

ЗА ДАННИТЕ И ПЪТЯ КЪМ НАУКАТА ЗА ДАННИТЕ

Павел Азълов
САЩ

Резюме. Първите прояви, в които могат да се търсят корените на науката за данни, са от около преди шестдесет години. Но тя израства и се формира в сегашния си вид през последните две десетилетия, силно повлияна и стимулирана от *големите данни*. В този смисъл, настоящата статия се явява естествено продължение на статията за големите данни от предишния брой на списанието.

Тук акцентът е върху хронологията на основните събития, които съществено допринесоха за утвърждаването на *науката за данните*. Успоредно с това е представена и кратка справка за автори от различни научни среди, обществени организации и държавни комисии, участвали в дебата при изясняване на въпроса „Какво е наука за данните?“. Специално е отбелязан подходът на James Gray като четвърти подход в научните открития. Накратко са представени водещите идеи при изграждането на специалността *Наука за данните* в няколко университета в САЩ.

Ключови думи: наука за данните; специалист по наука за данните; образование по наука за данните; големи данни; аналитика на данните

„Всъщност какво точно е науката за данните? Както може би се досещате, отговорът зависи от кого го искате.“

Xiao-Li Meng

1. Въведение

Търсенето на знания е вграден механизъм в природата на човека. Но често, поради липса на знания, той е безсилен да отговори на много въпроси, да обясни много факти и да разбере какви са връзките между тях. С развитието на информационните технологии в края на миналото столетие хората започват да осъзнават, че в данните има неразкрити знания, които в определени случаи могат да помогнат да се намерят отговорите на много въпроси. Ето в това трябва да се търси основата, върху която се е зародила и изградила съвременната *наука за данните*.

Първите прояви от зараждането ѝ са от около преди шестдесет години, но тя израства и се оформя в последните две десетилетия, силно повлияна и стимулирана от *големите данни*. Настоящата статия е естествено продължение на статията за големите данни от предишния брой на списанието (Azalov, 2020). Двете статии са обединени от темата за данните, и по-конкретно от компютърните технологии за тяхното първоначално натрупване, съхранение и обработка, както и от резултатите на специалистите от различни области, които изграждат алгоритми за извличане на знанията, „дълбоко скрити“ в тях.

Науката за данни се появи сред вече утвърдени науки и появата ѝ не се приема еднозначно. Това е точно синтезирано в мнението на Xiao-Li Meng, цитирано в началото. В това ни убеждават и редица специалисти от академичните среди, от бизнеса и от информационните технологии, отразени в статии, книги, учебници, университетски програми, интернет блогове, интервюта, обяви за позиции в областта на науката за данните и много други. В някаква степен то се потвърждава и от становищата, представени в тази статия. Съществено е да се отбележи, че цитираните автори са специалисти от най-различни научни области: математика и статистика, компютърни науки и информационни технологии, бизнес науки и инженерни науки, физика и астрономия, химия и биология, медицински науки и социални науки, дори и философия.

Акцентът в статията е върху хронологията на основните събития, довели до формирането на днешния облик на областта наука за данните. Едновременно с това е представено и разнообразието от вижданията на отделни специалисти, научни и обществени организации и държавни комисии по въпроса „*Какво е науката за данните?*“. Точно това е и причината след цитирането на всяко мнение да се представи и „визитната картичка“ на съответния автор. Съвсем преднамерено цитираните индивидуални мнения в статията не са коментирани.

Езикови бележки

Науката за данните се „ражда“ в САЩ. Това показват наличните документи. Ето защо и първите термини се появяват на английски език. Идеите в тази наука се възприемат бързо от специалистите в много държави и с това започва езиковата адаптация на съответната терминология. Подобен процес е характерен и за други науки. Не беше много отдавна, когато бяха адаптирани английските термини от компютърните науки. Добре известно е, че дословният превод невинаги е подходящ, понякога е твърде неприемлив, а понякога се губят нюанси, което съвсем не е без значение. По тази причина специалните термини от науката за данните, както и всичките имена на лица и съответните им научни степени и звания, имената на университетите и учебните дисциплини, а също и имената на компаниите са записани на оригиналния им език.

Каква е целта на статията?

Повод за тази статия е актуалността на темата, която става все по-популярна и в България^{25, 26}. Причина за това е преди всичко необходимостта от реални приложения на науката за данните. Силното развитие на съвременните компютърни и информационни технологии определено дава мощен тласък за бързото ѝ утвърждаване. И това е естествено, защото те са „технологичният гръбнак“ на самата наука за данните.

Числата, получени от справката в Google за срещането на термина *наука за данните*, записани на английски и български език, са показателни за актуалността на темата:

Data Science: About 3 970 000,000 results

Наука за данните: About 15 400 000 results

Статията не изчерпва темата за пътя към науката за данните. Съществуват много книги, статии и документи, посветени на науката за данните в по-обширен формат, включително и придружени с много исторически данни. Такива са например статиите в (Donoho, 2015) и (Cao, 2017).

По-долу е „дадена думата“ на учените, които първи посочват пътя към науката за данните, и на тези, които са допринесли за изграждането ѝ; на професорите, които вече я преподават в университетите; на тези, които успешно решават и преодоляват проблемите с *големите данни*, както и на тези, които я прилагат в научните лаборатории и бизнес компаниите.

2. Пионерите в науката за данните

Настоящият разказ започва с кратка справка за идеите и първите стъпки на тези, които преди няколко десетилетия първи виждат в съзнанието си очертанията на една нова наука – *науката за данните*. Те са пионерите, които посочват промените в класическата статистика и предсказват резултатите, които могат да се получат от натрупаните с години данни в компании, държавни агенции и научни лаборатории. Те полагат основите на обществото, което днес се управлява от данните.

1962: John Tukey

В статия си *The Future of Data Analysis* (Tukey, 1962) в обем от над 60 страници John Tukey описва идеите си за реформа в статистиката. Статията е публикувана в *The Annals of Mathematical Statistics* през март 1962 г. и в нея авторът обяснява, че обхватът на статистическите изследвания е вече твърде тесен, и посочва как трябва да бъде пренасочен. Вижданията, които J. Tukey представя в статията, са твърде революционни за времето си. Изминалите 60 години оттогава показват, че до голяма степен неговите идеи са напълно актуални и се вписват в това, което днес наричаме *наука за данните*. Нещо повече, J. Tukey казва, че с обновяването на *анализа на данните* тази област

може вече да се смята за нова самостоятелна наука, а не като отделен клон на математиката.



John W. Tukey (1915 – 2000)

John Tukey е виден американски математик. Получава магистърска степен от Brown University и докторска степен по математика от Princeton University. Работи известно време в AT&T Bell Laboratories. На 35 години е вече професор в Princeton University и основава Катедрата по статистика, на която става ръководител. Той е пионер в областта на анализа на данните, има принос в алгоритъма за бързо преобразоване на Фурие, в разработването на метода за графично представяне на групи от данни (box plot) и др.

Носител е на редица награди и отличия. През 1973 г. е награден с National Medal of Science in Mathematical, Statistical, and Computational Sciences, USA.

1974: Peter Naur

Peter Naur (1928 – 1916) е един от пионерите на компютърните науки. През 2005 г. той получава *ACM A. M. Turing Award*¹⁾ „...за фундаменталния принос към проектирането на езика за програмиране и компилатор на Algol 60 и за изкуството и практиката на компютърното програмиране“. Той въвежда термина *datalogy* (Naur, 1966), с който е именувал своя курс по компютърни науки, определяйки *datalogy* като наука за обработване на данните²¹⁾. Той въвежда и обяснява и термина *datamation*. Това е един първи опит за използване на термина *наука за данните*, но в по-различен смисъл от този, който се използва днес (Voulgaris, 2014).

1997: Chien-Fu Jeff Wu

Chien-Fu Jeff Wu има бакалавърска степен по математика от National Taiwan University и докторска степен по статистика от University of California, Berkeley. Понастоящем е Coca-Cola Chair in Engineering Statistics, Professor, at Georgia Institute of Technology. През ноември 1997 г. изнася встъпителна лекция в University of Michigan на тема *Statistics = Data Science*²²⁾. В нея той апелира за радикална промяна в съдържанието на статистиката като приложение и академична област. Настойчиво препоръчва и смяна на името ѝ.

