Potencjał współczesnego wnioskowania statystycznego. W setną rocznicę publikacji przez Ronalda A. Fishera monografii pt. *Statistical Methods for Research Workers*

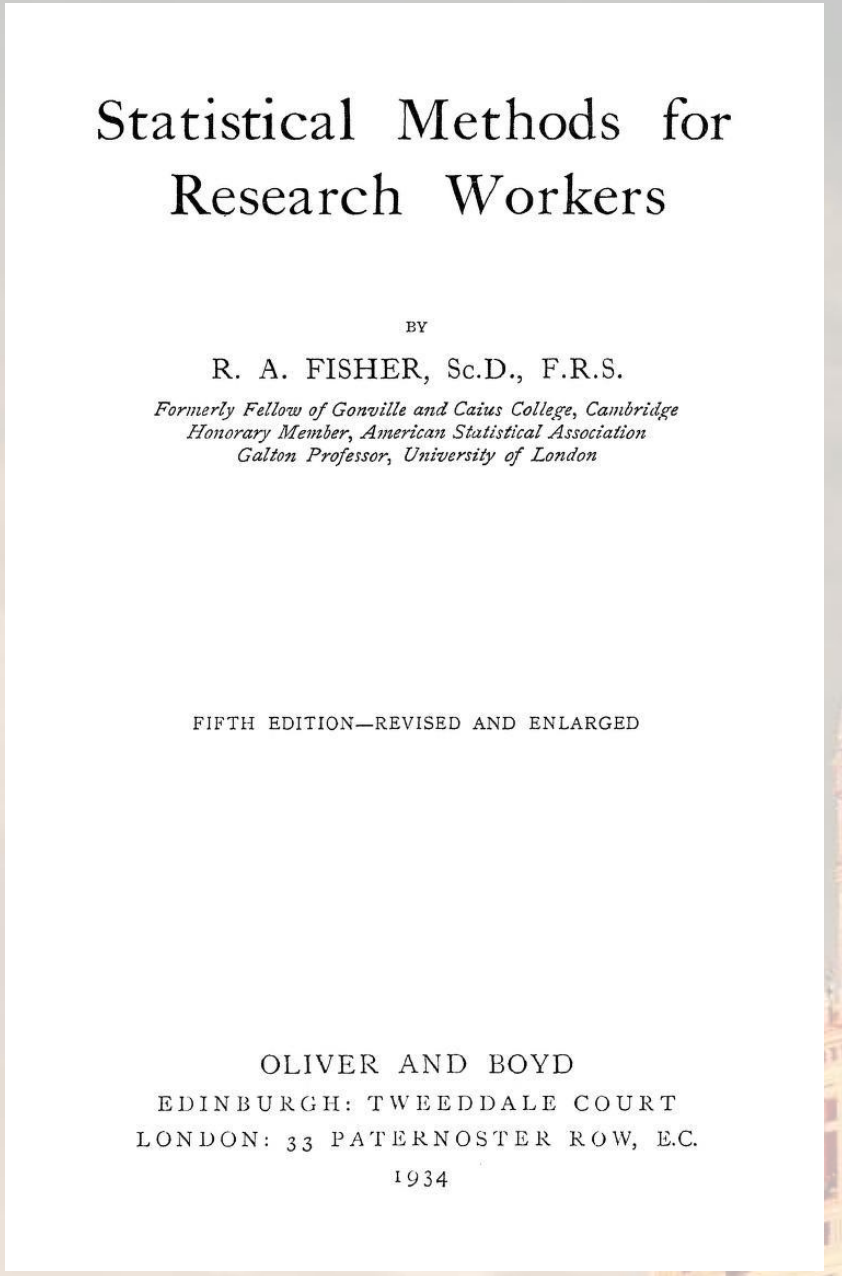
Even scientists need their heroes, and R. A. Fisher was certainly the hero of 20th century statistics.

