Paper 2781-2018

Demystifying Buzzwords:

Using Data Science and Machine Learning on Unsupervised Big Data

Ben Murphy, Zencos

ABSTRACT

Each month it seems there is a new technology introduced that can transform your organization in unimaginable ways. Each technology arrives with its own set of industry buzzwords that make it difficult to understand how your organization would benefit. We will go through the meaning of the buzzwords associated with machine learning, data science, artificial intelligence, and advanced analytics. We can immediately apply this knowledge to a discussion about how a team of data scientists can follow a process guided by answering business questions using the best tools for each task. We review the relative strengths and weaknesses of the various tools, techniques, and technologies associated with these buzzwords. You will walk away a winner by knowing how to put each to its best use.

INTRODUCTION

Let's be honest for a second, just you and me. We're going to talk about buzzwords around data science and many related topics, but almost every one of them boils down to some combination of math/statistics, computer science, and data. My goal is to help you navigate the problem-solving process, so the purpose of understanding what these terms mean is to help you feel more knowledgeable as you tackle the problems you face I'll try to keep you from getting sick of the word "buzzword" and bee-related metaphors, but if you learn something, that's what matters.

I'm guessing you came here because you've got a sense that you're going to get more useful results by collecting and analyzing data about what's going on with your organization. Maybe you've got the data you need already, maybe you don't, or maybe you're not sure. Perhaps you've got a sense for how to start analyzing your data, but you're uncertain what the next step should be. We're going to cover a variety of topics that will help you understand how to navigate the entire process starting from no data and ending with better results for you and your team.

Knowing the difference between, say, unstructured and structured data is only useful if you've got both types available, have the tools and techniques to analyze either or both, and possess an understanding of why you might choose each one depending on the situation. By the end of our little chat, I want you to feel confident that you can get the data you need, learn from the patterns in the data, and develop a game plan so that you're ready to act to make things better in your organization. We'll cover everything you need to know, and help you go find anything else you might need. Sound good? Let's get started.

WHAT'S ALL THE BUZZ ABOUT?

Buzzwords are nothing new, but sometimes it's hard to tell if people know what's going on or if they're just trying to sound smarter than they are. It's important to know the meaning of some of these terms and phrases, but that's not how problems get solved. Problems get solved when we can understand what decision makers can control, when we understand as much as possible about the different options available for each decision, and we can evaluate the potential outcomes of those options to pick the best path forward. We'll talk through the meaning of buzzwords as part of navigating the problem-solving process so that you feel educated on the definitions of the buzzwords and prepared to solve the problems facing your team.

WHICH WORDS BUZZ THE LOUDEST? THE BUZZIEST WORDS

The biggest buzzwords in the industry often apply to the large categories or descriptors of groups of things.

Data Science

The term “data science” applies to a wide variety of tools and techniques that are all about learning from data and solving problems with data. Like other scientific disciplines, data science is focused on the ways that people can understand data and use data to their benefit. Almost everything we’re going to discuss in this paper is part of data science as a discipline.

Big Data

The term “big data” is usually used to refer to data that is too large in size in some way (number of variables, number of observations, or complexity of data types being the most common) for analysis to be straightforward with traditional or more simplistic methods. In many ways, the term big data has been over used and lost some of its meaning. Some also feel that the enthusiasm around big data puts too much emphasis on the quantity of data without clarifying any qualities of the data, which is often more important. Spending a great deal of time or energy getting larger quantities of data is rarely a worthwhile pursuit, but going after high quality data is almost always useful. Additionally, instead of talking about big data recently, there’s more discussion around the way data is structured (more on that soon).

Artificial Intelligence

The term “artificial intelligence” is basically what it sounds like – electronics that resemble human intelligence in some way. The term artificial intelligence was coined decades ago when people first envisioned machines that could be designed with ways of operating like the human brain. These machines were to be engineered by humans and wouldn’t have any naturally occurring intelligence, but their human creators could design them in a way that they would be able to function in some capacity like a human brain. There are different varieties of artificial intelligence that are described based on whether the machine is proactive or reactive, and how broad or narrow the problems are that the machine can figure out.