2001: William S. Cleveland

От 2001 г. насам все по-често се говори за науката за данните като една нова научна област с много приложения. Заслугата за това се приписва на William Cleveland, който публикува статията *Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics* (Cleveland, 2001). Тя

представлява обръщение към академичната общност по статистика за преосмисляне и освежаване на статистиката чрез радикална промяна към по-широко определение на областта. В нея той представя и конкретен план за преориентиране на изследванията по статистика. Тези идеи доказват стойността си в течение на времето, защото и днес звучат актуално, въпреки че години наред са останали недостатъчно разбрани и са имали твърде малък ефект.



William S. Cleveland (1943)

William Cleveland е Shanti S. Gupta Professor of Statistics and Courtesy Professor of Computer Science at Purdue University, Department of Computer Science. Той е математик с диплома от Princeton University и доктор по статистика от Yale University. Бил е ръководил на департамента по статистика към Bell Labs, а след това е избран за професор по статистика и компютърни науки в Purdue University. Работил е върху методи за визуализиране на данни, изграждане на статистически модели и алгоритми за машинно обучение. Носител е на много

награди и отличия от реномирани американски и световни асоциации по статистика и информационни технологии⁴⁾.

3. За четвъртия подход към научните открития

Преди малко над десетина години се появяват публикации, в които се пише за четвъртия подход в научните открития. Началото се поставя през 2007 г., когато под редакцията на Т. Hey, S. Tansley и К. Tolle излиза от печат книгата *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Hey & al, Eds., 2009). Тя представлява сборник от статии на специалисти от много научни области, обединени от посланието на James Gray. По същество книгата е посветена на Grey, вдъхновена от неговите изследвания в областта на появилия се неотдавна термин *наука за данните*.

Кой е James Gray?

През януари 2007 г. James Gray изнася доклад на заседанието на NRC-CSTB (The National Academies of Science, Engineering, Medicine – Computer Science and Telecommunications Board¹⁶⁾. Темата е *eScience – A Transformed Scientific Method*. Докладът е впечатляващ и съдържа революционни за времето си идеи.

Следват някои от акцентите в доклада, резюмирани накратко в две групи. *Важни изводи, породени от големите данни* (Szalay, 2009):

- данните нарастват експоненциално в много науки;
- съществуват огромни масиви от данни, но няма решения за някои проблеми, свързани с тях;
- има нужда от нови инструменти: „микроскопи“ и „телескопи“ за данни;

- данните стават все по-отворени;
- науките се ръководят от данни (големи и малки);
- данните променят не само науката, но и обществото;
- големите данни сближават статистиката, компютърните науки, физическите науки и науките за живота;
- чрез данните се извършва научна революция в откритията.



James Nicholas „Jim” Gray (1944 – 2012)

James Gray е програмист, експерт по бази от данни, инженер и изследовател. Той е първият доктор по компютърни науки от University of California, Berkeley (1969 г.) За приносите му, свързани с транзакциите при базите от данни (ACID, OLAP и др.), и лидерството му при внедряването на софтуерни системи през 1998 г. е удостоен с престижната награда ACM A. M. Turing Award. Получава и наградата IEEE Charles Babbage. Работил е в големи високотехнологични компании като Bell Labs, IBM Research, DEC и накрая в Microsoft. Изследванията

му след 2002 г. са върху eScience проект, в който акцентът е върху решаването на научни проблеми чрез интензивното използване на данни. Съвместно с Alex Szalay – физик от Johns Hopkins University, Gray формулира „Четвъртата парадигма“ като подход за научни открития. На 28 януари 2007 г. Gray излиза с яхтата си в открито море, но не се завръща. На същата дата след пет години е обявено, че е починал.

Препоръки към NRC-CSTB за насърчаване на (Szalay, 2009):

- развитието на общи лабораторни системи за управление на данни;
- изследванията за управление на научните данни, за анализа на данните, за визуализирането на данните и разработването на нови алгоритми;
- създаването на дигитални библиотеки, които да поддържат с данни някои науки по същия начин, по който Националната медицинска библиотека подкрепя биологичните науки;
- разработването на нови инструменти за създаване на документи и модели за публикуване;
- развитието на библиотеки с научни цифрови данни и научна литература.

Точно това е грандиозната цел, към която се стреми Jim Gray – светът да бъде устроен по начин, по който цялата научна литература да е достъпна онлайн, всички научни данни да са онлайн и да се осъществи тяхната интеграция.

На корицата на цитираната по-горе книга (Hey et al, Eds., 2009) Бил Гейтс (Microsoft Corporation) пише:

| С идеите си Jim Gray насочи мислите на хората за това как данните и софтуерът могат да предефинират подходите в научните изследвания.

Висока е и оценката на редица специалисти, измежду които е и Douglas Bruce (Professor of Bioanalytical Sciences in the School of Chemistry at the University of Manchester) (Hey et al, Eds., 2009). Всеобщо е мнението, че едно от най-големите предизвикателства пред науката в XXI век е как да се реагира на ерата, в която науката се управлява от данни. Измежду утвърдените досега подходи за научни открития: *емпиричен* – чрез наблюдения и експерименти; *теоретичен* – чрез теории, и *изчислителен* – чрез модели и симулации, се налага и подходът чрез *интензивно използване на данни*, който вече се приема и широко навлиза в научните среди. Осъзнава се, че скоростта, с която ще се развива една научна дисциплина, до голяма степен ще зависи не само от нейното вътрешно развитие, но и от отварянето ѝ към използването на нови технологии, свързани с данните: бази от данни (малки и големи), визуализирането на данните, с облачните технологии и др. (Hey et al, Eds., 2009).

Подходът Data first

В публикацията си *Is the Fourth Paradigm Really New?* (Kastens, 2012) Kim Kastens (Lamont Research Professor, Lamont-Doherty Earth Observatory of Columbia University; Adjunct Full Professor, Columbia University Department of Earth & Environmental Sciences) описва няколко научни проекта, разработени преди около 40 години. При тях са изследвани данни, натрупани от много години. Като резултат изследването е потвърдило правилността на преди това създадени модели за световните океани. В заключение, авторката пише, че подходът на интензивното използване на данни в научните изследвания (тя използва термина *data-first*) може и да не е съвсем нов, но това не го омаловажава. Напротив, резултатите от проведените изследвания за „тектонната революция на плочите“ подкрепят посланието за четвъртия подход, както и това, че този начин на правене на наука може да бъде изключително продуктивен.

И още няколко мнения

По време на конференцията Nvidia GPU (GTC 2019) през април 2019 г. Jen-Hsun Huang (Huang, 2019) поднови десетилетния дебат за съвременната наука за данни, заявявайки, че тя е „четвъртият стълб“ като метод за научни открития. Практически той напълно подкрепя идеите на Jim Gray и най-общо добавя (Woodie, 2019), че сега методите на изследвания с данни придобиват сила благодарение на големите данни и на пробива в алгоритмите за машинното обучение:

...Те направиха възможно данните да са четвъртият стълб на научното откритие. Сега ние решаваме проблеми, които по-рано бяха невъзможни.

Тук само ще отбележим, че това са мисли, изказани от основателя и президент на компанията Nvidia¹⁷⁾.

Идеята за „четвъртия стълб“ е твърде революционна, за да не се прояви към нея известен скептицизъм. Ето две мнения.

Oliver Schabenberger: (Woodie, 2019)

...научният метод, който използвах с моите студенти, е близък и скъп за сърцето ми. Не виждам науката за данни като четвърти стълб на науката. Виждам това като метод за получаване на прозрение от данните. Преди я наричахме статистика. Преди я наричахме извличане на данни. И какво по същество е различното сега?

Moshe Vardi (Vardi, 2010), (Woodie, 2019)

Не ми е много удобно да смятам, че науката може да има нов крак на всеки няколко години... Всъщност вярвам, че науката все още има само два крака – теория и експерименти. Гледната точка за „четири крака“ изглежда предполага, че научният метод се е променил по фундаментален начин. Твърдя, че не научният метод се е променил, а по-скоро как се прилага.