(Efron, b. (1998). R. A. Fisher in the 21st century. "Statistical Science", 2 (13), s. 95)

Krzysztof Jajuga
Mirosław Szreder

V Kongres Statystyki Polskiej, 1-3 lipca2025 roku, Warszawa

Cel: wskazanie i scharakteryzowanie działań, jakie podjąć powinni statystycy, aby z jednej strony pełniej wykorzystać potencjał wnioskowania statystycznego w epoce big data i AI, a z drugiej silniej uświadomić i przekonać użytkowników statystyki o inherentnych elementach niepewności obecnych we wnioskowaniu statystycznym.



Statistical Methods for Research Workers

BY

R. A. FISHER, Sc.D., F.R.S.

Formerly Fellow of Gonville and Caius College, Cambridge
Honorary Member, American Statistical Association
Galton Professor, University of London

FIFTH EDITION—REVISED AND ENLARGED

OLIVER AND BOYD
EDINBURGH: TWEEDDALE COURT
LONDON: 33 PATERNOSTER ROW, E.C.
1934

PREFACE TO FIRST EDITION

For several years the author has been working in somewhat intimate co-operation with a number of biological research departments; the present book is in every sense the product of this circumstance. Daily contact with the statistical problems which present themselves to the laboratory worker has stimulated the purely mathematical researches upon which are based the methods here presented. Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data. Such at least has been the aim of this book.

I owe more than I can say to Mr W. S. Gosset, Mr E. Somerfield, and Miss W. A. Mackenzie, who have read the proofs and made many valuable suggestions. Many small but none the less troublesome errors have been removed; I shall be grateful to readers who will notify me of any further errors and ambiguities they may detect.

Rothamsted Experimental Station,
February 1925.

Statystyka – nauka badająca prawidłowości występujące w zjawiskach masowych.

Ronald Fisher dostrzegał trzy najważniejsze aspekty w rozumieniu statystyki, definiując ją jako:

- naukę o populacjach (the study of **populations**),
- naukę o zróżnicowaniu, zmienności (the study of **variation**),
- naukę o metodach redukcji danych (the study of methods of the **reduction of data**).

Dalej dodał (s. 3):

The conception of statistics as the study of variation is the natural outcome of viewing the subject as the study of populations.

Syntetycznie wyraził to profesor Harvardu Stephen Jay Gould:

Variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency.

(Gould, S.J. (1985). The median isn't the message. “Discover”, 6, 40–42)

Jednym ze współczesnych zagrożeń w posługiwaniu się wnioskowaniem statystycznym jest skłonność do niedostrzegania tkwiącej w nim niepewności (nie tylko w samym modelu).

Duży potencjał elektronicznych sposobów gromadzenia danych oraz łatwe w obsłudze oprogramowanie statystyczne – wymagające od użytkownika zaledwie podstawowej wiedzy statystycznej – tworzą wrażenie posiadania niezawodnego i wolnego od niepewności zbioru metod i technik badawczych.

Andrew Gelman:
„Wydaje się, że statystyka jest często sprzedawana jako rodzaj alchemii, która przekształca losowość w pewność”.

It seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an “uncertainty laundering” that begins with data and concludes with success as measured by statistical significance.

(Gelman, A. (2016). The Problems With P-Values are not Just With P-Values. *The American Statistician, Online Discussion*)

Pierwsze zagrożenie, to złudzenie, że programy obliczeniowe są w stanie nie tylko przetworzyć dane statystyczne, wydobyć z nich wiedzę i nadać wynikom odpowiednią interpretację, ale także uwolnić użytkownika od konieczności zajmowania się niepewnością ocen (estymacja) oraz niepewnością podjętych decyzji (weryfikacja hipotez).
Innymi słowy: przekonanie, że myślenie statystyczne oparte na prawdopodobieństwie (*probability-based thinking*) nie ma już takiego znaczenia jak w przeszłości.

