla różnych lokalizacji i różnych pór dnia i roku)
- Transport (np. analiza zagęszczenia ruchu drogowego na wybranych odcinkach w zależności od pory dnia oraz pracy sygnalizacji świetlnej)
- Logistyka (np. analiza wolumenu transportowego, jego szybkości i kosztów w zależności od rodzaju transportu)
- Medycyna (np. analiza danych pacjentów pod kątem optymalizacji zespołu wybranego szpitala pod kątem dobrania odpowiedniej liczby lekarzy danych specjalności)
- Badania kliniczne (np. analiza skuteczności nowych leków w zależności od charakterystyki pacjentów)
- Diagnostyka (np. opracowywanie zautomatyzowanych rozwiązań i metod diagnostycznych)
- Farmacja (np. przyspieszenie procesu wynajdywania nowych związków o potencjale leczniczym)
- Produkcja (np. usprawnianie i optymalizacja procesów i linii produkcyjnych)

- Motoryzacja (np. rozwój rozwiązań autonomicznych zwiększających bezpieczeństwo na drodze)
- E-Commerce (np. analiza danych klientów pod kątem usprawnienia personalizacji reklam, ofert i rekomendacji produktów)
- Cyberbezpieczeństwo (np. usprawnienie procesów wykrywania oszustw i exploitów)
- Finanse (np. analiza rynków, szacowanie ryzyka inwestycji, czy też choćby personalizacja ofert produktów finansowych dla klientów banków w zależności od ich profilu)
- I wiele innych… Gdzie dla przykładu możemy napomknąć o “sztucznych inteligencjach” wykorzystywanych obecnie niemal do wszystkiego, przez wszystkich i wszędzie (polecam przy okazji film “Wszystko wszędzie naraz” - perełka)

Jak więc widzimy danetykować możemy praktycznie w dowolnej branży, gdzie na podstawie danych tworzone są informacje. A z uwagi na fakt, że w dzisiejszych czasach robi się to niemal wszędzie, to i niemal wszędzie każdy znajdzie coś dla siebie. Innymi słowy, pracy nie zabraknie (no chyba, że wygryzie nas AI hue hue hue).

> Co ciekawe, ilość przetwarzanych danych na całym świecie w roku 2010 wynosiła 2 zetabajty. W roku 2022 było to już 97 zetabajtów. Natomiast przewiduje się (liberalnie), że w roku 2025 będzie to już 181 zetabajtów. Zwracam tutaj uwagę na to, że jeden zetabajt, to jeden bilion, czyli tysiąc miliardów gigabajtów. Także ten, no, sporo.

Czym jest przetwarzanie danych?

Rzadko kiedy dane (czyt. nigdy), na których mamy zrealizować jeszcze bliżej nieokreślone analizy, do tych analiz się w ogóle na początku nadają. Najczęściej bowiem jest tak, że najpierw musimy je przejrzeć, zrozumieć ich strukturę, scharakteryzować ich typy oraz zrozumieć ich znaczenie. Gdy już rozumiemy na co właściwie patrzymy na ekranach naszych komputerów, to należy to, na co patrzymy, doprowadzać do relatywnego stanu użyteczności. To znaczy, sprawdzamy dane pod kątem występowania w nich duplikatów, identyfikujemy i zaradzamy brakom w danych, a także dane agregujemy. Najogólniej rzecz biorąc ugniatamy dane do momentu, aż przyjmą formę, która da się w sposób sensowny przeanalizować. A co to znaczy? Cóż, ciężko jest to opisać bez konkretnych przykładów, co postaram się zapewnić w następnych materiałach. Jednak najczęściej realizujemy następujące kroki w ramach przetwarzania danych:

1. Sprawdzić/określić typy danych w poszczególnych kolumnach i zweryfikować zgodność wprowadzonych wartości z ich domniemanymi typami.
2. Poprawić błędy/literówki w wybranych wartościach, jeżeli takowe występują.
3. Poradzić sobie z brakującymi wartościami.
4. Poradzić sobie z duplikatami wartości.
5. Jeżeli zajdzie taka potrzeba, to przekonwertować wybrane typy danych na inne typy.
6. Przeprowadzić normalizację wybranych zmiennych, czyli sprowadzić je do wspólnej, porównywalnej skali.
7. Przeprowadzić kodowanie zmiennych kategorialnych, czyli np. dla zmiennej `płeć` zmienić wartości `kobieta` i `mężczyzna` na `0` i `1`.
8. Przefiltrować, przekształcić, zredukować dane do pożądanego zakresu.
9. Powtarzamy maglowanie do momentu, aż dane przyjmą sensowną, dającą się analizować formę.

Czym jest analizowanie danych?