Завършваме тази дискусия с мнението на голям колектив от Renaissance Computing Institute, University of North Carolina at Chapel Hill, отразено в публикацията *Scientific Discovery in the Era of Big Data: More than the Scientific Method* (Schmitt et al, 2015).

...Предлагаме широка рамка за съвместна работа, като се създаде мрежа от бази от знания за научни открития. В проекта ще бъдат включени не само научни експерти, но всички заинтересовани страни, включително политиците, представителите на индустрията и дори обикновените граждани, като по този начин ще се даде възможност и на нетрадиционните източници на данни да стимулират научните открития.

Не може да се отрече, че в тези редове, написани осем години по-късно, прозират идеите на Jim Gray за създаването на библиотеки от данни и научна литература, които да се използват за научни открития.

От приведените примери се вижда, че мненията за науката от данни като подход за научни открития са разнопосочни. Всъщност мненията за това какво представлява самата наука за данните, също не са еднозначни. Но това е и едната от целите на тази статия – да се представят различните гледни точки по тази актуална в днешно време тема „Какво е наука за данните?“.

4. Първи прояви на науката за данните

Идеите на Tukey и Cleveland не се възприемат масово и еднозначно. Въпреки това следващите години бележат първите, все още неукрепнали стъпки към науката за данните в няколко направления (Zhu et al, 2015), (Cao, 2017).

– **2002 г. Списания.** През 2002 г. CODATA публикува първото реферирано списание – *Data Science Journal*. (CODATA⁵ – Committee on Data of the International Science Council).

– **Лаборатории и институти по Data Science:**

2008 г. – The University of Technology Sydney преименува лабораторията Knowledge Discovery Lab в *Data Sciences & Knowledge Discovery Research Lab*;

2012 г. – създава се Institute for Data Sciences and Engineering at Columbia University, New York, USA;

2012 г. – създава се Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China;

2011 г. – **първа конференция по Data Science** в University of Illinois at Urbana-Champaign, USA;

2012 г. Първи университетски курс по Introduction to Data Science: Columbia University – преподавател е Rachel Schutt (Davidian, 2013). Тя е съавтор на учебника *Doing Data Science* (O' Neil, C., R. Schutt, 2013).

Да анализираме анализаторите

С течение на годините в статистическата колегия започва да се усеща известно неудобство от разширяващия се обхват на науката за данните. Появяват се нови методи и технологии, които все още не са в ежедневието „репертоар“ на много от класическите статистики. Естествено е, че новата роля на статистиката изисква нови знания и промяна в характера и стила на работа. Това добре се вижда от някои публикации през годините 2013 – 2014. Ето една извадка от (Harris et al., 2013):

Смятаме, че термини като Data Scientist, Analytics и Big Data са трудно разбираеми. Хората, които се занимават с тази работа, идват от по-традиционни и утвърдени области: статистика, машинно обучение, бази от данни, изследване на операциите, социални или физически науки и др. Всички тези професии имат ясни очаквания за това, на което практикуващият е способен да прави; добре са дефинирани образователните и кариерните пътеки, включително специализации, базирани на пресечната точка на наличните умения и нуждите на пазара. Това все още не е вярно за новите модни думи... Според нас една от причините, поради които „науката за данни“ и други модни думи се появили наскоро, е появата на нови евтини технологии и техники за работа с много големи масиви от данни... Нашите проучвания показват, че повечето специалисти от областта на науката за данните рядко работят с терабайти или по-големи данни.

Не сме ли ние науката за данни?

Мнението на авторите на книгата *Analyzing the Analyzers* (Harris et al., 2013) не се споделя от академичната общност по статистика и това е добре отразено в статията *Aren't we Data Science?* на Marie Davidian (Davidian, 2013).

С първите вълни на науката за данните, които вече навлизат в университетските аудитории, Marie Davidian е силно загрижена за съдбата на катедрите по статистика. Това е поводът за статията ѝ в AMSTAT NEWS. В една голяма част от нея тя преразказва разговора си с Rachel Schutt – асистент от Columbia University, изградила и провела първия курс по наука за данните (Davidian, 2013):

„**Науката за данните е тук, за да остане**“, казва Рейчъл... В заключение, тя отбелязва, че се страхува, че академичните катедри по статистика могат да бъ-

дат разглеждани като остарели и постепенно да бъдат премахнати през следващото десетилетие, ако не се развием и посрещнем това предизвикателство – науката за данни няма да изчезне... Тя ме убеди, че инициативата на ASA Big Data е съществена стъпка към справянето с някои от тези предизвикателства на ниво асоциация, към полагането на основите за усъвършенстване на учебните програми... Ние не сме наука за данните, но трябва да играем критична роля.

Коя е Marie Davidian? По времето, когато М. Davidian пише статията, тя е J. Stuart Hunter Distinguished Professor of Statistics at North Carolina State University, а през 2013 г. е президент на American Statistical Association (ASA).

Защо ни е нужен нов термин, когато статистиката съществува от векове?

Това заглавие е извадка от статията *Data Science and Prediction* на Vasant Dhar, публикувана в едно от най-реномираните списания по компютърни науки – Communications of the ACM (Dhar, 2013). В нея той пише:

Терминът „наука за данните“ се използва все по-често заедно с термина „големи данни“. Какво означава „наука за данните“? Има ли нещо уникално в това? Какви умения трябва да притежава специалистът data scientist, за да бъде продуктивен в настъпващата дигитална ера... Науката за данните е фокусирана около данните и в анализа на данните... **Защо тогава ни е нужен нов термин, когато статистиката съществува от векове?** Фактът, че сега разполагаме с огромни количества данни, сам по себе си не трябва да оправдава необходимостта от нов термин.“

На въпросите от по-горе авторът сам дава обстоен отговор, който се състои от отделни части. По-долу привеждаме част от представените аргументи.

Данните са най-важната част от науката за данните, а те вече не са само структурирани. Преобладаващата част от тях са полуструктурирани и неструктурирани. Техният анализ е различен и това налага използването на дисциплини от други науки. Такава дисциплина е например *лингвистиката*.

На базата на езици за работа с полуструктурирани данни като JSON, CSV и HTML компютрите могат да обменят данни и да общуват помежду си. Това позволява вземането на решения в по-големи мащаби и все по-често компютърът става инициатор и ръководител на процес, без да е подпомаган от човек.

Машинното (само)обучение е дисциплина от изкуствения интелект, т.е. от областта на компютърните науки, а тя вече има централно място в науката за данните. Знанията и методите, използвани в машинното обучение, попадат в три широки класа, единият от тях е статистика и вероятности. Вторият клас се отнася до компютърните науки и по-конкретно до алгоритми и структури от данни, бази от данни (малки и големи), езици за програмиране и други.

Към третия клас влизат уменията на специалистите по науката за данните. Тук става въпрос за хора с творческо и нестандартно мислене. Точно те са специалистите по данни, които успяват да предложат решения на трудни проблеми, от които да се получат интересни и неочаквани резултати.

Кой е Vasant Dhar? Той е професор в New York University, Stern School of Business and the Center for Data Science и основател на SCT Capital Management – хедж фонд в New York City, базиран на машинно обучение. Има бакалавърска степен по технологии от Indian Institute of Technology в Delhi, магистърска по философия от University of Pittsburgh и докторска по изкуствен интелект от University of Pittsburgh. Изследванията му са в областта на ИИ с основен въпрос „Кога да се доверявам на системите за ИИ за вземане на решения?“. Преведен на езика на финансите, този въпрос може да се преформулира така: „Ще доверите ли парите си на робот?“. Тези въпроси са особено важни за автономните системи, базирани на машинно обучение, които се учат и се адаптират към текущите данни⁶⁾.

