ГОЛЕМИТЕ ДАННИ И ОБРАЗОВАНИЕТО В ИНФОРМАЦИОННОТО ОБЩЕСТВО

Ренета П. Бърнева

State University of New York at Fredonia, USA,
barneva@fredonia.edu

***Резюме**: За големи данни се заговори сравнително отскоро. Под това понятие се разбират данни, които не могат да се обработват с традиционните алгоритми и софтуер. Има много прогнози, че анализът на големи данни ще стане необходим във всяка област, както днес са уменията да си служим с компютърна и комуникационна техника. В тази статия разглеждаме как да се подготвят специалисти, умеещи да анализират големи данни, като в частност разглеждаме обучението на професионално, университетско и училищно ниво.*

***Ключови думи**: анализ на големи данни, професионално образование, висше образование, средно образование.*

1. Въведение

Анализът на данни се е използвал много преди компютърната ера, но с развитието на компютърната и комуникационната техника, и особено с драстичното намаляване на цените им и с повсеместното им използване, се натрупаха огромни количества данни. Те идват от камерите за сигурност, от мобилните комуникации, електронната поща, социалните мрежи, записите на транзакции в банковата система и търговската мрежа, както и от измервателни уреди и устройства в метеорологията, астрономията, транспорта, медицината и др. Според някои източници, повече данни са били записани през последните няколко години, отколкото от началото на човешкото общество до преди няколко години [1]. Освен това, има тенденция съществуващите по-стари данни - документи, книги и фотографии да се сканират и също да се представят в цифрова форма, която освен удобства при съхранявянето, създава отлични условия за обработката и извличането на ценна информация от тези данни.

Оказва се, обаче, че поради големият им обем и съхраняването им в различни формати, обичайните алгоритми не могат да се прилагат директно за обработването им [2]. Има и допълнителни трудности, свързани със съхранението, организацията и унифицирането на данните. Ето защо новите модели и методи, свързани с големите данни, се превърнаха в обект на научните изследвания.

Според Кронке [3] нарастването на значението на анализа на данни е закономерно. Тази прогноза той прави въз основа на анализа на елементите на информационните системи - хардуер, софтуер, данни, процедури и хора - и

важността на всеки от тях, която нараства в този ред, в развитието на информационното общество. Според него, първият елемент на информационните системи - хардуерът - е от съществено значение през 60-те години на миналия век и компании като IBM, произвеждащи хардуер, станаха много влиятелни. По-късно софтуерът стана жизненоважен и през 80-те години компании като Microsoft, произвеждащи софтуер, придобиха голямо значение. Днес данните са от изключително значение и компании като Google и Yahoo придобиха голямо влияние. Очаква се скоро процедурите да станат най-значими в този процес.

Данните, съхранени в първичен вид, сами по себе си, не дават особено преимущество. Важно е да се извлече ценна информация от тях, което става чрез методи за анализ. Самият анализ зависи от спецификата на обекта, представен чрез данните. От голямо значение е анализиращият да има познания в конкретната предметна област – медицина, селско стопанство, астрономия, метеорология – за да знае коя информация би била ценна. Има дори прогнози, че способността да се анализират големи данни ще създаде разслоение в обществото – онези, които не са в състояние да го правят, ще стават все по-бедни, докато други ще станат по-богати и ще имат повече власт поради способността си да анализират данните [4]. Това вече може да се види в някои области, които на пръв поглед са далеч от анализа на данни, като спорта - в момента сериозните отбори трябва да наемат по няколко анализатора на данни, за да бъдат конкурентоспособни [5,6].

Ето защо има голяма нужда от експерти по анализ на данни. Според някои източници, към 2018 год. ще има недостиг от 140 000 - 190 000 специалисти [7]. Така пред образованието възниква задачата да се подготвят специалисти, които могат да се справят с анализа на големи данни. През последните 3-4 години университетите по света започнаха да работят в това направление и набързо бяха разкрити редица магистърски програми [8].

В тази статия споделяме някои мисли относно това, какво трябва да се изучава в областта на анализа на големи данни и на какво ниво. Според нас обучението в областта ще се развие подобно на обучението по информатика, затова в следващата точка припомняме как се е развило обучението в тази област. Последователно разглеждаме обучението на университетско ниво и в средното образование, като обръщаме внимание на един друг аспект на големите данни – как те могат да се използват за подобряване на самия процес на обучение. Също така отбелязваме и някои потенциални опасности при използване на големите данни в образованието.

