ISSN: 2473-3857
v3 n4346

A Comparison of Key Concepts in Data Analytics and Data Science

Kirby McMaster
kmcmaster@weber.edu

Brian Rague
brague@weber.edu

Computer Science
Weber State University
Ogden, UT 84408

Stuart L. Wolthuis
stuartlw@byuh.edu
Computer & Information Sciences
Brigham Young University-Hawaii
Laie, HI 96762

Samuel Sambasivam
ssambasivam@apu.edu
Computer Science
Azusa Pacific University
Azusa, CA 91702

Abstract

This research study provides an examination of the relatively new fields of Data Analytics and Data Science. We compare word rates in Data Analytics and Data Science documents to determine which concepts are mentioned most often. The most frequent concept in both fields is *data*. The word rate for *data* is more than twice the next highest word rate, which is for *model*. This contrasts sharply with how often the word *data* appears in most Mathematics books. Overall, we observed substantial agreement on important concepts in Data Analysis and Data Science. Eighteen of the 25 most frequent concepts are shared by both fields. One difference is that the words *problem* and *solution* had Top 25 word rates for Data Science, but not for Data Analytics. A close look at Statistics concepts suggests that Data Analytics is more focused on *exploratory* concerns, such as searching for patterns in data. Data Science retains more of the classical *inferential* activities that use sample data to draw conclusions about populations. Both fields deal with Big Data situations, but Data Scientists must continue to be prepared for traditional small sample applications.

Keywords: data, data analytics, data science, statistics, exploratory, inference.

1. INTRODUCTION

Several decades ago, one of the authors worked as a Statistician for a food manufacturing company. Duties included research design, data collection, data management, and data analysis. The research design component involved relatively small laboratory experiments and sample surveys.

Data collection consisted of cleaning and organizing data onto punched cards or into a single flat file. For data analysis, we used existing software to perform analysis of variance, multiple regression, and cross-tabulation. Occasionally, small Fortran programs were written to perform custom data analyses, such as providing univariate and bivariate descriptive statistics, scatter diagrams, contour plots, and quality control charts.

For each research study, the typical size of the resulting data set was measured in kilobytes. Data management activities were minimal. If several related studies were performed, the generated samples were kept separate, without combining them into a single overall file.

Big Data

In the current era, many organizations now routinely collect massive amounts of data, both for individual studies and during continuing operations. Sizes of data sets for companies such as Walmart and Amazon are measured in gigabytes, terabytes, and beyond. A widely accepted term for these large data collections is Big Data.

With so much data being recorded by companies, organizations, and government entities, the skill-set for a Statistician, who must now deal with Big Data, requires considerable expansion. New approaches have been devised for the management and analysis of extremely large data sets.

Data Analytics and Data Science

Two recent fields that deal with Big Data have been developed and are evolving rapidly--*Data Analytics* (DA) and *Data Science* (DS). Universities are adding programs in these fields at the undergraduate and graduate levels. One Internet listing of 23 such academic programs includes 14 described as Data Analytics, 8 described as Data Science, and 1 program using both names.

A close inspection of these programs through reading academic descriptions and by examining required courses indicates many similarities but some notable differences. Generally, each program provides a mixture of Statistics, Applied Mathematics, Computer Science, and substantive content (e.g. Business, Medicine).

The program name Data Science suggests a careful application of the scientific method, especially research design, sampling, and measurement. The name Data Analytics places more emphasis on ways to describe large data sets. The central goal of both programs is to obtain practical interpretations of data that can assist in making operational and strategic decisions.

Purpose of this Research

In this study, we compare the topics and tools that are presented in Data Analytics and Data Science programs. Our research is relevant to potential students who need to evaluate the knowledge and skills provided in competing academic programs. It is also of value to faculty and academic administrators who are asked to design and teach courses in these programs.

Our research approach involves performing a *content analysis* (Krippendorff, 2012) of selected documents that describe these fields. Words used frequently in each sample of documents allow us to infer which concepts are emphasized in each type of program.

2. METHODOLOGY

This section describes the methodology used to collect word frequency data from samples of Data Analysis and Data Science documents. Special attention is focused on words that are relevant to Big Data issues.