Да притежаваме науката за данни

В поздравителното си слово към участниците в годишната конференция на IMS (Sydney, 9 – 14 юли 2014 г.) президентът на IMS Bin Yuse се обръща към участниците с доклада *Let us own Data Science*. Тя произнася пламенна реч, в която апелира за пренасочване на изследванията в катедрите по статистика в духа на науката за данните:

Сега „Статистиката“ обхваща голям набор от дейности, така че не е много позната за външните хора. Те и до днес я разбират в първоначалния ѝ смисъл – броене или преброяване, въпреки че сега много статистики са част от екипите за изследванията на рака, геномиката, невронауката и астрономията. За разлика от това свързаните с данните подобласти на компютърните науки приеха нови имена за направления: Artificial Intelligence, Data Mining, Machine Learning, Deep Learning.

За Bin Yu науката за данните е естествено сливане на компютърното и статистическото мислене в ерата на големите данни и те, статистиките, трябва да „притежават“ тази наука. Data Science е новоприетият термин за справяне със съвременните проблеми с данните в тяхната цялост и начинът да го притежаваме, не е труден (Yu, 2014):

Просто трябва да следваме стъпките на Turner, Carver, Hollerith, Cochran and Tukey и да работим върху реални проблеми. Съответната методология/теория ще дойде по естествен път. Истинските проблеми на днешния ден идват от геномиката, невронауката, астрономията, нанонауките, компютърните и социалните науки, персонализираната медицина/здравеопазването, финансите, правителството . . .

Коя е Bin Yu? Тя е Professor of Statistics, UC Berkeley (2001 –), Professor of Electrical Engineering & Computer Science, UC Berkeley (2006 –), Chair, Department of Statistics, UC Berkeley (2009 – 2012), President, IMS (2013 – 2014)²⁴⁾.

5. Един въпрос с много отговори

„Какво е наука за данните?“ е основният въпрос на статията, но той има немалко отговори. За да стане по-ясно защо тази тема продължава да се дис-

кутира, ще представим гледните точки и мненията на специалисти от академичните среди, бизнеса и сектора на ИТ, прелиствайки страници от книги, статии и документи. Някои от становищата са изказани само преди няколко месеца.

Wikipedia

Wikipedia е най-популярната и лесно достъпна платформа за справки и затова се обръщаме най-напред към нея. Тези, които отдавна не са отваряли тази страница, може би ще забележат, че дефиницията е съществено обновена. В сегашния ѝ вариант четем следния текст:

Науката за данните е интердисциплинарна област, която използва научни методи, процеси, алгоритми и системи за извличане на знания и прозрения от много структурирани и неструктурирани данни. Науката за данните е свързана с извличането на данни и на „големи данни“. Науката за данните е „концепция за обединяване на статистиката, анализа на данни, машинното обучение и свързаните с тях методи“ с цел „разбиране и анализ на действителните явления“ с данни. В нея се използват техники и теории от много други области като математика, статистика, компютърни и информационни науки.

Data Science Association (DSA), САЩ

DSA е нестопанска професионална асоциация на работещите в областта на науката за данните⁷⁾. Тя издава списанието *International Journal of Data Science*, в което се публикуват статии с широк спектър от теми на науката за данните. Публикуват се и позиции за работа на специалисти по науката за данните, информация за конференции и др. DSA има изградени комисии по различни направления. Една от тях е комисията за стандарти, отнасящи се до науката за данните. Комисията е разработила и кодекс на асоциацията, първият раздел на който включва дефиниции на основни понятия. Ето тяхната дефиниция, която е много кратка, но и твърде обща:

Науката за данните е термин, който включва научно проучване за създаване, валидиране и трансформиране на данни с цел извличане на стойност от тях.

National Institute of Standards and Technology (NIST), САЩ

На уеб-страницата на NIST¹⁴⁾ е записано, че основна цел на Института е да се насърчават иновациите в САЩ и индустриалната конкурентоспособност чрез подобряване на науката, стандартите и технологиите. В документа “*Big Data Interoperability Framework, V7*”¹⁵⁾ е дадена следната дефиниция за термина *наука за данните*:

Науката за данните е извличането на полезни знания директно от данни чрез процес на откриване или на формулиране и тестване на хипотези.

Науката за данните – пресечна точка на три области

The Data Science Design Manual е учебник, в който авторът представя много теми от науката за данните, като въвежда и съответния апарат по математика, статистика, машинно обучение, визуализиране на данните, големи данни и др. Още в началото след кратък анализ на съвременната роля на големите данни и облачните технологии авторът пише (Skiena, 2017):

Какво е науката за данните? Като всяка нова развиваща се област, тя все още не е напълно дефинирана... Мисля, че науката за данните лежи в пресечната точка на компютърните науки, статистиката и съответните области на приложение. От компютърните науки идват машинното обучение и високоефективните изчислителни технологии за справяне с мащаба. От статистиката произлизат дългогодишната традиция в проучвателния анализ на данните, тестовите за значимост и визуализирането на данни. Конкретните области на приложения на бизнеса и на различните науки определят специфичните предизвикателства, стоящи пред науката за данните.

Авторът на учебника – Steven Skiena, дава отговор и на въпроса „*Защо науката за данните стана актуална сега, когато изграждащите я други области са утвърдени отдавна?*“. Той вижда отговора най-вече в прилагането на новите технологии за съхранение и управление на големите данни, в развитието на облачните технологии, в разработването на нови алгоритми за машинно обучение, включително и в средата на големите данни, в напредъка в компютърното зрение, в обработването на естествените езици и др.

Кой е Steven Skiena? Той е професор, доктор по компютърни науки в Stony Brook University и директор на Institute for AI-Driven Discovery and Innovation, който е в рамките на College of Engineering and Applied Sciences. S. Skiena е съосновател на General Sentiment – компания за социални медии и анализи на новини. Той е разработил технологична платформа на система за обработка на естествен език и система за анализ на настроеността, наречена *Лидия*. Понастоящем работи върху проектирането на алгоритми с приложения в биологията.

В много статии и книги специално се дискутират изискванията към кандидатите за работа в областта на науката за данните. По-долу са представени вижданията за професионалната характеристика на двама специалисти – единият от сферата на информационните технологии, а другият от академичните среди.

Позицията специалист по науката за данните във Facebook

Една от първите компании, която обявява позиция за специалист по наука за данните (*data scientist*), е Facebook. Ето и текста, в който се представят изискванията към специалиста по наука за данните (Hammerbacher, 2009):

Във Facebook почувствахме, че традиционните наименования като Business Analyst, Statistician, Engineer и Research Scientist съвсем не съответстват на

специалността на човек, когото искахме за нашия екип. Натовареността, съответстваща на позицията, е разнообразна: всеки ден може да се наложи един от членовете на екипа да създава последователност от операции като част от многостепенна обработка на голям обем от данни с езика Python, да проектира тест за проверка на хипотеза, да извършва регресионен анализ на проби от данни с езика R, да проектира и реализира определен алгоритъм за обработка на данни или определена услуга в Hadoop, както и да представя резултатите от нашите анализи на други членове на организацията по ясен и кратък начин. За да опишем този списък от умения, необходими за изпълнение на това множество от задачи, създадохме позицията Data Scientist.

Този текст е извадка от книгата *Information Platforms and the Rise of the Data Scientist*, в която Jeff Hammerbacher е един от редакторите ѝ и е автор на пета глава.

Кой е Jeff Hammerbacher? Той има бакалавърска степен по математика от Harvard University. Едва 23-годишен, една година преди да завърши, той е вече на работа във Facebook. Mark Zuckerberg – президент и изпълнителен директор на Facebook, му дава високото звание *главен научен изследовател*, определя му екип от 100 служители и му възлага да разработи информационната платформа на компанията. Основната цел е анализът на данните в социалната мрежа. Броят на потребителите на Facebook е много голям, а обемът на данните е огромен. Задачата е трудна, но в рамките на проекта Jeff Hammerbacher създава две системи за работа с големи данни. На базата на Hadoop са разработени системите с отворен код Hive¹⁰⁾ и Cassandra³⁾. През 2008 г. той напуска Facebook и става съучредител на Cloudera – световно известна софтуерна платформа за данни.

А защо Марк Зукърбърг възлага тази отговорна задача точно на Jeff Hammerbacher? Отговор на този въпрос е представен по-долу като сбит преразказ на епизод от ранните му години, които определят неговото бъдеще на изявен специалист по науката за данни, в направлението на големите данни.