Machine Learning

The term “machine learning” is meant to refer to an ever-growing list of algorithms and models that are designed to leverage computing power to improve the process of learning from data. Machine learning is a subset of artificial intelligence in that the term does mean, in some sense, exactly what it sounds like – teaching machines to be able to learn. There is a wide variety of types of tasks that machine learning methods might accomplish, and we’ll explore some of them in more detail later. Machine learning algorithms are generally geared toward prediction, which is part of why they are so popular.

Predictive Analytics

The term **“**predictive analytics” has a self-evident definition – a type of analysis that’s meant to predict something. With such a generic meaning, this phrase is used often and in most cases, it could be more beneficial to use a term that more specifically clarifies exactly what kind of methodology or algorithm is being used.

Data Mining

Methods of “data mining” are like machine learning because they are designed to discover patterns in data, usually with the use of a computer. Usually the implication with the term data mining is that all the discovery is driven by a person, which is one slight contrast between machine learning and data mining, as many of the algorithms or methods are similar between the two.

CONTEXT IS EVERYTHING

For these more general terms, having a precise definition is not as important as a general understanding for each term and how they relate to each other and related terms. Indeed, searching the Internet for definitions and Venn diagrams could turn up dozens of possibilities for each buzzword. Instead of focusing too much on trying to gather some consensus about specific definitions, let's focus on how understanding these terms helps us solve problems in real-world applications.

To be honest, it would be awesome to have a graphic that helped capture the relationships for these buzzwords, and if you search online you're likely to find many of the same ones that I did – check out the recommended reading section. Unfortunately, even when some of these relationships are straightforward, like machine learning being a subset of artificial intelligence, and deep learning being a subset of machine learning, the graphics don't really help convey anything beyond describing it. Showing you 3 concentric circles wouldn't add to your understanding of these 3 terms, and the other terms aren't related in a way that's easy to visualize. When you find some of those visualizations online, they often seem to highlight how messy, convoluted, and distracting these buzzwords can be.

In our context here, data science is the most general term and artificial intelligence, big data, machine learning, predictive analytics, and data mining are all parts of data science. Big data is a sweeping way to describe information that probably requires a modern computer to analyze. Predictive analytics includes a large variety of algorithms and methods, some of which are examples of artificial intelligence, machine learning, or data mining. Machine learning is a subset of artificial intelligence focused on simulating rational thought in computers so that they can draw conclusions based on information. Data mining is a set of algorithms that include many things that are also examples of machine learning.

WHERE'S THE HONEY? PEOPLE, PROCESS, TECHNOLOGY!

If your hive is buzzing about some of these terms, then you can help them stay focused on thinking about how to get to the honey, or the actionable solutions, which are the biggest benefit of data science.

People: Do you have to find a rainbow unicorn?

Ideally, data scientists have some blend of experience and knowledge in things related to computer science, math and statistics, and the problems at hand -- subject matter expertise. Applied data science is about combining these skills and knowledge areas to utilize computers to analyze information and come up with interesting new insights. The best data scientists are also capable of handling all the important communication necessary to understand the problems at hand and then explain the solutions and findings once analysis is ready.

Some data scientists are more general and have a broad understanding of many of the ways to find patterns in data, and others are more specialized in specific algorithms or methodologies. The various data science skills – computer science, math and statistics, subject matter expertise, and communication – are all important for people based on their role. The best teams of data scientists often include a few generalists, especially in leadership roles, and as many specialists as required.

Process: Can you keep the focus on business questions?

Once the right people are on board the data science team, it's useful to have a robust and consistent process that can be followed to ensure that the work efforts of your data science team are useful to you, to your internal stakeholders, and to your customers. A discussion about buzzwords isn't the setting to describe a full process for problem solving for a data science team, but the process can be critical to success of the data science efforts, just like understanding the buzzwords and what they mean. Data science endeavors should be geared around answering questions pertinent to the business. Understanding the decisions that an organization can make and the options for each decision will guide data science work towards being actionable and providing value.

Technology: What tools are available?

If the right people are going through a well-guided process, the last step is to have the right tools and technology available. Although this is for SAS® Global Forum, I'm not here to sell you SAS software as

the technology you would use. The beauty of the Power of SAS (especially with the recent efforts in Viya® and the related tools in Visual Analytics®, Visual Statistics®, and all the classic SAS software like Enterprise Miner®) is that it provides a wealth of options for data scientists to find the tools they most prefer. If you or your team wants to work in open source, SAS supports full integration to R and Python through Viya. If you or your team are comfortable in classic data mining software tools, Enterprise Miner is a best in class option. If you haven't gotten used to any specific tool yet, SAS has built up a robust suite of options available through Visual Analytics and Visual Statistics called Visual Data Mining and Machine Learning that are easy to pick up and get results quickly.