Drugie zagrożenie łączy się z samym prawdopodobieństwem jako miarą niepewności.
Częstościowa interpretacja prawdopodobieństwa staje się niewystarczająca → powszechne staje się łączenie (integracja) różnych źródeł danych o populacji, gwałtownie rośnie liczba rożnych źródeł informacji, rosną wskaźniki odmów (*nonresponse*), często sięga się do wiedzy ekspertów, ….

Ronald Fisher, aczkolwiek akceptujący częstościową interpretację prawdopodobieństwa, dobrze zdawał sobie sprawę z jej ograniczeń.

Filozofia Fishera charakteryzuje się serią błyskotliwych kompromisów między punktami widzenia bayesowskim i częstościowym.

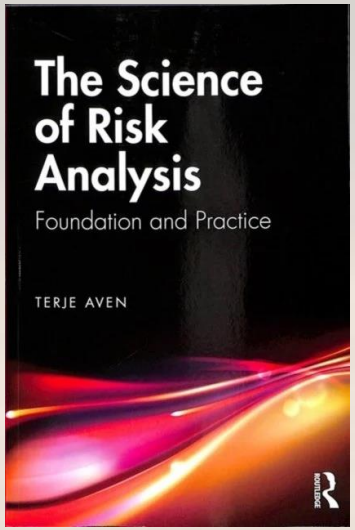
(Efron, B. (1998). R. A. Fisher in the 21st Century. "Statistical Science", 2 (13), s. 95)

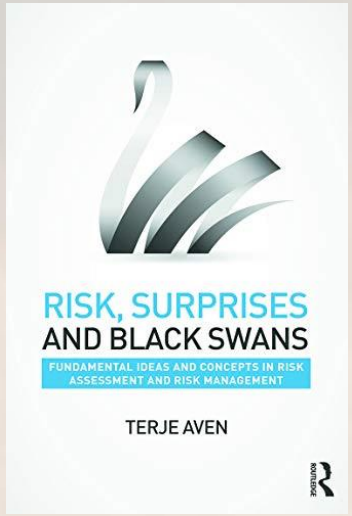
Fisher jest autorem tzw. **fiducialnego prawdopodobieństwa** (*fiducial probability*), mającego być przeciwwagą dla koncepcji rozkładów a priori.

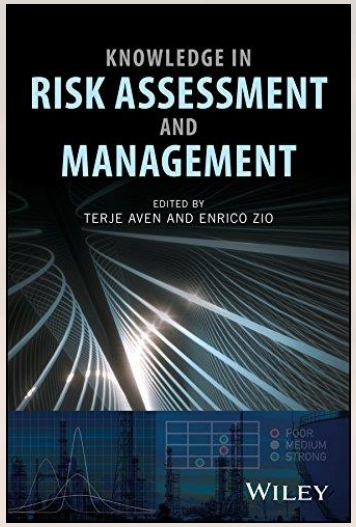
W niektórych jego pracach pojawia się prawdopodobieństwo rozumiane jako *numerical measure of rational belief* (Fisher, 1930, s. 532).

Czyli dopuszczał on interpretowanie prawdopodobieństwa jako *a state of mind* (korespondującego z personalistyczną interpretacją prawdopodobieństwa), a nie wyłącznie jako *a state of nature* (odpowiadającego częstościowej interpretacji).

Uważamy, że rosnąć będzie znaczenie personalistycznej interpretacji prawdopodobieństwa, która dopuszcza zarówno wiedzę ekspertów – ich racjonalny stopień przekonania (*rational degree of belief*), jak i kombinacje różnych źródeł informacji i ich łączenie np. z wykorzystaniem tw. Bayesa.







Norweski uczony Terje Aven w kilku swoich pracach słusznie przekonuje, że potrzebne jest powiązanie szacunków prawdopodobieństwa z ocenami „siły wiedzy”, na której te szacunki są oparte.

Wiedza wspierająca prawdopodobieństwo jest równie ważna, jak samo prawdopodobieństwo.