Два месеца преди да завърши първата си година като студент Jeff е уволнен като касиер в един магазин. Но в това уволнение той вижда и една добра възможност. Вместо да каже на родителите си, че е уволнен, той продължава да излиза от къщи всеки ден, облечен както преди: черни панталони, черни обувки, бяла риза и т.н. В действителност той посещава градската библиотека и чете книги и списания.

Още на 17 години Jeff е безумно любопитен и иска да научи как работи човешкият мозък, как заболява и как се възстановява. Той се интересува не само от способността на мозъка да реагира към вътрешни и външни сигнали (данни), но и от ролята му да съхранява информация, да генерира планове и да създава нова информация. Той осъзнава, че библиотеките и мозъците на хората са примери за информационни платформи, и естествено, когато през 2006 г. се присъединява към екипа на Facebook, е натоварен да изгради информационната платформа на компанията.

Накрая на този кратък разказ Jeff не пропуска да сподели, че хитрият му план да посещава библиотеката вместо магазина, от който е уволнен, не проработва дълго. Една вечер на излизане от библиотеката не намира автомобила си. Впоследствие разбира, че майка му е разкрила неговата схема и тайно се е погрижила да го премести. По време на дългата разходка до дома си той се замисля за това колко е важно човек да гледа със скептицизъм на собствените си решения. И още нещо, той съветва никой никога да не се опитва да надхитри майка си.

Разказът е много поучителен. Още като студент у него живее идеята за изграждане на разпределена информационна платформа, която е аналог на възможностите на много човешки мозъци. Делят го по-малко от десетина години от деня, когато той реализира идеята си.

Специалист, който задава уникални въпроси за данни

Авторът на *Data Science: Theories, Models, Algorithms, and Analytics* Sanjiv Ranjan Das (Das, 2016) казва, че науката за данните е областта, в която теориите се реализират с помощта на данни, а понякога и с големи данни. Той посочва, че областите, в които бъдещите специалисти по науката за данните би трябвало на насочат усилията при подготовка си, са компютърните науки, бизнесът, икономиката, статистиката и конкретна област, свързана с прилагането ѝ. Интересно е да се спомене мнението му относно някои изисквания към специалистите по наука за данните:

Значи искате да бъдете „специалист в областта на науката за данните“? Няма широко прието определение кой е такъв специалист. Гледната точка в тази книга е, че такъв специалист е човек, който задава уникални, интересни въпроси за данните, основани на формална или неформална теория, за да генерира точни и полезни идеи.

Тук авторът съвсем точно е цитирал мисъл на Georg Cantor „*Да формулираш въпрос в математиката, има по-голяма стойност от предлагане на решението му*“.

Не по-малко важна и интересна е и бележката му за екипната работа при изследвания в областта на науката за данните:

Добрите въпроси идват от много дисциплини и най-добрите отговори вероятно ще дойдат от хора, които имат интереси в множество области, или от екипи, които добре съчетават разнообразни знания и умения.

Кой е Sanjiv Ranjan Das? Той е Professor of Finance and Data Science, Santa Clara University, Leavey School of Business Department of Finance. Има докторска степен по финанси от New York University (1994). През 2000 г. получава магистърска степен по компютърни науки от University of California, Berkeley.

Представяме и мненията на двама специалисти в областта на науката за данните, които са професори в различни научни области. Статиите са публикувани в Harvard Data Science Review (HDSR).

Анализът е в основата на науката за данните

Темата за анализа на данни е добре представена в стотици публикации. Тук се спираме на кратката статия *The Data Life Cycle* на Jeannette Wing, която описва всяка една от фазите на жизнения цикъл на данните: генериране, събиране, обработване, съхранение, анализ, визуализиране и интерпретиране. След последната фаза тя представя ролята на човека, който по съответен начин ще използва резултата от изследването. По-долу специално е даден коментарът за фазата *анализ* (Wing, 2019):

Когато повечето хора се замислят какво представлява науката за данните, за тях това означава анализ на данни... изчислителни и статистически техники за анализ на данни, алгоритмите и методите, които са в основата на ИИ, извличането на данни, машинното обучение и статистическите изводи... Със сигурност анализът на данните е в основата на науката за данни.

Коя е Jeannette M. Wing? Тя е Avaneessians Director of the Data Science Institute и Professor of Computer Science at Columbia University. Била е Corporate Vice President of Microsoft Research, докато е преподавател в Carnegie Mellon University, а след това е била също и Assistant Director of CISE at NSF.

Какво не е науката за данни?

Започваме с нестандартния подход на Xiao-Li Meng за дефиниране на термина наука за данните, представен в *Data Science: An Artificial Ecosystem* (Meng, 2019).

Всъщност какво точно е науката за данни? Както може би се досещате, отговорът зависи от кого го искате. Някои казват, че науката за данните е компютърната наука. Други смятат, че науката за данни е просто статистиката... Когато нещо е твърде обширно, можем да използваме една математическа стратегия и да го опишем чрез допълнението му – тоест това, което не е. Сега ще изброя пет неща, които не са наука за данните.“

По-долу са записани петте твърдения, всяко от които е добре обосновано и обяснено в статията, включително и с примери.

Първо, науката за данните не е само машинно обучение или просто статистика.

Второ, науката за данните не се отнася само за прогнозиране.

Трето, науката за данните не се отнася само до анализ на данни.

Четвърто, науката за данните не е дисциплина, която се намира само в полето на STEM (Science, Technology, Engineering and Mathematics).

Пето и най-критичното, науката за данните дори не е една дисциплина.

Кой е Xiao-Li Meng? Той е Whipple V.N. Джоунс професор по статистика в Harvard University. Получава бакалавърска степен по математика от Fudan University, Китай (1982 г.) и докторска степен по статистика от Harvard University (1990 г.). Бил е професор в University of Chicago. В Harvard University

е бил ръководител на катедрата и декан на Graduate School of Arts and Sciences (2012 – 2017 г.). Избран е за най-добър статистик на възраст под 40 години от COPSS (Committee of Presidents of Statistical Societies). Той е главен редактор на Harvard Data Science Review²³).

Науката за данни през следващите 40 години

Книгата *Foundations of Data Science* е университетски учебник, в която авторите Avrim Blum, John Hopcroft и Ravindran Kannan представят теорията на науката за данните, изградена върху основите на математиката. И да си припомним – върху основите на математиката възникнаха и компютърните науки. В началото на 60-те години на миналия век в катедрите по математика на много университети по света започват да се преподават и първите курсове по компютърни науки. Ето какво е записано в предговора на учебника (Blum et al, 2020):

Докато традиционните области на компютърните науки остават изключително важни, все повече изследователската работа в бъдеще ще бъде свързана с използването на компютрите за разбиране и извличане на полезна информация от големите данни, възникващи в приложенията, а не само с разработването на компютърни системи, решаващи конкретни добре дефинирани задачи. Имайки предвид това, в тази книга ние написаме теорията, която очакваме да бъде полезна през следващите 40 години, точно както въвеждането на теорията на автоматите, алгоритмите и свързаните с тях теми даде предимство на студентите през последните 40 години. Една от основните промени е увеличаването на акцента върху **теория на вероятностите, статистиката и числените методи**.

Книгата вече е адаптирана като университетски курс в някои водещи университети с наименованието *Математически основи на науката за данните*.

Кои са авторите на учебника?

John Edward Hopcroft (1939 г.) е специалист по теоретични основи на компютърните науки. Автор е на книги и учебници, от които са се учили няколко поколения специалисти по компютърни науки. Завършил е електроинженерство и има бакалавърска степен (1961 г.) от Seattle University, магистърска степен (1962 г.) от Stanford University и докторска степен (1964 г.) от Stanford University. Той е IBM Professor of Engineering and Applied Mathematics, Cornell University. Носител е на няколко десетки награди от най-висок ранг в областта на компютърните науки. За фундаментални постижения в проектирането и анализа на алгоритмите и структурите от данни през 1986 г. е удостоен с ACM A.M. Turing Award.