NEXT STEPS: WHERE DO WE GO FROM HERE?

We will walk through some of the specific terms that can help you navigate applied data science, and if you want more information about them or any others, there's a wealth of information online.

WHAT ARE THE WORKER BEES TALKING ABOUT?

Although we can't possibly go through and define and discuss every term under the data science umbrella, there are a few more specific terms that apply to problem solving in data science that are important.

Supervised / Unsupervised Algorithms

"Supervised" / "Unsupervised Algorithms" are probably the most misunderstood buzzwords we will discuss—and I poked fun at the term in the title by using it incorrectly. This is a case where the language definition of the word does not line up well with the applied meaning in this context. Categorization of supervised or unsupervised has nothing to do with whether a human is assisting or providing oversight for the computer or the algorithm itself. Instead, algorithms are considered supervised when the goal of the algorithm is a specific target or prediction and algorithms are considered unsupervised when the algorithm does not have a specifically labeled target to predict. This differentiation means that accuracy measures are appropriate for supervised algorithms but not for unsupervised ones. Some examples of supervised algorithms include regression or decision trees and examples of unsupervised algorithms include clustering and association patterns.

Structured / Unstructured Data

"Structured data" usually has a defined data model and the way that different data elements are related is well defined, which usually also involves metadata about data types and such. "Unstructured data" has less specific relationships between data elements and may not have any metadata or a data model at all. Unstructured data is also typically more text heavy. Examples of unstructured data include books, journals, audio, video, and pictures. Both structured and unstructured data can include a variety of types of information, such as dollar amounts, dates, and text. The distinction, then, is about whether there's a well understood relationship between the various amounts, dates, and text.

Deep Learning / Neural Networks

"Deep learning" is a term to describe a subset of machine learning that's designed to have the algorithm perform feature extraction tasks without human assistance. In many machine learning algorithms, the person programming the algorithm is required to identify the things that might help make decisions or improve performance, but in deep learning, the algorithms can perform feature extraction and learn to focus on the right features, which reduces the need for human assistance and makes these algorithms extremely powerful. "Neural networks" are an example of how deep learning can be implemented, and they're designed based on an understanding of how the human brain works to build algorithms that can replicate the learning and processing power of human brains.

Cloud Computing / Distributed Computing

As technology has progressed and as demand for analytics has increased, one of the biggest shifts has been in how the hardware –the physical computer – is setup and configured. The industry has migrated from having servers owned by an organization to having companies that provide server grade computing

available to their customers. Companies that want to perform analytics may decide that they want to take advantage of having the physical hardware hosted by another corporation, which is generally referred to as “cloud computing.” Relatedly, as cloud computing became more extensive and algorithms were developed that required tremendous computing resources, hardware that is organized into a cluster to create distributed systems that can manage tasks by treating a network of machines as one large computing unit to better accomplish significant work efforts. Both concepts of cloud computing and distributed computing can be setup internally for an organization or through a third party.

Internet of Things

“Internet of Things” (or IoT) is a term to describe the idea that common everyday objects are connected to networks and generating data in ways that did not exist until recently. Things like vehicles, home appliances, and other physical devices are being built with sensors, actuators and connectivity so that devices can share all the data collected by those sensors and actuators. As a relatively new wealth of information and data becomes available from all these connected devices, the term Internet of Things helps describe all the interconnected devices and the data they generate.

HOW CAN I WALK AWAY A WINNER? PUTTING IT ALL TOGETHER

We’ve covered several buzzwords and talked through some of how organizations might want to put together teams of people, build processes to solve problems, and leverage technology to do so as efficiently as possible. Let’s review the relative strengths and weaknesses of the various tools, techniques, and technologies associated with data science.

Tools:

Tools available for data science include a wide variety of commercial software like SAS, SPSS®, and Stata®, as well as open source software like R and Python. Each of these options will have relative strengths and weaknesses. The key factors to consider are the time and money required to implement solutions, but the decision may also be influenced by other factors, such as the ability to integrate with the existing system and the availability of resources and people who have experience with each option.