2. Кратък преглед на обучението по информатика и паралел с обучението по анализ на големи данни

Първоначален интерес към нови области се заражда обикновено у някои учени. Често новите области са с интердисциплинарен характер и възникват в резултат на сътрудничество между учени от различни дисциплини. След това към изследванията им се привличат аспиранти и евентуално се защитават дисертации. Междувременно, ако областта е перспективна и продължава да се развива, се подготвя курс или курсове на магистърско ниво.

По този начин се развиват повечето от новите дисциплини, но те често се оказват доста тесни и образованието се изчерпва до нивото на отделен избирателен магистърски курс, четенето на който след известно време се преустановява.

Информатиката, обаче, се оказа област, която силно се разви за много кратко време. Появиха се различни направления в нея – изкуствен интелект, компютърна графика, езици за програмиране и транслатори, паралелно програмиране, и др., които бяка предпоставка за утвърждаване в средата на 80-те години и на програми на бакалавърско ниво. През 80-те и 90-те години други дисциплини – математика, физика, инженерни науки – включиха в програмите си задължителни курсове по информатика.

Трябва да се отбележи, че България е една от първите страни в света, които въведоха програми по информатика и в резултат на това нашите специалисти в тази област са на много високо ниво. Нещо повече, през 80-те години се въведе задължително обучение по информатика и в средното училище – нещо, за което се говори едва сега в Съединените щати. Особено добра подготовка по информатика на ниво средно образование се дава в специализираните математически гимназии и техникуми по електроника или сродни дисциплини. Важно е да се отбележи, че средното образование в България не целеше да подготви програмисти, а по-скоро да развие умения за използване на компютърна и комуникационна техника, които да се използват в различни професии или в ежедневието.

В края на 90-те години, когато компютрите навлязоха широко в бита и на работното място, в обявите за работа се появи изискването за "компютърна грамотност", като се имаше предвид използването на системи за текстообработка, електронни таблици и електронна поща. В началото на 21 век, обаче, тези изисквания отпаднаха, тъй като това е нормално очакване, както очакването за традиционна грамотност.

Според нас обучението в областта на големи данни ще се развие по подобен начин. Много университети вече предлагат курсове по методи за анализ на данни; десетки университети откриха магистърски програми. Все още са рядкост бакалавърските програми в тази област – по скоро се правят

специализирни направления в програмите по информатика. Тези програми, обаче, изискват знания по висша математика и програмиране и поради тази причина не са подходящи за специалности с хуманитарна или артистична насоченост. Вероятно тенденцията ще бъде да се създадат системи, така че всеки да може да извършва анализ на данни. Самите данни ще са достъпни онлайн, така че локално да не е необходима голяма изчислителна мощност. Поради това ще бъде от по-голямо значение да се внедри обучение по анализ на данни във всички дисциплини, отколкото да се създават чисти специалисти по анализ на данни. Последните ще се реализират по-скоро в изследователската работа за развиване на нови методи, алгоритми и системи за анализ на данни, отколкото пряко в различните приложни области. Поради тази очаквана тенденция, ще е важно да се въведе и обучение в средното образование по анализ на данни.

3. Обучение по анализ на данни на университетско ниво

Ед Лазовска [9] нагледно описва знанията на специалистите на бъдещето като имащи формата на буквата П, докато знанията на специалистите сега имат формата на буквата Т. В "Т" вертикалната черта означава конкретната област, в която човек получава университетско образование, а хоризонталната е области, към които знанието се разширява. В "П" втората вертикални черта отговаря на уменията да се анализират данни, на които ще се опира разширението в други области. С други думи, Лазовска смята, че всеки човек ще се нуждае от знания как да анализира данни.

По-долу ще разгледаме какво трябва да включва обучението по анализ на големи данни в университетските програми.

3.1. Магистърски и докторски програми

В тези програми се очаква да се подготвят специалисти, които ще разработват методи и софтуер за анализ на данни. Подходящо е да се включат курсове по статистика, извличане на знания, бази от данни и знания, изкуствен интелект и самообучаващи се системи, езици като R, Hadoop, Scala, визуализация, системи и интерфейси като Apache Spark, Knime, SAS, SPSS, оптимизационни модели и алгоритми върху тях. Също така е целесъобразно да се разгледат приложения в различни предметни области като сигурност, медицина, метеорология, селско стопанство и др.

Освен тези магистърски програми, целящи да обучават специалисти за развитие на теорията и инструментите за анализ на данни, е уместно да се разработят и програми за обучение на специалисти в дадена област на социалните и хуманитарните науки или изкуствата, които да обработват данни в тяхната специфична предметна област. Такива програми са значително по-

малко (вж. напр. [10]). Целта там е да се подготват специалистите да работят със софтуер за обработка на данни, да се запознаят с методи за визуализация на данните и изготвяне на отчети, да се изучат различни математически модели и практическите задачи, които могат да се сведат до тях, както и софтуера за решаване на задачи, сведени към такива модели.