Avrim Blum (1966 г.) има докторска степен от MIT (1991 г.). Бил е Professor and Chief Academic Officer of Computer Science at Carnegie Mellon University (1991 – 2017 г.). От 2017 г. е Professor and Chief Academic Officer at Toyota Technological Institute at Chicago. Научната му интереси са в областта на теоретичните основи на компютърните науки и по-конкретно – по машинното обучение и теорията на алгоритмите.

Ravindran Kannan (1953 г., Индия) има докторска степен от Cornell University. Бил е William K. Lanman Jr. Professor of Computer Science and Professor of Applied Mathematics at Yale University. Преподавал е в MIT и CMU. Kannan е Principal Researcher at Microsoft Research India, където ръководи изследователска група по алгоритми. Научните му интереси са в областта на теоретичните основи на компютърните науки и по-конкретно – по алгоритми, дискретна математика и оптимизации.

6. Науката за данните – университетска специалност

Въпросът, свързан с образованието по науката за данните, е изключително важен и в редица университети започват да се откриват отделни специализации, а на други места – цели специалности по науката за данните.

През септември 2015 г. University of Michigan обявява инициативата „Наука за данните“ (DSI) и планира да инвестира 100 милиона долара, както и да разкрие 35 нови преподавателски позиции. На откриването на учебната година вицепрезидентът по академичните въпроси Martha Pollack (Pollack, 2015) казва:

Науката за данните се превърна в четвърти подход към научното откритие в допълнение към експериментирането, моделирането и изчислението. За да стимулира иновациите, DSI ще стартира инициативи в четири критични интердисциплинарни области, които надграждат нашите силни страни в транспортните изследвания, науките за здравето, аналитиката на обучението и социалните науки.

Коя е Martha E. Pollack? Тя е професор по компютърни науки, специалист по ИИ и е 14-ият президент на Cornell University, където работи от април 2017 г. Била е в борда на директорите на IBM¹⁹⁾.

Ролята на академичната общност в образованието по науката за данните

Статията *The Role of Academia in Data Science Education* е изключително актуална. Публикувана е само преди няколко седмици в Harvard Data Science Review (Irizarry, 2020). В нея е засегната група от въпроси, но тук накратко се представя само част от тези въпроси.

В статията се отбелязва, че науката за данните е интердисциплинарна област и че тя е „чадър“ – термин, въведен в (Meng, 2019). Под този чадър са се „настанили“ няколко области. Едната се занимава с хардуер и проблеми, свързани с инфраструктурата за съхранение на данни (*data engineering*). Другата е насочена към анализа на данни и машинното обучение. В третата област попадат софтуерните специалности. Техните системи са от голяма полза, защото съществено улесняват работата по разнообразни проекти.

Друг акцент в статията, който всъщност е основният, пада върху образованието по науката за данните. „Повечето дефиниции не водят до консенсус

между учените по науката за данните относно основните ѝ принципи, а такава дефиниция е необходима, ако университетите ще предлагат научни степени по наука за данните“, пише авторът и продължава:

Най-напред трябва да се дефинира ядрото на съответна учебна програма. Следдипломните специализации в направленията, свързани с анализа и управлението на данни и със статистиката и компютърните науки, не дават необходимия практически резултат при работа по реални проекти... Разработените магистърски програми генерираха приходи и това беше първият им приоритет. Като резултат днес имаме десетки университети, които предлагат тези степени. Но за какво точно са подготвени тези студенти? Какво предлагат тези нови програми, че досегашните не са го постигнали? Като се има предвид, с някои изключения, че при създаването на нови програми не са постъпили на работа нови преподаватели и в много случаи са разработени малко или никакви нови учебни материали, не е ясно дали магистърската степен по науката за данните, както се предлага в момента, представлява това, което работодателите търсят.

В статията се правят редица препоръки за разработването на магистърски програми, като например:

- обучението трябва да е в няколко различни направления – да не забравяме, че науката за данните е „чадър“;
- курсовете по статистика и машинно обучение трябва да са с практическа насоченост, т.е. фокусът да не е в теорията. Курсовете по компютърни науки по алгоритми, оптимизация и структури от данни да се фокусират върху практически задачи и съответна програмна реализация;
- студентите трябва да получат практически умения по програмиране. Курсовете по програмиране да се фокусират върху езици като R, Python, Spark, Keras и TensorFlow за целите на машинно обучение, както и към езици като C++, необходими за изграждането на инфраструктура на проектите. Като се има предвид масовото използване на базите от данни, всички студенти трябва да са запознати поне със SQL;
- да се разработят курсове, в които се разглеждат реални проекти и в които се решават въпроси от отворен тип;
- особено полезни са крайните проекти, които трябва да се ориентират в разработването на софтуер, приложим в областта на науката за данните.

Кой е Rafael Irizarry? Той е Professor in Biostatistics at Harvard T. H. Chan School of Public Health; Professor at Biostatistics and Computational Biology, Dana-Farber Cancer Institute; Chair, Biostatistics and Computational Biology, Dana-Farber Cancer Institute. Измежду основните курсове по науката за данните има и такива, които са специализирани в областта на биомедицината: Professional Certificate in Data Science, HarvardX Biomedical Data Science Online Curriculum, Statistics and R for the Life Sciences Online Course, Data Analysis for the Life Sciences XSeries, Genomics Data Analysis XSeries, Introduction to Data Science и др.

Специалността „Наука за данните“ в Georgetown University

Програмата е структурирана в групи по следния начин⁸⁾:

- основни курсове (6 на брой);
- четири групи от избираеми курсове: аналитика на данните (12); математика (6); компютърни науки (13); допълнителни избираеми курсове (7) и група от курсове, ориентирани предимно към области, в които се прилага науката за данните.

Следва списъкът от всички курсове, структуриран по групи.

Основни курсове

- Advanced Programming Topics
- Introduction to Data Analytics
- Massive Data Fundamentals
- Scientific and Analytical Visualization
- Probabilistic Modeling and Statistical Computing
- Statistical Learning

Аналитика на данните ● избираеми курсове

- *Effective Presentation for Technology & Science*
- *Computational Linguistics* ● *Advanced Python*
- *Data Ethics Privacy & Security*
- *Structures and Algorithms for Analytics*
- *Time Series*
- *Optimization*
- *Natural Language Processing for Data Analytics*
- *Neural Networks and Deep Learning*
- *Storytelling, Data Communication and Decision Science*
- *Advanced Machine Learning*
- *Relational and Semi-Structured Databases and SQL Programming.*
- *Internship*

Математика – избираеми курсове

- Mathematics of Climate
- Applied Longitudinal Analysis
- Mathematics of Social Networks
- Stochastic Simulation
- Bayesian Statistics
- Categorical Data Analysis

Компютърни науки ● избираеми курсове

- *Image Processing*
- *Statistical Machine Translation*
- *Intro to Cryptography*
- *Data Privacy*
- *Network Security*
- *Advanced Algorithms*
- *Probabilistic Proof Systems*
- *Theory of Computation*
- *Automated Reasoning*
- *Statistical Machine Learning*
- *Computer Vision*
- *Advanced Data Bases*
- *Web Search and Sensemaking*

Допълнителни избираеми курсове

- Machine Learning for Bioinformatics
- Clinical Trials and Experimental Design
- Social Network Analysis
- Regression Analysis
- Database Development and Management
- GIS and Spatial Data Modeling for Public Policy
- The Policy Issues of Big Data and Artificial Intelligence

Курсове по компютърните науки в специалността Наука за данните¹⁸

Магистърската програма по Information and Data Science at the UC Berkeley заслужава специално внимание²). Тя е изградена на две нива, които в тази статия не са представени. Това, което не може да не ни впечатли, е списъкът със софтуерните системи, които студентите изучават и прилагат в курсовете от специалността. Тези курсове включват: (1) програмиране, предимно на Python и R и техните библиотеки; (2) системи, ориентирани за работа с бази от данни; (3) среди за работа с големи данни; (4) системи за визуално представяне на данни; (5) системи за анализ на данни и аналитика на големи данни; (6) облачни технологии; (7) системи за анализ в специфични области. Ето и пълния списък от курсове:

Python (basic programming, libraries for linear algebra, plotting, machine learning: numpy, matplotlib); R and R libraries; Relational Javascript; databases; Hadoop; MapReduce; Spark; AWS; Tableau; D3; Illustrator; R; ggplot2; Highcharts; Visit; Distributed Storage; Ethereum Blockchain; Docker; CouchDB; Cassandra; OpenStack Swift; Apache Solr; BVLC Caffe; Nvidia/Digits; Keras; IBM Watson; GATK.