Commercial and proprietary software, like SAS, tend to have more robust user interface design, wizard style approaches to guide users through designing and building algorithms, and offer customer service and support. The commercial software also tends to require money when acquiring the software so that these companies can maintain the business of developing their software and providing useful tools and support to the users.

Open source software, like R and Python, tends to have very flexible programming options, they usually support the latest cutting-edge methodology as early as possible, and they are almost always free of monetary costs to install and run. Open source software is likely to require more programming time to design and develop models since they don’t often have highly flexible wizard style guided interfaces, and the support comes from a community of users – searching the Internet for answers, basically – instead of a commercial customer service team.

Choosing the right tools amidst options from commercial software and open source software depends on whether you’d rather spend time or money, and whether you are comfortable with not having commercial customer service support or guided interfaces to help your data scientists. As part of their development of Viya, SAS has done an excellent job in giving users the option to use both SAS software and open source software together, which is a terrific way to balance the best of both. In each situation, weigh whether you’d rather buy commercial software to gain the ease of use and save time with helpful interfaces and powerful functionality or get open source software for lower monetary outlay if you are willing to put in additional time and energy to get the same result.

Techniques:

There are a wide variety of techniques available for identifying patterns in data and making predictions. The nature of the data and the types of patterns and predictions that will help decision makers can help identify candidates for which techniques may suit the situation. The biggest tradeoff to consider when

narrowing the options is between how easy it is to explain what the technique is doing and the predictive power. Essentially, more complicated and intricate techniques will be harder to explain but will often offer better predictive power.

In the simplest cases, things like linear regression are the easiest to explain and understand. It's easy to give transparency around exactly how it works, exactly how the pieces fit together and exactly how the predictions follow from building the regression. At the other end of the spectrum for this tradeoff are deep learning methods, like neural networks, where they are designed to minimize human interaction in design and development and they are amongst the most powerful techniques currently available but they're basically black boxes when it comes to understanding how they work. In between these two end points is where many techniques are designed to strike a balance between transparency and accuracy.

If we are going by the guidelines that we discussed previously where we want to deliver something to decision makers to help enable them to make better decisions, then being able to explain how things work can often be critically important. Many times, people building predictive models underestimate the value in being able to understand how predictions are made. In each situation, determining the best model to use probably involves an understanding of the intended audience and their need to understand the way the model works, and the extent to which each option fits the data and creates accurate predictions. In situations where there's multiple techniques that might be comparable in those factors, my advice is to value transparency and in general, the simplicity of a technique, above the predictive power. Starting with the simplest method that leads to an improved understanding and a better ability to make decisions is often the best course of action.

Technologies:

We've covered a few types of technology when talking about cloud and grid computing. Without being able to delve into immense technical detail, it's important to know that often the tradeoff between time and money for hardware technology is similar to the tradeoff for software that we discussed in the tools section. Typically, the fastest and highest performing hardware costs the most money. There are other concerns, like security and stability, that should also be weighed, but those don't necessarily come at the direct tradeoff of other features or functionality.

As technology evolves and innovative ways of performing intense number crunching are developed, figuring out which technology to use gets more and more difficult. One example is the development of using graphics processing units (GPUs) to leverage their massively parallel architecture to perform parallel calculations much faster than traditional central processing units (CPUs) may be able to do so. The reason that GPU processing is so intriguing, though, is the hope that it will be an exception to the typical tradeoff in hardware technology between time and money if purchasing the necessary GPU hardware is cheaper than similar CPU hardware in addition to the possibility that in these applications, GPUs can also be faster.

As the internet of things starts to trend toward the network of everything, and all devices are collecting and sharing data, it will only get more difficult to figure out the optimal technology to use for your data science team. The flexibility of utilizing third parties for hardware and software installation and configuration is hard to beat, especially as the cost for using third party services has gotten much more comparable to the overall cost of acquisition internally when considering time and money. This has also given rise to multiple service offerings for providing "(blank) as a service" for things like "Data Science as a service" (DSaaS) and "Machine Learning as a service" (MLaaS). For many organizations, the complexity of installation and configuration coupled with the wide array of choices means that using a third party will yield a more effective system that's more quickly deployed, ultimately proving more beneficial than going the internal route.