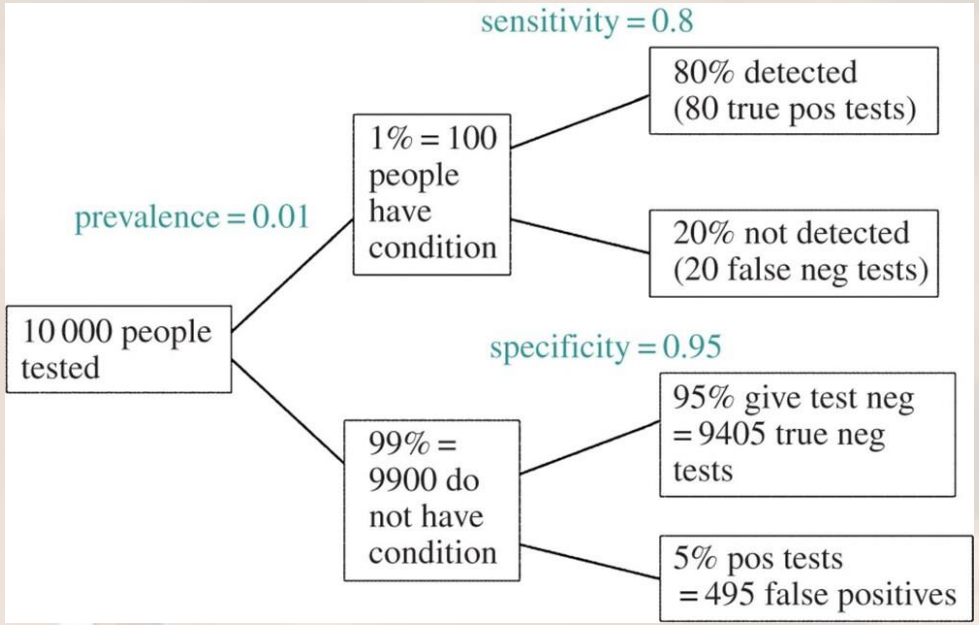
(Aven T. (2014), *Risk, Surprises and Black Swans: Fundamental Ideas and Concepts in Risk*, Routledge)

Specyficzne problemy niepewności tkwiącej w weryfikacji hipotez i kategorii statystycznej istotności

Problem 1

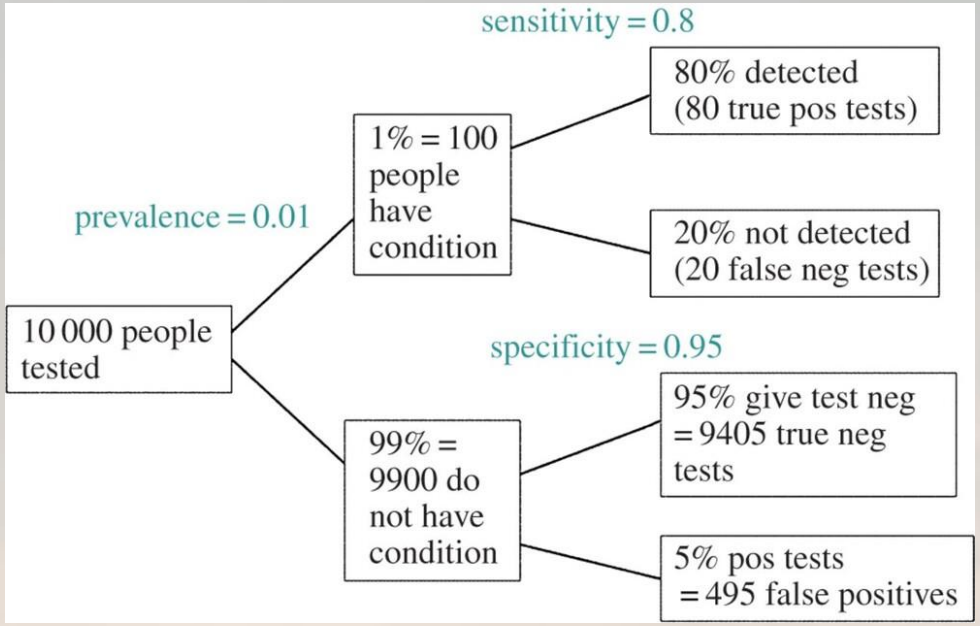
Wciąż ważna kwestia odsetka fałszywie pozytywnych decyzji

Na podst. Colquhoun D. (2014), An investigation of the false discovery rate and the misinterpretation of p-values. “Royal Society Open Science”, https://doi.org/10.1098/rsos.140216)



Pozytywny wynik otrzyma: 495+80=575 osób.