3.2. Бакалавърските програми

Наскоро се появиха и бакалавърски програми по анализ на данни, въпреки, че засега броят им е по-ограничен. Подобно на началния етап в развитието на програмите по информатика, те обикновено се комбинират със съществуващи програми по статистика, информатика или бизнес и се преподават от съответната катедра [11]. Освен изброените по-горе предмети, които се изучават в магистърските програми, са включени и дисциплини на по-ниско ниво, като въведение в информатиката, въведение в статистиката, алгоритми и структури от данни, езици за програмиране, основни математически курсове, бизнес стратегия, управление на бизнес процеси, бизнес и маркетингово разузнаване, изследване на операциите и други. Конкретните дисциплини зависят от насочеността на програмата и катедрата, която я ръководи – математическите катедри наблягат на статистика, комбинаторика и изследване на операциите; информатичните катедри – на езици за програмиране, изкуствен интелект, методи за извличане на знания, бази от данни, самообучаващи се системи и др.; а бизнес катедрите – на използването на системи за анализ на данни, на визуализацията, анализа на данни за повишаване на конкурентоспособността и сигурността на данните. Трябва да се отбележи, че бизнес организациите бяха едни от първите, които започнаха да използват анализи на големи данни, натрупани от транзакции, за да получат предимство пред конкурентите си на пазара и в резултат много катедри по бизнес обучаваха студентите на така наречената business intelligence.

Тъй като се очаква в скоро време анализът на големи данни да стане необходим във всяка област, както понастоящем са необходими уменията за работа с компютри и компютърни системи, е важно университетите да въведат съответни курсове, достъпни за студентите от всички дисциплини. Тези курсове не трябва да изискват задълбочени знания и умения по математика, статистика и/или информатика, а по-скоро да наблегнат на работата със системи за анализ на данни, системи за визуализация и изготвяне на доклади. Важно е да се разбере какви възможност дават системите, като специално внимание се обърне на видовете модели и графики и за какви видове задачи могат да се използват. Университетите, които въведат навреме такива курсове в образованието по различни специалности, ще имат определено предимство пред останалите.

3.3. Професионално обучение

Освен подготовката на обучаваните в момента, трябва да се помисли за големия брой специалисти, които вече работят в различни области, които се нуждаят от подготовка в използването на големи данни. Някои от тях, обикновено в началото на професионалната си кариера, вероятно биха се записали в магистърски програми, но повечето хора в средата на кариерата си и особено старшите специалисти едва ли биха имали тази възможност. За тези случаи е добре да се използват кратки професионални курсове. Те трябва да са тясно-профилирани, специализирани за всяка конкретна област и специфичните цели в нея. Както и при общообразователните университетски курсове, ударението трябва да пада на използването на софтуерни системи. Тези курсове, обаче, могат да са с по-конкретна приложна насоченост, тъй като се очаква да се преподават на специалисти от една област и дори от едно предприятие.

Особен интерес представляват курсовете за техническия персонал, който ще подготвя отчети за мениджърите. Там е целесъобразно да се обърне внимание на средствата за визуализация на данните, на съхраняване на данните, сигурност и достъп до тях и подготвяне на интерактивни отчети.

Професионалните курсове са обикновено с гъвкав формат – интензивни, в рамките на един ден, или по-продължителни, като на всяка тематична единица се дават сведения за отделен независим модул, тъй като се предполага, че не всички участници могат да присъстват на всички занятия. Може да се очаква поява на необходимост и от по-продължителни курсове за преквалификация на кадри.

3.4. Осигуряване на ресурси за обучението

Докато въпроса за ресурсите обикновено не е актуален при професионалното обучение, тъй като то се покрива от фирмите или от специални фондове, често стартирането на нови университетски програми, особено на бакалавърско ниво е проблематично, поради нуждата да се наемат преподаватели и подготвят курсове преди да са записани студенти за програмата. Освен това, като правило, отначало програмите започват с неголям брой студенти. По-долу даваме няколко идеи, които могат да се използват в този случай.

- *Комбиниране на съществуващи курсове за други дисциплини.* Това е тактика, която широко се използва при създаването на нови университетски програми. Все пак, необходимо е да се разработят и специфични курсове. На по-късен етап, когато новата програма е набрала достатъчно студенти, курсовете се преработват само за нея. Трудността в случая е координацията, понеже анализът на данни има

мултидисциплинарен характер и много от курсовете ще се предлагат от различни катедри.