CONCLUSION

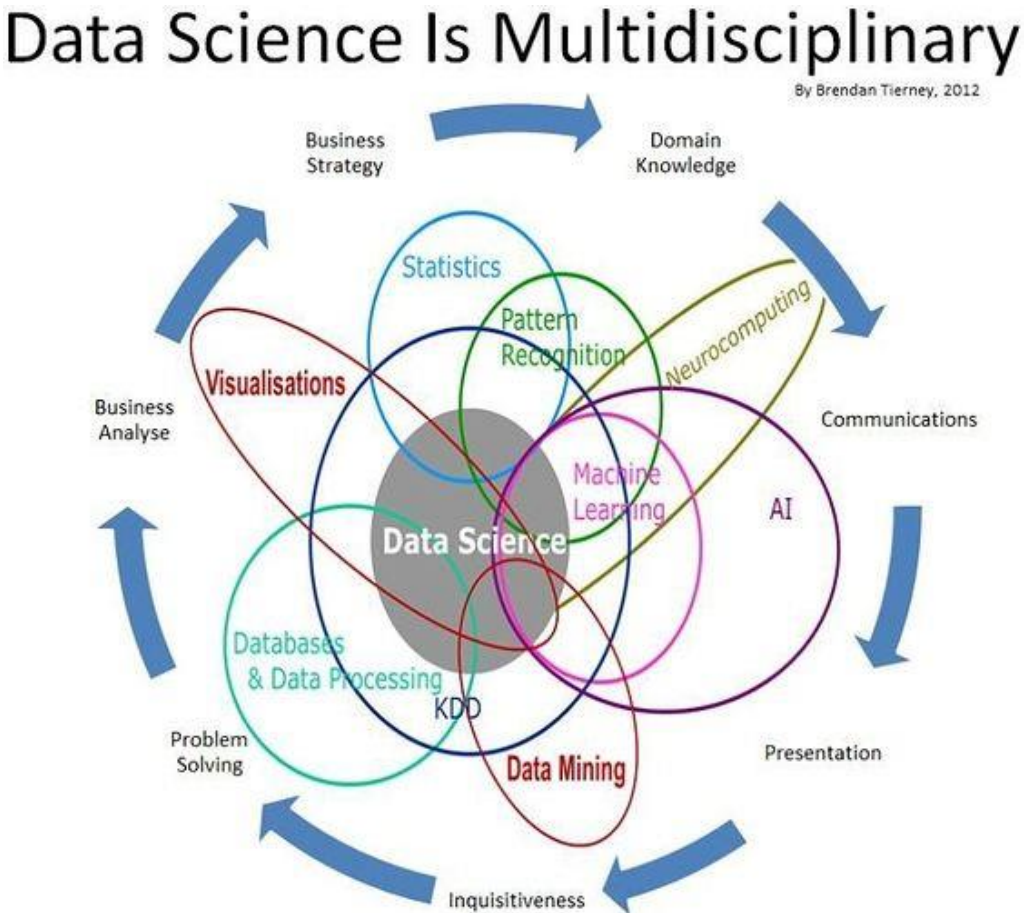
Thanks for taking the time to chat with me today. I hope our conversation helped you understand all the buzzwords that are flying around conversations in your organization. I also hope that beyond equipping you with the knowledge to participate in conversations and use those buzzwords correctly, you are also excited to be as busy as a bee, tackling problems in your organization and delivering actionable solutions. At the end of the day, being able to make better decisions is the best way to add value with data science.