Najogólniej rzecz biorąc celem analizowania danych jest zapewnienie na ich podstawie użytecznych informacji. Najczęściej więc zaczynamy od wspomnianej już wcześniej eksploracyjnej analizy danych (EDA), w ramach której wykorzystujemy statystyki opisowe oraz różnego rodzaju wizualizacje (np. histogramy, wykresy skrzynkowe, wykresy rozrzutu). Celem EDA jest charakteryzacja rozkładów danych, jak również identyfikacja potencjalnych trendów i zależności. Jeżeli znajdziemy w danych coś ciekawego, to następnym krokiem jest przeprowadzenie wnioskowania statystycznego, które pozwala zwalidować nasze założenia względem danych i odpowiedzieć na posiadane pytania badawcze. W ramach EDA oraz wnioskowania statystycznego najczęściej realizujemy następujące kroki:

1. Analiza miar częstości (liczebność, częstość względna, procent), tendencji centralnej (średnia, mediana, moda), dyspersji (wariancja, odchylenie standardowe) oraz asymetrii (skośność, kurtoza).
2. Wizualizacja danych (np. histogram, wykres skrzynkowy / rozrzutu).
3. Wnioskowanie statystyczne (np. chi kwadrat, korelacja, test t, analiza wariancji).
4. W zależności od potrzeb powtórzenie EDA lub nawet powtórzenie procesu przetwarzania danych przed kolejną iteracją EDA.

Czym jest modelowanie danych?

Gdy już poznaliśmy w sposób wystarczający dane dzięki ich przetwarzaniu i analizie, to możemy pójść o krok dalej i spróbować stworzyć model te dane opisujący. Po co? Przychodzą mi teraz do głowy dwa główne powody. Po pierwsze, po to, aby móc dokonywać predykcji na temat danych nieistniejących na podstawie danych istniejących - czyli tak zwane modelowanie predykcyjne (ang. *predictive modelling*). Po drugie, aby uzyskać jeszcze głębszy wgląd w dane i dostrzec istniejące w nich wzorce i zależności na poziomie, jakiego nie mogliśmy osiągnąć z wykorzystaniem EDA oraz wnioskowania statystycznego - czyli tak zwane modelowanie opisowe (ang. *descriptive modelling*).

> Możemy rozróżnić modelowanie predykcyjne oraz modelowanie opisowe.

Wśród najpopularniejszych technik wykorzystywanych w modelowaniu danych możemy wymienić:

1. Regresje (ang. *regression*)
2. Drzewa decyzyjne (ang. *decision trees*)
3. Losowy las decyzyjny (ang. *random forest*)
4. Sieci neuronowe (ang. *neural networks*)
5. Głębokie sieci neuronowe (ang. *deep neural networks*)

Danetyka kilka lat temu i dzisiaj

Jeszcze kilka lat temu mało kto mówił w Polsce o *data science*, o czym niech świadczą statystyki wyszukiwania tej frazy w wyszukiwarce Google dla naszego kraju.

Wzrost zaczął się dopiero w okolicach roku 2016, co niekoniecznie dobrze świadczy o naszym pięknym kraju (tutaj bez sarkazmu - Polska jest piękna). Pozwólcie bowiem, że przedstawię Wam krótką listę wydarzeń ze świata analizy danych tylko z ostatnich kilkunastu lat.

- Rok 2002 - Powołanie do życia pierwszego czasopisma naukowego dedykowanego *data science* przez *Committee on Data Science and Technology* o nazwie *Data Science Journal*.
- Rok 2008 - Termin *data scientist* staje się światowym *viralem* (bynajmniej nie u nas).
- Wielkie korporacje, jak Google, czy Facebook zaczynają opierać swoje istnienie na pośredniczeniu między reklamodawcami a swoimi użytkownikami. Przetwarzanie wielkich wolumenów danych, ich analiza i budowanie modeli predykcyjnych staje się nieodłączną częścią działalności korporacji.
- Rok 2010 - Powstaje platforma Kaggle, zrzeszająca osoby zajmujące się *data science*.
- Rozkręca się *hype* na sztuczną inteligencję (ang. *artificial intelligence* - AI).
- Rok 2011 - Liczba ofert pracy dla danetyków skacze o 15 000%.
- Sztuczna inteligencja stworzona przez IBM wygrywa program *Jeopardy!*
- Rok 2012 - *Data scientist* zostaje okrzyknięte najseksowniejszym stanowiskiem pracy na świecie przez Harvard [3].
- Rok 2013 - Według firmy IBM 90% wszystkich danych na świecie zostało wytworzonych w ciągu ostatnich dwóch lat.
- Rok 2014 - *Data scientist* zostaje okrzyknięte najseksowniejszym stanowiskiem pracy na świecie przez magazyn *Forbes* [4].
- Rok 2015 - Dzięki wykorzystaniu uczenia głębokiego (ang. *deep learning* - jednej z form AI) efektywność systemu rozpoznawania mowy firmy Google skacze o 49%. Google

zwiększyło również wykorzystanie uczenia maszynowego z dotychczasowego sporadycznego, do ponad 2 700 wewnętrznych projektów.