Специалността „Наука за данните“ в Pennsylvania State University (Penn State)

Науката за данни е мултидисциплинарна област и е естествено специалистите, които работят в тази област, да са от различни дисциплини. Прекалено амбициозно е да се мисли, че човек може да се обучава така, че да бъде експерт по науката за данните и да е в състояние да се справя с всички предизвикателства, възникващи по проекти от областта на науката за данните. Учебната програма, представена по-горе, е отличен пример за реализиране на обучението на специалистите по науката за данните. Обединяването на основните дисциплини в една група и на избираемите дисциплини в отделни групи дава възможност за профилиране вътре в самата специалност.

Идеята за профилиране може да има и друга реализация. Например в Penn State специалността „Наука за данните“ се предлага под формата на *Intercollege Undergraduate Major* и е изградена в три варианта, които са в рамките на три колежа, всеки от които включва няколко факултета¹⁸).

- **Applied Data Sciences**: изучават се принципите, методите и инструментите за управление, интегриране, анализ, визуализиране и прогнозно моделиране на масивни и сложни данни.

- **Computational Data Sciences**: изучават се дисциплини от компютърните науки, които са в основата на науката за данните, включващи проектирането, внедряването и анализа на софтуер, който управлява големите масиви от данни и използва изчислителната мощ на многоядрен хардуер.

– **Statistical Modeling Data Sciences:** изучават се статистическите модели и методи, които са необходими за откриване и валидиране на модели в наборите от данни с всякакви размери.

Kaggle – платформа за обучение и състезания по науката за данните

Kaggle¹²⁾ е технологична платформа за сътрудничество, която обединява хора с интереси в областта на науката за данни. Тя предоставя възможности за обучение, обмен на информация, обява на работни места и участия в *състезания по машинно обучение*. В Kaggle членуват милиони специалисти от цял свят с различни нива на подготовка.

Задачите в състезанията са реални проблеми и се предлагат от компании, които определят конкретните правила и наградата за победителя. За да се почувства сериозността на проблемите, които се решават в състезанията, по-долу накратко е представен проектът Zillow Prize²⁸⁾. Това е конкурсната тема, дадена преди няколко години. Резултатът, който се очаква от състезанието, е алгоритъм за машинно обучение относно предсказването на цената на недвижими имоти.

Компанията, формулирала темата, е Zillow. За пресмятането на цената на даден имот тя използва модела Zestimates²⁷⁾. Първоначалната грешка на модела⁹⁾ е била 14%, но след поредица от подобрения е сведена до 5%. С крайния резултат от това състезание компанията цели да подобри модела си за предсказване. Наградата за победителя е *един милион долара*. Той се определя няколко месеца след крайния срок за предаване на проектите, за да има достатъчно време за тестването на моделите с реалните данни от пазара на недвижимо имущество.

Компанията Kaggle е създадена през 2010 г. в Австралия от Anthony Goldbloom. В момента тя е дъщерна компания на Google.

7. Заключителни бележки и опит за обобщение

От направения преглед се вижда, че няма единно мнение относно дефиницията на термина *наука за данните*, но различията не са съществени. От приведените мнения на различни специалисти се оформя изводът, че те се дължат на това, че науката за данните се формира в рамките на вече утвърдили се научни области. Процесът на сходимост към единно мнение вероятно няма да е бърз. От една страна, необходимо е време да се приемат минималните изисквания към учебните програми (стандарт) за университетско образование по *наука за данните*. От друга страна, трябва да се инвестира в обновяването на преподавателските екипи, които да реализират тези стандарти, а добре е известно, че това е трудна задача.

Но още през 2012 г. Rachel Schutt казва (Davidian, 2013): „*Науката за данни е тук, за да остане*“. А тя не само остава, но се утвърждава, разширява се и вече е неразделна част от обществения живот и бизнеса на всички нива.

Подобни заключения се съдържат в редица публикации, в които се твърди, че обхватът и въздействието на науката за данните ще продължат да се разширяват значително през следващите десетилетия, тъй като научните данни стават повсеместно достъпни. А това всъщност включва част от препоръките на Jim Gray за развитието на библиотеки с научни цифрови данни и научна литература, казани по повод на четвъртия подход за научните открития (Heu et al., Eds., 2009).

Нека още веднъж да погледнем университетската учебна програма от по-горе. Тя оформя една добра рамка за обучение по науката за данните, така че точната дефиниция на термина *наука за данните* вече не е толкова съществен въпрос. Това е така, защото по-важно е да се определи какво се влага в образованието на младите хора, които са утрешните специалисти по науката за данните. Те ще имат не само отлична представа за нейната дефиниция, но и професионални умения, за да я прилагат, и едва ли между тях ще има спор относно дефиницията на понятията, защото всички те ще са специалисти от една наука – науката за данните.

Науката за данните е нова и модерна и предлага добра бъдеща кариера. Такава е оценката по света на много държавни агенции, академични и университетски институти, технологични и бизнес компании. Не на последно място, това се потвърждава и от високите финансови стимули, които се предлагат на добрите специалисти по науката за данните^{11, 20}. Всичко това е повод младите хора да бъдат насърчавани към избор на една добра професионална кариера в тази област.

Послеслов

Работата върху настоящия текст вече привършваше, когато COVID-19 вирусът започна да „набира скорост“ по цял свят. Една сутрин в един от българските онлайн вестници беше публикувана кратката статия *Белият Дом призова да се използва изкуствен интелект в борбата с коронавируса*. Ето няколко съществени извадки¹³⁾:

Вашингтон призова изследователите в САЩ и други части на света да използват изкуствен интелект за намиране на отговорите на ключовите въпроси за вируса... В съобщение Белият дом поиска експертите да разработят нови технологии за извличане на текст и данни, за да се намерят отговори на „високо приоритетните“ научни въпроси относно COVID-19... Белият дом обединява сили с технологичните гиганти Microsoft и Google за събиране на широка база от данни за вируса с надежда компютрите да открият информация, която е пропусната от човека ... Американското правителство поиска да се използва изкуствен интелект за анализ на 29 000 научни статии за вируса... Около 13 000 статии за коронавируса са включени в новата база от данни във формат, който улеснява анализирането им със софтуер. Базата от данни съдържа частични текстове, такива като обобщения от други 16 000 статии.“

С това привършва и разказът ни за пътя към *науката за данните*. На ход са „интелигентните“ алгоритми и специалистите по науката за данни от цял свят.