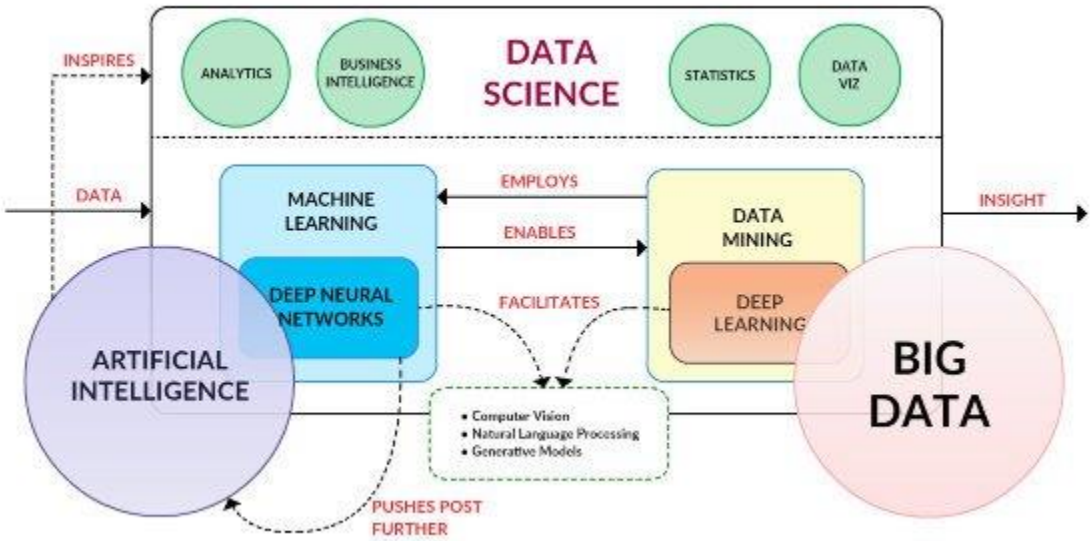
REFERENCES / RECOMMENDED READING

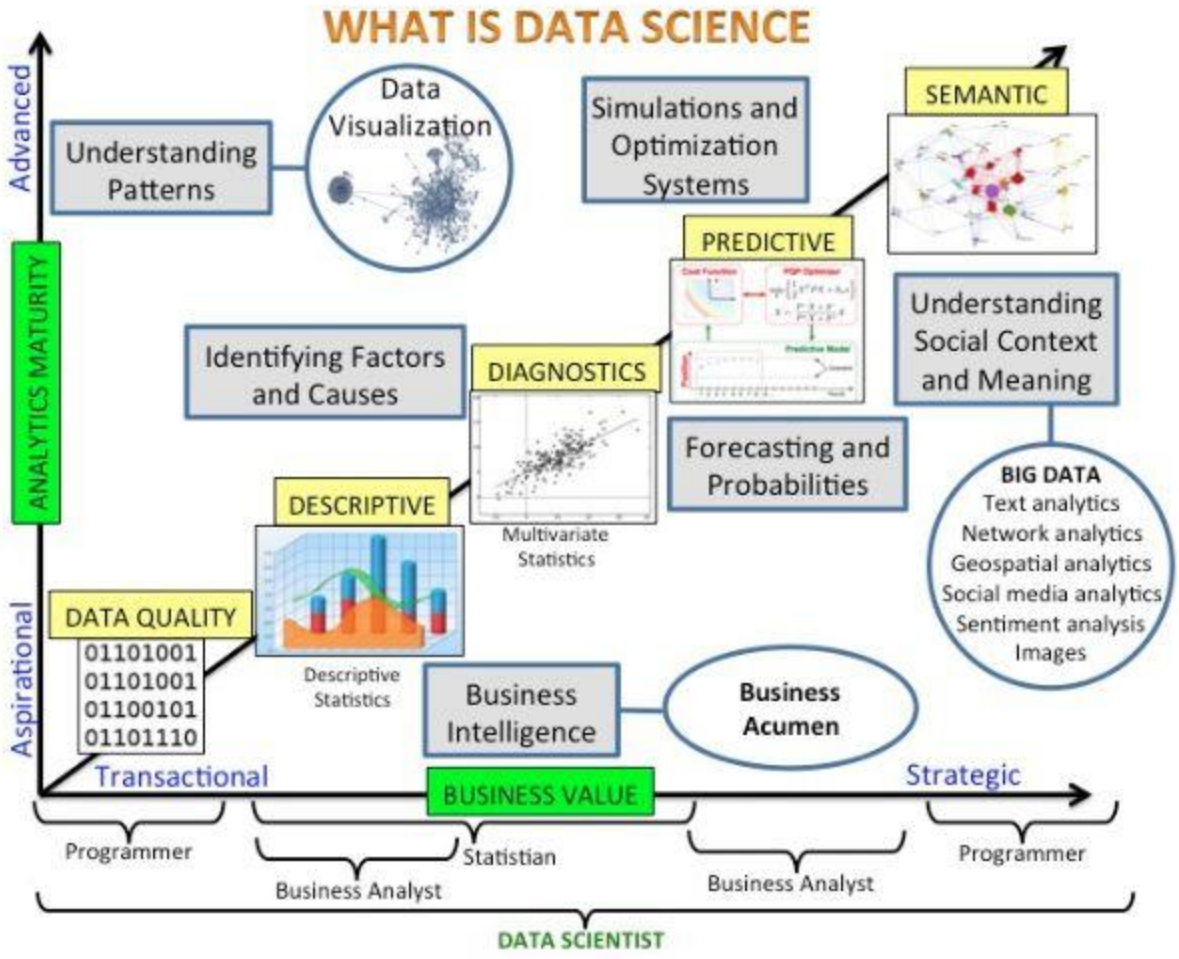
The following online resources were helpful while writing this paper, either as examples of buzzwords that are being discussed, the types of discussions happening about buzzwords, or ways to think about data science buzzwords and how they're related. In general, the meaning of these terms is based on common usage or specific definitions, so these aren't references in a classic sense, but they are likely to be interesting to you!