Samples of Documents

Using the Internet, we collected a sample of 14 Data Analysis documents and a second sample of 12 Data Science documents that explain the nature of these fields. We chose documents that are available on the Internet and can be downloaded as PDF files (which are easily converted to text files). The size of the individual documents varied, but the total number of words in each sample was approximately the same.

Convert PDF files to Text Files

Documents in PDF file format are not convenient for performing repeated word searching and counting. Fortunately, Adobe Reader includes an

ISSN: 2473-3857
v3 n4346

option to convert the contents of a PDF file to a text file. We used Adobe Reader to create a text file for each of the 26 documents in our study.

Identify Individual Words in Documents
We observed that the document text files included many character strings that contain digits, punctuation, and other non-alphabetic symbols. They also contained a large number of common English words (e.g. "the", "and") which were not of interest for this study.

To simplify our counting of concept words, we wrote a short Python program that performed the following "data cleansing" tasks.

(1) Our program first changed all letters to *lower-case*.

(2) The program then removed all *non-letter* symbols and replaced them with blanks.

(3) The program converted most *plural* nouns and verbs to *singular* form.

(4) Finally, our program removed approximately 120 common English words that appear on Fry's Lists (1993).

We used our Python program to obtain a filtered set of text files that ultimately consisted of lower-case letters and blanks, the singular form of nouns and verbs (but allowed different verb tenses), and excluded many common English words.

Perform Word Counts
We used a popular program called TextSTAT (Huning, 2007) to obtain word counts for all words in our "cleansed" text files. With TextStat, you first define a "Corpus" which holds a list of text files. We defined one corpus for the 14 DA files and a separate corpus for the 12 DS files.

To perform a word search, a separate TextSTAT screen allows the user to specify search options. Most of the time, we used the option to include all words, with the words and their counts presented in decreasing frequency order. We would then go through the output and record word counts for the most frequent words.

Occasionally, we would enter a short string (e.g. *statistic*) to search for all words that contain the string (e.g. *statistic*, *statistical*, *statistician*).

Word Groups for Concepts
A single data-oriented concept can often be expressed by an author in more than one form. For example, nouns and verbs can be presented in singular or plural form. Verbs can also be written using various tenses. Sometimes, the same concept is described by both a noun and a verb (e.g. "sample", "sampling"). In some cases, *synonyms* representing similar ideas are used to represent a concept (e.g. "algorithm", "method"). Some concepts are written not as a single word but as a word phrase (e.g. "big data").

Our intent was to count how often authors referred to DA and DS concepts. However, TextSTAT was designed to count individual words. For this reason, we defined a *word group* for each concept. A word group is comprised of either a single word or a set of words that represent the same concept. To get a textbook count for a concept, we added the frequencies for each of the words in the word group. This was the most time consuming part of our data analysis.

Convert Word Counts to Word Rates
Because the DA and DS samples of documents contain different numbers of words, the actual word counts for a concept are not comparable across samples. To standardize the counts, we converted each word count for a concept to a *word rate*. The rate we chose was "per 100,000 words". Word rates were calculated for each concept in each set of documents.

3. ANALYSIS OF DATA

Our primary objective in this research is to examine the fields of Data Analytics and Data Science through the prism of selected documents which describe these fields. Which concepts do they share? In what ways are they different? To answer these questions, we compared word rates for concepts in several different ways, as is shown in the following tables.

Most Frequent Words
Table 1 provides listings of the 25 concepts with highest word rates for DA and DS. A separate list of concepts, ordered by decreasing word rate, is shown for each set of documents.

Eighteen concepts occur on both lists, but in different orderings. The most frequent word on both lists is *data*, which might be expected based on the names for the two fields. The second most frequent concept is *model*.

The DA and DS word rates for *data* are more than twice as high as the corresponding word rates for *model*, with word rates declining gradually for the remainder of the concepts. Other concepts with high word rates on both lists include: *value*, *variable*, *function*, *set/element*, and *cluster*.