БЕЛЕЖКИ

1. ACM Turing Award: The A.M. Turing Award, sometimes referred to as the *Nobel Prize of Computing*, was named in honor of Alan Mathison Turing (1912 – 1954) – a British mathematician and computer scientist. He made fundamental advances in computer architecture, algorithms, formalization of computing, and artificial intelligence. <https://amturing.acm.org/>
2. Master of Information and Data Science at the UC Berkeley. <https://datascience.berkeley.edu/academics/curriculum/>
3. Cassandra: <http://cassandra.apache.org/>
4. Cleveland, W. (2004) <https://www.cs.purdue.edu/people/faculty/wsc/>
5. Data Science Journal (DSJ) <https://datascience.codata.org/>
6. Dhar, V. <https://www.stern.nyu.edu/faculty/bio/vasant-dhar>
7. DSA: Data Science Association <https://www.datascienceassn.org/>
8. Data Science at Georgetown. https://analytics.georgetown.edu/?gclid=EAIaIQobChMIxPSSoMvn5QIVGovICh2RdQHNEAAYAiAAEgJzR_D_BwE#
9. George Box: Всички модели са грешни, но някои са полезни.
10. Hive: <https://hive.apache.org/>
11. How much does a Data Scientist make in the United States? <https://www.indeed.com/career/data-scientist/salaries>
12. Kaggle <https://www.kaggle.com/>
13. News.bg (2020) <https://news.bg/world/beliyat-dom-prizova-da-se-izpolzva-izkustven-intelekt-v-borbata-s-koronavirusa.html>
14. NIST Mission, Vision, Core Competencies, and Core Values <https://www.nist.gov/about-nist/our-organization/mission-vision-values>
15. NIST Big Data Interoperability Framework: Volume 7, Standards Roadmap <https://www.nist.gov/publications/nist-big-data-interoperability-framework-volume-7-big-data-standards-roadmap-version-2>
16. NRC-CSTB: The National Academies of Science, Engineering, Medicine – Computer Science and Telecommunications Board. <https://sites.nationalacademies.org/CSTB/index.htm>
17. Nvidia <https://www.nvidia.com/en-us/>
18. Data Science PSU <https://datasciences.psu.edu/>
19. Pollack, M. https://en.wikipedia.org/wiki/Martha_E._Pollack
20. Data Scientist Salary <https://www.ziprecruiter.com/Salaries/Data-Scientist-Salary>
21. Sveinsdottir, E., Frøkjær, E. (1988). Datalogy – The copenhagen tradition of computer science. BIT 28, 450–472. <https://doi.org/10.1007/BF01941128>

22. Wu, J. Statistics = Data Science. <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjhv oTcuJXoAhX9hXIEHTCbCmoQFjAAegQIBRAB&url=https%3A%2F%2Fwww2.isye.gatech.edu%2F~jeffwu%2Fpresentations%2Fdatascience.pdf&usq=AOvVaw1h4MxES-MZwgxNTyChcY6Q>
23. Xiao-Li (2020) <https://datascience.harvard.edu/xiao-li-meng>
24. Yu, B. (2014) <https://www.stat.berkeley.edu/~binyu/Site/Welcome.html>
25. Sofia University: Големите данни в полза на интелигентно общество. <https://projects.uni-sofia.bg/ProjectDetails.aspx?ProjectId=2778>
26. Sofia University. https://www.dnevnik.bg/bulgaria/2019/10/07/3973039_sofiskiiat_universitet_shte_analizira_golemi_danni_zal/
27. Zestimate <https://www.zillow.com/zestimate/>
28. Zillow Prize <https://www.kaggle.com/c/zillow-prize-1>
29. Снимките в текста са от архива на интернет.
30. Всички интернет страници, цитирани в статията, са посетени и отворени за последен път на 03.30.2020 г.

ЛИТЕРАТУРА

Азълов, П. (2020). За данните и пътя до „големите данни“. *Математика и информатика*, година LXIII, брой 3, стр. 235 – 270.

REFERENCES

- Azalov, P. (2020). About the data and the road towards the big data. *Mathematics and Informatics*, vol. LXIII, issue 3, pp. 235 – 270.
- Blum, A., J. Hopcroft, and R. Kannan. (2020). *Foundations of Data Science*. Cambridge University Press. ISBN: 9781108755528.
- Cao L. (2017). Data science: A comprehensive overview. *ACM Computing Surveys* 50, 3, Article 43. DOI: <http://dx.doi.org/10.1145/3076253>
- Cleveland, W. (2001) Data Science: an action plan for expanding the technical areas of the field of statistics. *International statistical review*, 69(1), pp. 21 – 26.
- Davidian, M. (2013) Aren't We Data Science? Column of ASA President Marie Davidian in *AmStat News* <http://magazine.amstat.org/blog/2013/07/01/datascience/>
- Das, S. (2016). Data Science: Theories, Models, Algorithms, and Analytics *Licensed under the Apache License*, Version 2.0.
- Dhar, V. (2013). Data Science and Prediction *Communications of the ACM*, vol. 56 no. 12, pp. 64 – 73.
- Donoho, D. (2017) *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, 745 – 766. <https://doi.org/10.1080/10618600.2017.1384734>

- Hammerbacher, J. (2009). Information Platforms and the Rise of the Data Scientist. Beautiful Data, Edited by Toby Segaran and Jeff Hammerbacher. O'Reilly Media.
- Harris, H., S. Murphy, and M. Vaisman (2013) *Analyzing the Analyzers. An Introspective Survey of Data Scientists and Their Work*. O'Reilly Media ISBN: 978-1-449-36824-1.
- Huang, J. (2019) https://en.wikipedia.org/wiki/Jensen_Huang
- Irizarry, R. A. (2020). The Role of Academia in Data Science Education . *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.dd363929>
- Hey, T., S. Tansley, and K. Tolle. Eds. (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Eds., Redmond, VA: *Microsoft Research*, ISBN 978-0-9825442-0-4. <http://fourthparadigm.org>.
- Tukey, J. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*. vol. 33, No. 1, 1962, pp. 1 – 67.
- Kastens, K. (2012). Is the Fourth Paradigm Really New? <https://serc.carleton.edu/earthandmind/posts/4thparadigm.html>
- Meng, X. (2019). Data Science: An Artificial Ecosystem. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.ba20f892>
- Naur, P. (1966). The Science of Datalogy. *Communications of the ACM*, vol. 9, No 7, p. 485 <http://www.tamerkhraisha.com/files/datalogy.pdf>
- O'Neil, C. and R. Schutt. (2013). *Doing Data Science*. ISBN: 978-1449358655 O'Reilly Media
- Pollack, M. (2015). U-Michigan launches \$100 million Data Science Initiative <https://news.umich.edu/u-michigan-launches-100-million-data-science-initiative/>
- Schmitt, C., Cox, S., Fecho, K., Idaszak, R., Lander, H., Rajasekar, A. and Thakur, S. (2015). Scientific Discovery in the Era of Big Data: More than the Scientific Method. A RENCI *White Paper* vol. 3, no. 6 <http://dx.doi.org/10.7921/G0C82763>
- Skiena, S. (2017). *The Data Science Design Manual*. Library of Congress Control Number: 2017943201 ISBN 978-3-319-55443-3 DOI 10.1007/978-3-319-55444-0.
- Szalay, A. (2009). *Data Driven Discovery in Science: The Fourth Paradigm*. Alex Szalay, The Johns Hopkins University.
- Vardi, M. (2010). Science Has Only Two Legs. *Communications of the ACM*, vol. 53 no. 9, p. 5 https://www.win.tue.nl/~wstomv/edu/hci/downloads/2010-2011/Science_Has_Only_Two_Legs.pdf DOI:10.1145/1810891.1810892
- Voulgaris, Z. (2014). *Data Scientist: The Definitive Guide to Becoming a Data Scientist*. Technics Publications ISBN: 9781935504696.

- Wing, J. M. (2019). The Data Life Cycle. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.e26845b4>
- Woodie, A. (2019). Is Data Science the Fourth Pillar of the Scientific Method? <https://www.datanami.com/2019/04/15/is-data-science-the-fourth-pillar-of-the-scientific-method/>
- Yu, B. (2014). IMS Presidential Address: Let us own Data Science. <http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/>
- Zhu, Y and Y. Xiong, (2015). Towards Data Science. *Data Science Journal*, 14: 8, pp. 1 – 7 <https://datascience.codata.org/article/10.5334/dsj-2015-008/>

ABOUT THE DATA AND THE ROAD TOWARDS THE DATA SCIENCE

Abstract. The roots of data science started to grow approximately sixty years ago. The current state, however, began to form only two decades ago, highly influenced by the emergence of *big data*. As such, this article is a natural continuation of the similarly themed big data article, published in the preceding issue of this journal.

The focus of this article is on the events that materially contributed towards the prominence of data science. Additionally, the article discusses the debate around *What is data science?* had by academics, public organizations and government entities over the years. An emphasis is given to the James Gray paradigm, known as the fourth approach to scientific discoveries. The leading ideas in establishing Data Science as a major in several US universities are outlined briefly.

Keywords: data science; data scientist; data education; big data; data analytics

✉ **Dr. Pavel Azalov**

ORCID iD: 0000-0003-4313-3383

Professor Emeritus of Computer Science

Penn State, Hazleton, U.S.A.

E-mail: pkal0@psu.edu