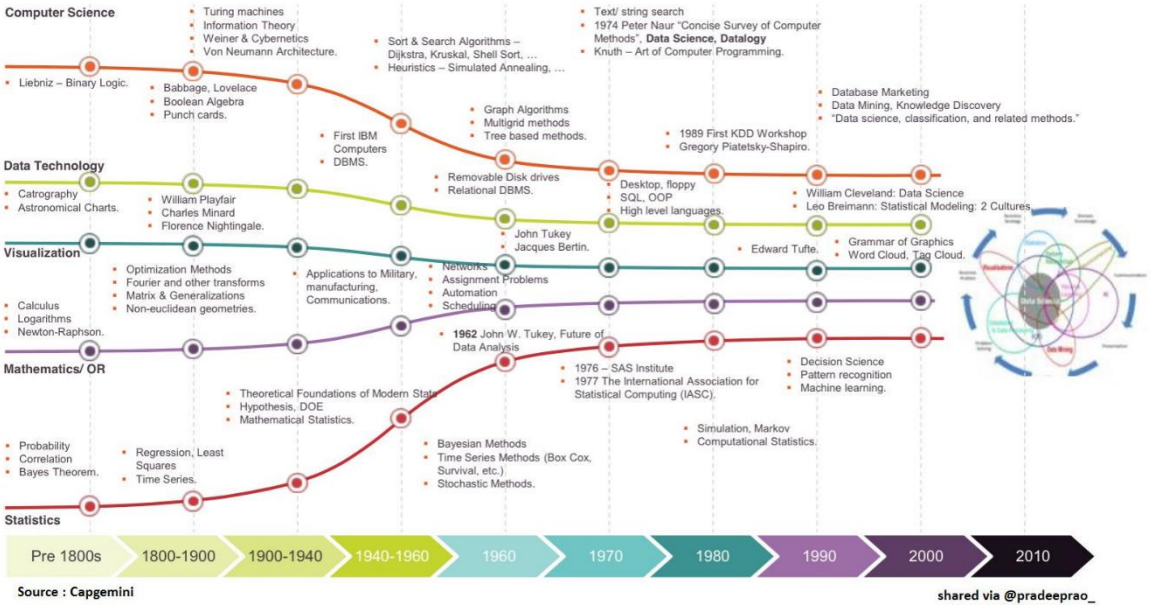
- https://indico.io/blog/data-science-buzzwords-demystified/
- https://www.forbes.com/sites/valleyvoices/2017/03/16/the-3-tech-buzzwords-every-ceo-should-know/
- https://www.datasciencecentral.com/profiles/blogs/word-clouds-of-big-data-data-science-and-other-buzz-words
- https://www.linkedin.com/pulse/data-science-buzzwords-part-1-chett-rubenstein/
- https://revenueanalytics.com/machine-learning-data-science-and-operations-research-whats-the-difference/
- https://www.reddit.com/r/statistics/comments/4w32o5/much_of_data_science_and_associated_buzzwords_are/
- https://www.datapine.com/blog/business-intelligence-buzzwords-2016/
- https://www.kdnuggets.com/2015/02/data-science-confusing-jargon-abused.html
- https://www.dialogtech.com/blog/marketing-analytics/ai-data-science-buzzwords
- https://www.forbes.com/sites/bernardmarr/2016/12/06/what-is-the-difference-between-artificial-intelligence-and-machine-learning/#3ca99c512742
- https://medium.com/iotforall/the-difference-between-artificial-intelligence-machine-learning-and-deep-learning-3aa67bff5991
- https://www.quora.com/What-are-the-main-differences-between-artificial-intelligence-and-machine-learning-Is-machine-learning-a-part-of-artificial-intelligence
- https://stats.stackexchange.com/questions/5026/what-is-the-difference-between-data-mining-statistics-machine-learning-and-ai