- *Създаване на между-университетски програми.* Комбинирането на курсове и специалисти от различни университети облекчава нуждата от ресурси. Като част от обучението може да се прибавят и различни форми на стажове в предприятия. Курсовете в този случай се предлагат чрез синхронно или асинхронно дистанционно обучение. Тази идея и конкретната програма, която сме предложили и реализирали в Университета на Щата Ню Йорк, сме описали подробно в [12].
- *Привличане на спонсори от индустрията, финансови или търговски институции.* При този модел спонсорите могат да се възползват от безплатно обучение на свои служители, както и от практиканти, които да работят в предприятието в рамките на стаж. Някои научни фондации също предлагат начално финансиране на такива програми [13]. Този модел е подходящ за стартиране на програма. След 2-5 години програмата трябва да постигне самофинансиране.

4. Анализ на данни в средното образование

Както отбелязахме по-горе, България беше пионер в обучението по информатика в училище и това доведе до отлични резултати в подготовката на специалисти в областта. На сегашния етап би било целесъобразно да се въведе и обучение по анализ на данни. Първоначално това би могло да стане под формата на уроци в различни предмети. Учениците могат да получат задания да открият някакви факти или знания, използвайки Интернет. Подходящо е заданието да е формулирано така, че отговорът да не може да се намери чрез проста заявка към система за търсене. Заданията трябва да се усложняват постепенно, като се включат по-сложни търсения, включително и в бази от данни. Интересни и занимателни задачи могат да се дават от областта на спорта, където онлайн съществуват детайлни бази от данни. В уроците по математика и информатика биха могли да се съчетаят елементи на статистиката с чертане на различни графики, отначало ръчно, а след това и чрез софтуер. На по-късен етап по различите предмети биха могли да се задават проекти, в които да се анализират и обработват голямо количество данни и да се изготвя доклад, в който извлечените знания да се визуализират.

Особено внимание може да се обърне на достоверността на данните и източниците, от които те се извличат, как да се подхожда, в случай, че има липсващи данни, как да се унифицират данните, и т.н. Също така е добре да се обърне внимание на информацията, която учениците оставят за себе си в социалните мрежи, форумите, или чрез използване на сайтове в Интернет-пространството и как тази информация би могла да се използва, включително и от злонамерени лица.

Може да се разгледат и някои етични въпроси и да се дискутира правото на собственост върху данните, кой би трябвало да има достъп до тях, в кои случаи е добре да се забрани събирането и/или анализа на данни и как това може да стане на практика, как да се избегне фалшивата информация разпространявана по интернет и подмяната на информацията и много други.

5. Използване на големи данни за подобряване на обучението

С навлизането на безплатни електронни средства за обучение в образованието, както и на стандартизираните тестове, се натрупаха големи обеми от данни, които могат да се използват за подобряване на обучението. Тези данни съдържат информация за скоростта на четене на материала, времето за отговор, най-често посещаваните страници, отговорите на въпроси от стандартни тестове и др., както на индивидуалния студент/ученик, така и сумарни данни за всеки урок и всеки тест. Тези данни могат да се анализират и от тях да се извлекат знания за успеваемостта на всеки отделен учащ, на група от студенти, обединени по определен признак (възраст, пол, географски район, семеен доход и др.) или за урока/курса. Съответно курсовете и уроците могат да се оптимизират, като се преработят местата, които не се възприемат добре от учениците. С други думи, процесът на обучение може да се оптимизира – нещо, което досега се е извършвало от преподавателя само по интуиция.

Нещо повече, благодарение на анализа на данните в реално време, уроците могат да се индивидуализират и адаптират, като студентите получават повече обяснения, помощ и насочващи въпроси върху частите от урока, които ги затрудняват и обратно, трудността на материала се повишава, когато учащият се справя успешно. Така могат да следват курсовете с индивидуална скорост подходяща за всеки студент или ученик. По този начин се преодолява един от главните недостатъци на груповото обучение, че учителят трябва да преподава със "средна скорост" и да налага уравниловка в класа.