- Rok 2016 - Sztuczna inteligencja o nazwie *AlphaGo* stworzona przez firmę Deep Mind należącą do firmy Google pokonuje mistrza świata w grze *Go*.
- Zaczyna się zainteresowanie tematem *data science* w Polsce na poważnie.
- Rok 2021 - Sztuczna inteligencja o nazwie *AlphaFold* przewiduje struktury białek nieporównywalnie efektywniej niż dotychczasowe metody komputerowe nadzorowane przez człowieka. Z tego co się orientuję, to na chwilę obecną przewidziała struktury już praktycznie wszystkich białek znanych nauce. Warto napomknąć, że jeszcze do niedawna niczym niezwykłym było robienie doktoratu w ramach którego odkrywano i badano strukturę JEDNEGO białka. Także ten...
- Rok 2022 - Jason Allen wykorzystując obrazy wygenerowane przez sztuczną inteligencję *Midjourney*, wygrywa konkurs artystyczny, pokonując artystów tworzących swoje prace ręcznie [5].
- Przełom roku 2022 i 2023 - Pojawia się *ChatGPT* firmy/fundacji OpenAI oparte o ich model GPT-3 (teraz już GPT-4, a niedługo GPT-5), przez co wielu ludzi wykonujących wiele różnych zawodów zaczyna... trochę się martwić.

Oczywiście lista istotnych z punktu widzenia branży wydarzeń w ostatnich 20 latach jest o wiele obszerniejsza, ale chciałem jedynie wywrzeć na Was pewne określone wrażenie. A mianowicie, świat danych i ich przetwarzania galopuje jak oszalały i w gruncie rzeczy nie wiadomo, co przyniesie jutrzejszy dzień. Co prawda co bardziej pesymistyczne głosy mówią, że danetyka popełnia rozciągnięte w czasie samobójstwo z uwagi na rozwój AI, która może ostatecznie sprawić, że dane "będą analizowały się same". Uważam jednak, że nawet jeżeli AI trafi pod danetyczne strzechy na dobre, to będzie miejsce dla ludzi, którzy potrafią z niej korzystać, rozumieją jej działanie oraz to, co "wypluwa" w wyniku swojego działania. Innymi słowy będą niejako pośredniczyć między światem zewnętrznym a technologią.

No to czym jest ta danetyka?

Uf, no właśnie - to czym jest ta danetyka? Na pewno jedną z najszybciej i najbardziej dynamicznie rozwijających się dziedzin. W niespełna 50 lat zdążyła powstać, rozwinąć się i zagrozić swemu istnieniu (w wykonaniu człowieka) przez swój bezprecedensowy rozwój. Oczywiście przesadzam i warto, aby spoglądając trzeźwo na otaczającą nas rzeczywistość rozumieć, że przetwarzanie, analiza i modelowanie danych, czy to własnoręcznie, czy przy użyciu AI, będzie najważniejszą dziedziną działalności człowieka w XXI wieku. Danetyka, to tłumaczenie danych na informacje, to optymalizowanie procesów, zwiększenie ich efektywności, a co za tym idzie zmniejszenie kosztów finansowych i energetycznych. Przyszłość należy do danych. Przyszłość należy do danetyki.

Autopromocja 🌱

Zachęcam do zainteresowania się moimi szkoleniami, które tworzę dla Strefy Kursów:

- Fundamenty programowania w Python ✨**nowość**✨
- Fundamenty języka Java

Oceny i liczba opinii na dzień 14 kwietnia 2023.

Od każdego zakupu mojego szkolenia po wejściu przez w/w linki dostaję kilka złotych prowizji. Nie musicie jednak dzięki temu płacić za szkolenia pełnej kwoty, bo możecie skorzystać z przygotowanych dla Was kodów rabatowych -30% `PL30FB` (ważny przez ograniczony czas) oraz -10% `ANco2` (nagroda pocieszenia, gdy pierwszy kod utraci ważność).

Z góry dziękuję za zainteresowanie.

Bibliografia

1. Tukey JW (1962) The Future of Data Analysis. Źródło.
2. Naur P (1974) Concise Survey of Computer Methods.
3. Davenport TH & Patil DJ (2012) “Data Scientist: The Sexiest Job of the 21st Century”. Harvard Business Review. Źródło.
4. Magyar J (2014) “Data Scientist: Sexiest Job Of The Century?”. Forbes. Źródło.
5. Harwell D (2022) “He used AI to win a fine-arts competition. Was it cheating?” Washington Post. Źródło.

Podobał się artykuł? Udostępnij go! 👇👇👇

popularnonaukowy