ISSN: 2473-3857
v3 n4346

No.	DA Word	Rate	DS Word	Rate
1	data	1880	data	2517
2	model	753	model	845
3	point	572	**science**	654
4	value	565	algorithm/ method	523
5	mean/ average	555	set/element	521
6	function	543	value	512
7	variable	472	probability	458
8	**regression**	461	function	453
9	set/element	454	variable	438
10	cluster	439	cluster	422
11	matrix	386	**user/ customer**	416
12	distribution	383	point	413
13	algorithm/ method	380	**solution/ result**	399
14	**analytics**	364	mean/ average	380
15	**estimate**	334	number	378
16	node/vertex	319	**random**	351
17	**big**	318	node/vertex	349
18	number	317	vector	341
19	probability	308	edge/line	333
20	vector	307	matrix	322
21	**linear**	305	**statistic**	312
22	**sample**	295	**graph**	306
23	edge/line	290	**problem**	287
24	analysis	286	analysis	284
25	**tree**	280	distribution	265

Table 1: Top 25 Words - DA vs. DS
Words on a single list are in **bold**.

Seven concepts (shown in **bold**) are unique to each list. *Science* is high on the DS list, whereas *analytics* is relatively high on the DA list. This is not a surprise, and it attests in a minor way to the validity of the data.

Big (data), *regression*, *estimate*, and *sample* are among the Top 25 DA concepts. *Statistic*, *random*, *graph*, and *user/customer* are Top 25 DS concepts. Reasons for these differences will be discussed later in the paper.

Because DA and DS are often described as interdisciplinary fields, we divided many of the concepts into separate tables according to four subject matter categories. Our groups include: Computational Mathematics, Statistics, Discrete Mathematics, and Software Development. These choices reflect our opinion that DA and DS adopt concepts to varying degrees from each of these fields. Some concepts are favored by DA, and others by DS. We added an extra table to present concepts that apply specifically to DA or DS (e.g. *analytics*, *science*).

Computational Mathematics
Some concepts apply to more than one field. For example, *analysis* can refer to an early stage in Software Development, or it can specify a particular Statistics methodology (e.g. analysis of variance).

In an earlier study (McMaster, 2007), we searched 56 Mathematics books for concepts that are common throughout Applied Mathematics. We examined textbooks representing Linear Algebra, Differential Equations, Discrete Mathematics, Statistics, Probability, and Operations Research. Our choice of Math fields was guided by the curriculum in the Applied and Computational Mathematics program at Princeton University.

We found 9 main concepts that are used broadly in Applied Mathematics. These concepts, along with their DA and DS word rates from the current study, are presented in Table 2.

No.	CM Word	DA Rate	DS Rate
1	model	**753**	**845**
2	value	**565**	**512**
3	function	**543**	**453**
4	algorithm/ method	**380**	**523**
5	variable	**472**	**438**
6	solution/result	229	**399**
7	problem	171	**287**
8	system	178	248
9	condition/ constraint	197	118

Table 2: Computational Math Words
Top 25 Word Rates are in **bold**.

The Computational Math concepts are listed in decreasing order by the larger of the DA and DS word rates. The high rates for the Computational Math concepts indicate that DA and DS use these mathematical abstractions frequently to define and represent data.

A *variable* is an abstraction for a set of measurements (*values*) that become data. *Functions* and *models* describe patterns and

ISSN: 2473-3857
v3 n4346

relationships in data. *Algorithms* define calculations that can be used to identify a specific model for a data set.

The higher DS word rates for *problem* and *solution/result* suggest that DS pays more attention to *problem solving*. DA might be more interested in finding data patterns to assist people in making decisions in a variety of situations, rather than in solving one particular problem.

A focus on problem solving in Mathematics books, even in books on Applied Mathematics, is not as common as one might expect. The majority of advanced Math books tend to organize material logically in a more familiar (to mathematicians) *theorem-proof* format. Polya's (1945) "How to Solve It" is one of the earliest and best known Math books having a clear emphasis on problem solving. This book is still in print and is highly regarded today.