Тук ще посочим и някои проблеми, възникващи при използването на големите данни в образованието. О'Нийл описва в книгата си Weapons of Math destructions: How Big Data Increases Inequality and Threatens Democracy [14] как в столицата на САЩ, Вашингтон, се е приела система за подобряване на обучението, разработена от Университета в Станфорд [15]. В частност, данните от стандартизираните тестове на ученици са се използвали за оценяване на качеството на учителите. Един от критериите при изчисляване на ефективността е бил, че колкото по-високи са оценките на учениците на тези тестове, толкова по-добър е учителят. Данните от тестовете са се анализирали и ефективността на учителя се е пресмятала автоматично, като определен процент от учителите с най-ниски точки са се уволнявали. Целта е била да се осигурят способни учители за всички ученици. Известно е, обаче, че има училища, в които учениците са с по-ниска успеваемост. В резултат на това

оценяване, в тези училища е станало невъзможно да се назначат не само добри учители, а никакви учители – никой не е искал да рискува да бъде уволнен и в досието му да пише, че това е станало поради лошо справяне с работата.

И така, анализът на големи данни трябва да се извършва, прилагайки здрав смисъл и с представа за реалността, а не само с прилагане на математически методи и пускане на данни в системи.

Заключение

Във връзка с нарастването на обемите от данни и тяхната роля при планиране и взимане на решения в различни области на науката и практиката, в тази статия дискутирахме отношението между образованието и анализа на данни. От една страна разгледахме обучението на студенти и ученици, както и професионалното образование на специалисти в други области, застъпвайки мнението, че в близко бъдеще всеки човек ще се нуждае от умения да анализира големи данни. От друга страна, обсъдихме как големите данни могат да се използват за подобряване и индивидуализиране на обучението.

Благодарности

Тази работа е частично спонсорирана от 2017 Cooperative Research Project at Research Center of Biomedical Engineering and Research Institute of Electronics, Shizuoka University. Авторът благодари на Валентин Бримков и Маргарита Бърнева за полезните забележки.

Литература

1. Big Data, for better or worse: 90% of world's data generated over last two years, Science Daily, https://www.sciencedaily.com/releases/2013/05/130522085217.htm (Посетен на 10 юни 2017).
2. Big Data, или до Големите данни и обратно. ComputerWorld, 18 май 2013.
3. Kroenke D.: Using MIS: A problem solving approach. Pearson, 2005.
4. Building a smarter university: Big data, innovation, and ingenuity. New York, NY, USA, October 29-30, 2013, http://www.suny.edu/sunycon/2013/ (Посетен на 10 юни 2017).
5. Brousell, L.: 8 Ways big data and analytics will change sports, CIO, 2014.
6. McLaughlin, M.: Sports teams and leagues in search of the big data boost. BizTech, 2016, http://www.biztechmagazine.com/article/2016/03/sloan-2016-sports-teams-and-leagues-search-big-data-boost (Посетен на 10 юни 2017).
7. Manyika, J., et al.: Big Data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, 2011, http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation (Посетен на 10 юни 2017).
8. 23 Great Schools with Master's Programs in Data Science. http://www.mastersindatascience.org/schools/23-great-schools-with-masters-programs-in-data-science/ (Посетен на 10 юни 2017).

9. Lazowska, E.: Big Data, Enormous Opportunity. Inaugural lecturer in the Critical Conversations University at Buffalo. http://www.cs.fredonia.edu/singh/Big_Data.pdf (Посетен на 10 юни 2017).
10. Rochester Business Journal Staff: St. John Fisher College launching applied data science graduate program. Rochester Business Journal, June 7, 2017.
11. Bachelors in Computer Science & IT. Data Science & Big Data. https://www.bachelorsportal.com/disciplines/282/data-science-big-data.html (Посетен на 10 юни 2017).
12. Barneva, R.P., V.E. Brimkov, J.O. Carbonara, J. Favata, B. Sherman, K. Kanev: Innovative way of offering master's program on data analytics with minimal resources, Japanese Journal of Applied Physics Conference Proceedings, Vol. 4, 2016, ID 011617.
13. NSF: Transdisciplinary Research in Principles of Data SciencePhase I (TRIPODS), https://www.nsf.gov/pubs/2016/nsf16615/nsf16615.htm (Посетен на 10 юни 2017).
14. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016.
15. District of Columbia Public Schools: IMPACT: An Overview. https://dcps.dc.gov/page/impact-overview (Посетен на 10 юни 2017).

BIG DATA AND EDUCATION IN THE INFORMATION SOCIETY

Reneta P. Barneva

State University of New York, Fredonia, USA
barneva@fredonia.edu

Abstract: *The big data issue is relatively recent. Under this term are meant data that cannot be processed with the traditional algorithms and software. There are many predictions that big data analysis will become necessary in every field of computer activity, as nowadays it is important to possess abilities to use computer and communication technology. The article discusses ways to educate specialists to analyze big data, with a focus on life-long learning, graduate, undergraduate, and secondary education.*

Keywords: *big data analysis, life-long learning, higher education, secondary education.*