Odsetek fałszywych pozytywnych: 495/575 = **86%**.



Pozytywny wynik otrzyma: 495+80=575 osób.

Odsetek fałszywych pozytywnych: 495/575 = **86%**.

$$P(ILL|POS) = \frac{P(POS|ILL) \cdot P(ILL)}{P(POS|ILL) \cdot P(ILL) + P(POS|ILL^{not}) \cdot P(ILL^{not})}$$

$$P(ILL|POS) = \frac{0.8 \cdot 0.01}{0.8 \cdot 0.01 + 0.99 \cdot 0.05} = 0.139 \approx 0.14$$

Reszta pozytywnych wskazań (86%) jest fałszywa.

Problem 2

Prezentowanie teorii weryfikacji hipotez w podręcznikach i niektórych pracach naukowych w taki sposób, że nie w pełni zostają wyartykułowane wszystkie ważne założenia i źródła niepewności tkwiące w tej teorii.

Problem 3 (praktyka stosowania wnioskowania)

Wieloletni proces upraszczania i redukowania złożoności matematycznego modelu testowania hipotez do prostego zagadnienia decyzyjnego, w którym cała niepewność zawarta jest w poziomie istotności (prawdopodobieństwie popełnienia błędu I rodzaju).

Ronald Fisher dość wcześnie zdał sobie sprawę z tego, że zaproponowana przez niego metodyka testowania hipotez może być przez badaczy traktowana jako uwolniona od niepewności procedura decyzyjna. Być może dlatego w 13-tym wydaniu swojej książki „Statistical Methods for Research Workers” z 1958 roku dodał następujące zdanie:

„Testy istotności są używane jako pomoc w osądzie i nie powinny być mylone z automatycznymi testami akceptacji bądź z funkcjami decyzyjnymi”.

Staramy się uzasadnić, że współczesna praktyka wnioskowania statystycznego z powodu obecnych w niej dychotomii nie ujawnia wszystkich zawartych w nim źródeł niepewności.

Dychotomie te dotyczą przede wszystkim dwóch kategorii: statystycznej istotności i nieistotności, interpretacji wskaźnika p-value.

R. Fisher, podobnie jak W.S. Gosset i K. Pearson, nie postulował bezkompromisowego, jednoznacznego podziału wyników badania statystycznego na statystycznie istotne i nieistotne.

In Fisher's system, the P value was to be used as a rough numerical guide of the strength of evidence against the null hypothesis. There was no mention of "error rates" or hypothesis "rejection.

(Goodman S. (2008), A Dirty Dozen: Twelve P-Value Misconceptions. Semin Hematol 45:135-140, Elsevier Inc.)

Przed 100 laty pisał (s. 45):

Wygodnie jest przyjąć punkt 0,05 jako granicę w ocenie czy odchylenie ma być uważane za istotne czy nie.

O ile w 1925 r. Fisher uznawał 0,05 za standardową, typową wielkość wskazującą na istotność odchyleń, niezależnie od rodzaju badania, to w swojej ostatniej, najbardziej filozoficznej książce z 1956 roku pt. „Statistical methods and scientific inference" pisał:

Żaden badacz nie ma stałego poziomu istotności przy którym z roku na rok, we wszystkich okolicznościach odrzuca (zerową) hipotezę; zamiast tego koncentruje się raczej na konkretnym przypadku w świetle swoich dowodów i przemyśleń.