By comparison, a more recent book on "How to Prove It" (Velleman, 1994) does not appeal to a wide audience. Perhaps this is because most Mathematics books are already based on the "how to prove it" framework.

Statistics

A list of 18 Statistics concepts, many with high word rates, is given in Table 3. This table does not include Computational Math concepts presented in Table 2, even though some of these concepts apply to Statistics. As in the previous table, the concepts are listed in decreasing order by the larger (DA or DS) word rate.

No.	ST Word	DA Rate	DS Rate
1	data	**1880**	**2517**
2	mean/average	**555**	**380**
3	regression	**461**	179
4	probability	**308**	**458**
5	cluster	**439**	**422**
6	distribution	**383**	**265**
7	random	168	**351**
8	estimate	**334**	126
9	statistic	233	**312**
10	sample	**295**	157
11	analysis	**286**	**284**
12	test	264	203
13	predict	254	223
14	error	251	154
15	plot	228	107
16	variance	221	91
17	component	198	133
18	density	198	68

Table 3: Statistics Words
Top 25 Word Rates are in **bold**.

In Table 3, 9 DA concepts and 8 DS concepts have Top 25 word rates. The word having the highest rate on both lists is *data*. We consider *data* primarily as a Statistics concept, even though it is used frequently in computing and in applied fields (e.g. science, business, government). From our previous research, we found that the word *data* appears infrequently in most Math books, including Applied Math books.

The 6 concepts having Top 25 word rates for both DA and DS are: *data*, *mean/average*, *probability*, *cluster*, *distribution*, and *analysis*. *Regression*, *estimate*, and *sample* are Top 25 concepts for DA. *Random* and *statistic* are Top 25 concepts for DS.

The field of statistics can be divided into *exploratory* and *inferential* activities. Exploratory methods search for patterns in the sample data, with less regard to the source of the data and the manner of sampling. Inferential statistics uses sample data to evaluate claims (hypotheses) about the population from which the sample was drawn.

Inferential statistics requires *probability* models based on how the data is collected. Usually, the basis is random sampling in surveys or randomization in experiments. Observe in Table 3 that the word rates for *probability* and *random* are higher for DS than for DA, since DS focuses more on inference.

On the other hand, the word rates for *regression*, *mean/average*, *estimate*, and *sample* are higher for DA. These concepts describe characteristics of the sample data.

Discrete Mathematics

Discrete Mathematics is a topic taught to Mathematics students and Computer Science students. In the CS curriculum, the course is often called Discrete Structures. Table 4 lists 12 Discrete Math concepts, along with their DA and DS word rates. Again, the concepts are listed in decreasing order by the larger (DA or DS) rate.

Discrete Math models are consistent with the discrete nature of data in a computer. Continuous Math models such as differential equations require floating point numbers and careful computation techniques employing numerical methods.

The word rates for DA and DS are surprisingly similar. Nine of the concepts are Top 25 DA words. Eight of the concepts are Top 25 DS

ISSN: 2473-3857
v3 n4346

words. The first 7 concepts on the list are Top 25 words for both fields.

No.	DM Word	DA Rate	DS Rate
1	point	**572**	**413**
2	set/element	**454**	**521**
3	edge/line	**290**	**333**
4	matrix	**386**	**322**
5	number	**317**	**378**
6	node/vertex	**319**	**349**
7	vector	**307**	**341**
8	graph	243	**306**
9	linear	**305**	172
10	tree	**280**	150
11	dimension	191	237
12	distance	133	215

Table 4: Discrete Math Words
Top 25 Word Rates are in **bold**.

Most of the word rate differences are relatively small. The largest differences are for *point*, *linear*, and *tree*, with DA having the higher rates.

Most of the Discrete Math concepts on the list define data structures (*matrix*, *vector*, *point*), finite models for data (*graph*, *tree*), and special features of the models (*node/vertex*, *edge/line*, *dimension*, *distance*). These models and data structures tend to be applied to sample data patterns, rather than to draw inferences to populations.

Software Development

Table 5 lists 10 concepts that relate to the creation of software. We call this process Software Development.