The following images are examples of how messy some of the visualizations of data science can be. The source is indicated when available, but for some, the original source was unclear.






WHAT IS DATA SCIENCE
Advanced
ANALYTICS MATURITY
Aspirational
Understanding Patterns
Data Visualization
Simulations and Optimization Systems
SEMANTIC
PREDICTIVE
Understanding Social Context and Meaning
Identifying Factors and Causes
DIAGNOSTICS
Forecasting and Probabilities
BIG DATA
Text analytics
Network analytics
Geospatial analytics
Social media analytics
Sentiment analysis
Images
DESCRIPTIVE
Multivariate Statistics
DATA QUALITY
01101001
01101001
01100101
01101110
Descriptive Statistics
Business Intelligence
Business Acumen
Transactional
BUSINESS VALUE
Strategic
Programmer
Statistian
Business Analyst
Business Analyst
Programmer
DATA SCIENTIST


Computer Science
Liebniz – Binary Logic.
Turing machines
Information Theory
Weiner & Cybernetics
Von Neumann Architecture.
Babbage, Lovelace
Boolean Algebra
Punch cards.
Sort & Search Algorithms – Dijkstra, Kruskal, Shell Sort, …
Heuristics – Simulated Annealing, …
Graph Algorithms
Multigrid methods
Tree based methods.
First IBM Computers
DBMS.
Text/ string search
1974 Peter Naur "Concise Survey of Computer Methods", Data Science, Datalogy
Knuth – Art of Computer Programming.
1989 First KDD Workshop
Gregory Piatetsky-Shapiro.
Database Marketing
Data Mining, Knowledge Discovery
"Data science, classification, and related methods."
Data Technology
Catography
Astronomical Charts.
William Playfair
Charles Minard
Florence Nightingale.
Removable Disk drives
Relational DBMS.
Desktop, floppy
SQL, OOP
High level languages.
William Cleveland: Data Science
Leo Breimann: Statistical Modeling: 2 Cultures
Visualization
John Tukey
Jacques Bertin.
Edward Tufte.
Grammar of Graphics
Word Cloud, Tag Cloud.
Optimization Methods
Fourier and other transforms
Matrix & Generalizations
Non-euclidean geometries.
Applications to Military, manufacturing, Communications.
Networks
Assignment Problems
Automation
Scheduling
Calculus
Logarithms
Newton-Raphson.
Mathematics/ OR
1962 John W. Tukey, Future of Data Analysis
1976 – SAS Institute
1977 The International Association for Statistical Computing (IASC).
Decision Science
Pattern recognition
Machine learning.
Theoretical Foundations of Modern Stats
Hypothesis, DOE
Mathematical Statistics.
Simulation, Markov
Computational Statistics.
Probability
Correlation
Bayes Theorem.
Regression, Least Squares
Time Series.
Bayesian Methods
Time Series Methods (Box Cox, Survival, etc.)
Stochastic Methods.
Statistics
Pre 1800s
1800-1900
1900-1940
1940-1960
1960
1970
1980
1990
2000
2010
Source : Capgemini
shared via @pradeeprao_

ACKNOWLEDGMENTS

Thanks to the Zencos team for their help with ideas for content, editing support, and conversations about the best way to deliver this information. Special thanks to those that helped with editing and reviewing this paper, including Tricia Anderuud, Sean Ankenbruck, Reid Baughman, and Ivan Gomez. Thanks, as always, to Jared Weiss for helping promote bee references, editing, and guidance.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ben Murphy
Zencos
bmurphy@zencos.com
www.zencos.com