Współcześnie następuje fetyszyzacja wielkości 0,05, jako ostatecznego rozstrzygnięcia w procesie weryfikacji wszelkich hipotez statystycznych, uznanie statystycznej istotności za synonim wysokiej wartości naukowej uzyskanych wyników wnioskowania, a w dalszej kolejności przyjęcie, że niewarte uwagi są wyniki badań, którym nie udało się przypisać terminu „statystycznie istotne”.

Zmuszanie do wyboru pomiędzy wynikiem istotnym a nieistotnym przesłania niepewność obecną wszędzie tam, gdzie wnioskujemy na podstawie próby.

(Altman D.G. (1991). Practical statistics for medical research, Chapman and Hall, s. 169).

In most scientific settings, the arbitrary classification of results into ‘significant’’ and ‘non-significant’ is unnecessary for and often damaging to valid interpretation of data.

(Greenland et al. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. “European Journal of Epidemiology” 31(4), s. 338)

Warto pamiętać, że statystyczna istotność nie odnosi się ani do wielkości efektu (*size effect*), którego dotyczy hipoteza zerowa, ani do właściwości populacji, z której pochodzi próba, lecz do konkretnego wyniku testu opartego na wynikach pojedynczej próby.

Uważamy, że usilne dążenie części badaczy do osiągania statystycznej istotności w testowaniu stawianych hipotez – co często jest wstępnym warunkiem zainteresowania wynikami badań przez redakcje czasopism naukowych – staje się w dobie *big data* groźniejsze dla nauki niż było w przeszłości.

1. Duże zbiory danych umożliwiają uzyskanie statystycznie istotnych wyników nawet dla bardzo niewielkich i merytorycznie nieznaczących efektów,
2. Zaawansowane oprogramowanie statystyczne znacznie uprościło wielokrotne testowania, czyli nierzetelną praktykę, której celem jest uzyskanie statystycznie istotnego wyniku: *p-hacking* (albo *data-dredging*, *snooping* czy *significance-chasing*).

Przykład praktyki *p-hacking*

Przyjmując poziom istotności 0,05, przeciętnie raz na 20 losowań pojawiać się będzie próba nietypowa, upoważniająca do odrzucenia hipotezy zerowej.

Jeżeli badacz w 19 próbach uzyska nieistotną wielkość efektu, a poinformuje odbiorców jedynie o wyniku dwudziestej próby (uznanym za statystycznie istotny), to rzeczywiste prawdopodobieństwo błędu pierwszego rodzaju, polegającego na odrzuceniu hipotezy prawdziwej, nie będzie wynosiło 0,05 lecz aż 0,64:

$$1 - 0{,}95^{20} \approx 0{,}642$$

W czasach szerszej dostępności do dużych baz danych wzrastać będzie znaczenie **bayesowskiego podejścia do testowania hipotez**.
Głównie z tego powodu, że coraz większa jest potrzeba włączenia do wnioskowania wiedzy już posiadanej, i to w formie, która pozwala na przypisanie jej określonego stopnia niepewności.

Podejście oparte na twierdzeniu Bayesa – zasadniczo różne od koncepcji Fishera – jest jednak bliskie tej koncepcji w jednym punkcie – w formułowaniu ostatecznej konkluzji, która nie jest wyrażona w kategoriach podejmowania decyzji.
Jej istotą jest wskazanie, jak silne są przesłanki na rzecz hipotezy alternatywnej w stosunku do hipotezy zerowej.

Tę relację wyraża w bayesowskim podejściu tzw. czynnik bayesowski BF (*Bayes Factor*):

$$BF = \frac{p(D|H_1)}{p(D|H_0)}$$

Relacja ta odnosi się bezpośrednio do hipotez, a nie do danych, jak wskaźnik p-value, a ponadto uwzględnia nie tylko informacje z próby (D), lecz także wcześniejszą wiedzę (a priori) na temat obu hipotez.