No.	SE Word	DA Rate	DS Rate
1	user/customer	216	**416**
2	class/object	275	182
3	case	215	165
4	input/output	198	204
5	code/software	194	183
6	table	138	161
7	type	126	158
8	attribute	143	53
9	database	134	121
10	file	91	109

Table 5: Software Development Words
Top 25 Word Rates are in **bold**.

Software Delopment allows us to see data and algorithms from a computer's point of view, which can improve our understanding of DA and DS. The value of computers in DA and DS is not limited to the ability of computers to transform data rapidly. Practitioners also benefit when they are able to translate data structures and algorithms into a language that is understandable to the computer (Knuth, 2008).

In Tables 2 thru 4, over half of the concepts have Top 25 word rates for DA and DS. In Table 5, the only Software Development concept that has a Top 25 word rate is *user/customer* for DS. The user/customer usually provides the problem for the Data Analyst or Data Scientist to solve, plus a request for software.

The concept *class/object* barely misses having a Top 25 word rate for DA). This is a foundation concept for developing software components using object-oriented programming (OOP).

The remaining concepts in Table 5 have word rates above 100 for DA and/or DS. These topics are discussed in our sample documents, but at a lower rate than most concepts in earlier tables. This indicates that these Software Development concepts are relevant to DA and DS, but do not receive the same level of coverage by the authors.

Most DA and DS academic programs require at least one programming course. However, the amount of programming that is required of students in a "non-programming" course can vary greatly.

We note that the concepts *database*, *table*, and *attribute* have fairly low word rates in Table 5. Their low word rates do not diminish the importance of database principles and software for managing Big Data.

Data Science and Data Analytics
In our last table, we highlight concepts that are important to DA and DS, but do not fit well into any of our previous categories. We think of these words as DA-specific or DS-specific. Nine of these concepts are presented in Table 6.

We have already mentioned the Top 25 word rates for *science* (DS) and *analytics* (DA). The word *big* (usually stated as *big data*) has a Top 25 word rate only for DA. The remaining 6 concepts describe models and methods for DA and DS. Four of the concepts have higher DS rates (*learning*, *visualization*, *training*, and *machine*). Two of the concepts have higher DA rates (*hadoop* and *mining*).

No.	SE Word	DA Rate	DS Rate
1	science	58	**654**
2	analytics	**364**	141
3	big (data)	**318**	90
4	learning	97	237
5	visualization	29	190
6	training	145	187
7	hadoop	143	31
8	machine	59	134
9	mining	134	21

Table 6: Data Science and Data Analytics
Top 25 Word Rates are in **bold**.

Machine and learning are usually expressed as the single concept *machine learning*. Visualization and mining are usually combined with the word data, as in *data visualization* and *data mining*. Hadoop is widely-used open source software for the management and parallel processing of big data.

The bottom 6 concepts in Table 6 apply in varying degrees to both DA and DS. The differences in their low word rates could partially be due to our small samples of documents.

4. CONCLUSIONS

This research study provides an examination of the relatively new fields of Data Analytics and Data Science. We compared word rates for concepts mentioned most often in samples of DA and DS documents. Our analysis of word rates leads us to the following conclusions.

First, there is substantial agreement on the most important concepts in DA and DS. The 25 most frequent concepts in each field share 18 of these concepts.

The most frequent concept in both fields is *data*. The word rate for data is more than twice the second highest rate, which is for *model*. Given the "D" in the names of the DA and DS fields, the frequent mention of data is not surprising. However, in earlier research (McMaster, 2007) we found that books on Mathematics topics often favor a logical framework (*theorem*, *proof*) over an empirical approach (*data*). You can think of DA and DS as leading a renaissance for data.

Second, when the concepts in our documents were grouped into the categories Computational Math, Statistics, and Discrete Math, the concepts with highest rates tended to be the same for DA and DS.

In the Computational Math category, *variable*, *value*, *model*, *function*, and *algorithm/method* had high rates for both DA and DS, but *problem* and *solution/result* had noticeably lower rates for DA. This suggests that DS places more emphasis on *problem solving*.