W kontekście big data i narzędzi sztucznej inteligencji wnioskowanie statystyczne nabiera innego (szerszego znaczenia).

Klasyczne określenie, w którym jest mowa o identyfikacji zależności istniejących w populacji podlegającej pewnemu rozkładowi prawdopodobieństwa można tu zastąpić bardziej ogólnym określeniem: wnioskowanie statystyczne jest to proces zastosowania metod analizy danych w celu identyfikacji zależności istniejących w zbiorze danych. Te zależności mogą stanowić podstawę sformułowania pewnych hipotez lub weryfikacji istniejących hipotez.

W przypadku stosowania narzędzi uczenia maszynowego z wykorzystaniem big data, jakość badań prowadzących do naukowych konkluzji zależy od obu filarów: danych i metod.
Pierwszy element jakości narzędzi uczenia maszynowego to jakość danych. Jest to kluczowy element, z uwagi na to, że pozyskiwane big data mogą mieć różną postać. Część alternatywnych danych może być przekształcona na zwykle dane numeryczne (liczbowe), co może ułatwić wnioskowanie.

Jednak jest coraz więcej danych w postaci tekstowej. Tutaj stosowane są duże modele językowe (Large Language Models), które analizują kontekst całych fragmentów tekstu. Jest to istotna zmiana w porównaniu ze starszymi modelami (znanymi jako modele Natural Language Processing), działającymi na zasadzie „worka słów” (bag of words), w przypadku których słowa mogły być przekształcone na kategorie liczbowe. Przedstawiony fakt dotyczący zaawansowanych dużych modeli językowych oznacza brak możliwości stosowania klasycznej teorii wnioskowania statystycznego.

Inny problem związany z jakością danych stosowanych w uczeniu maszynowym, to ich wiarygodność (określona jako veracity – jeden z elementów 5 V). Dotyczy to przede wszystkim danych pozyskiwanych z internetu, zwłaszcza za pośrednictwem części mediów społecznościowych. Często dane te są sfałszowane lub zniekształcone. Sytuacja jest tu zupełnie inne niż w przypadku tradycyjnych danych (np. pozyskiwanych w ramach działalności urzędów statystycznych), których wiarygodność jest wysoka. Trudno tu również mówić o próbach losowych (nieistotne, o jakim schemacie losowania mowa). Również to oznacza brak możliwości stosowania klasycznej teorii wnioskowania statystycznego.

Drugi element jakości narzędzi uczenia maszynowego to same metody. Tutaj kluczowe jest zastosowanie metod właściwych do rozwiązania danego problemu. W przypadku znanych od wielu lat metod uczenia maszynowego, takich jak metoda wektorów nośnych, drzewa decyzyjne, metoda najbliższych sąsiadów, itp., jest to relatywnie proste. Metody te są transparentne, zrozumiałe, interpretowalne, a zatem ekspert może bez problemu dokonać oceny adekwatności danej metody z punktu widzenia analizowanego problemu naukowego.

Inaczej jest w przypadku sztucznych sieci neuronowych, będących częściami składowymi dużych modeli językowych. Jest tam widoczny brak transparentności (a zatem również brak wyjaśnialności), co oznacza duża niepewność w stosowaniu tych modeli do konkretnych problemów badawczych.

Wnioski:

1. Integracja nie tylko zbiorów danych (a także wyników badań) wymagać będzie częstszego odwoływania się do personalistycznej interpretacji prawdopodobieństwa kosztem interpretacji częstościowej.

2. Potrzebne jest bardziej kompleksowe podejście do wnioskowania oraz bardziej zniuansowane w odniesieniu do testowania hipotez:

(***The tests themselves give no final verdict but as tools help the worker who is using them to form his final decision*** – pisali w 1928 r. Jerzy Spława-Neyman i Egon Pearson)

3. Metody z grupy *data analytics* stanowią nie alternatywne, lecz komplementarne rozwiązanie w analizach prawidłowości występujących w zbiorach danych.

DZIĘKUJEMY ZA UWAGĘ