We included a category for Software Development concepts, since DA and DS can be viewed as a blend of Statistics, Mathematics, and Computer Science. Almost all of the Software Development concepts had low word rates in the DA and DS documents. The explanation for low word rates might be partially due to the choice of documents in our samples. On the other hand, data analysts and data scientists do not write most of the software they use, so less emphasis on programming is understandable.

However, three of the Software Development concepts with low word rates--*database*, *table*, and *attribute*--are only indirectly involved in writing code. DA and DS without databases would be ineffective, so the lack of discussion about databases is disappointing.

Third, a closer look at Statistics concepts with differing DA and DS word rates suggests that DA places more focus on *exploratory* concerns, such as searching for patterns in data. DS retains more of the classical *inferential* activities that use sample data to draw conclusions about populations.

One reason that DS retains more focus on inferential statistics is due to sample size considerations. Both DA and DS deal with Big Data situations. DA has a higher word rate for *big* (data), but data scientists must also be prepared for traditional small sample problems.

Inferential statistics requires probability models based on the data collection methodology. The probability distribution for a statistic (*sampling distribution*) varies with the sample size. In many cases, the variance of the statistic is inversely proportional to the sample size. An extremely large sample size will yield a very small variance for the statistic. When the sample size is large, a "significant" (but small) difference in the sample may be unimportant for practical reasons. Thus, in Big Data cases, the sample can be considered to be the entire population, making inference irrelevant.

Future Research
Future research is planned for the following Big Data studies:

ISSN: 2473-3857
v3 n4346

1. Repeat this study with larger and more representative samples of documents. The literature on Data Analytics and Data Science is growing rapidly. In addition, the fields themselves are evolving in goals, methods, and applications.

2. Perform a comparison of program outlines and course descriptions for the ever-increasing number of graduate and undergraduate degrees offered in Data Analytics and Data Science. We would record which courses form the core of the programs and which topics are available as electives.

3. Perform an analysis of several Big Data projects to examine what types of applications are represented, what methodologies they employ, and how they measure "success".

5. REFERENCES

Fry, E.G., Kress, J. E., & Fountoukidis, D.L. (1993), *The Reading Teacher's Book of Lists* (3rd ed). Center for Applied Research in Education.

Huning, M. *TextSTAT 2.7 User's Guide. TextSTAT*, created by Gena Bennett, 2007.

Knuth, D. (2008), "Donald Knuth: A Life's Work Interrupted." Communications of the ACM, Vol. 51, No. 8.

Krippendorff, K. H. *Content Analysis: An Introduction to Its Methodology*, 3rd Ed. SAGE Publications, 2012.

McMaster, K., Hadfield, S., Wolthuis, S., & Sambasivam, S. (2007), "Two Gestalts for Mathematics: Logical vs. Computational." Proceedings of ISECON 2007, Vol. 24.

Polya, G. (1945). *How To Solve It*. Princeton University Press.

Velleman, D. (1994). *How to Prove It: A Structured Approach*. Cambridge University Press.

Data Analytics documents (partial list)

Hadoop: Big Data Analysis Framework. Tutorials Point, 2014.

Ledolter, J., *Data Mining and Business Analytics With R*. John Wiley & Sons, 2013.

Shalizi, C. R., *Advanced Data Analysis from an Elementary Point of View*. Spring 2013.

Wesler, M., *Big Data Analytics For Dummies*. John Wiley & Sons, 2013.

Zaki, M., & Wagner M., *Data Mining and Analysis*. Cambridge University Press, 2014.

Data Science documents (partial list)

Grus, J., Data Science from Scratch. O'Reilly Media, 2015.

Hopcroft, J., & Kannan, R. *Foundations of Data Science*, Draft, 2014.

Herman, M., Rivera, S., Mills, S., Sullivan, J., Guerra, P., Cosmas, A., Farris, D., Kohlwey, E., Yacci, P., Keller, B., Kherlopian, A., & Kim, M. *The Field Guide to Data Science*. Booz, Allen, Hamilton, 2013.

Pierson, L. *Data Science For Dummies*. John Wiley & Sons, 2015.

Zumel, N., & John M., *Practical Data Science with R*. Manning Publications, 2014.