Preface

Note. The authors of *Mathematical Foundations of Big Data Analytics* (Springer-Verlag, 2021), Vladimir Shikhman and David Müller, state that this book is the result of the teaching of a class of the same name as part of a Master of Data Science program at Chemnitz University of Technology in Chemnitz, Germany (30 or so miles from both Leipzig and Dresden); see page vii. They describe the necessary background as “standard courses in higher mathematics” which, as I read it, is meant to include three courses in calculus (at ETSU, this would be Calculus 1 [MATH 1910], Calculus 2 [MATH 1920], and Calculus 2 [MATH 2110]), a course in linear algebra (at ETSU, this is Linear Algebra [MATH 2010]; a more advanced option is Theory of Matrices [MATH 5090]), and a calculus-based statistics class (at ETSU, this would be Foundations of Probability and Statistics–Calculus Based [MATH 2050], though Mathematical Statistics 1 [MATH 4047/5047] would be preferable). Some exposure to numerical techniques would be useful, such as ETSU’s Numerical Linear Algebra [STAT 4267/5267] (Numerical Analysis [MATH 4257/5257] would also be useful, but our models will be more based on linear algebraic concepts).

Note. The list of chapters of *Mathematical Foundations of Big Data Analytics* is:

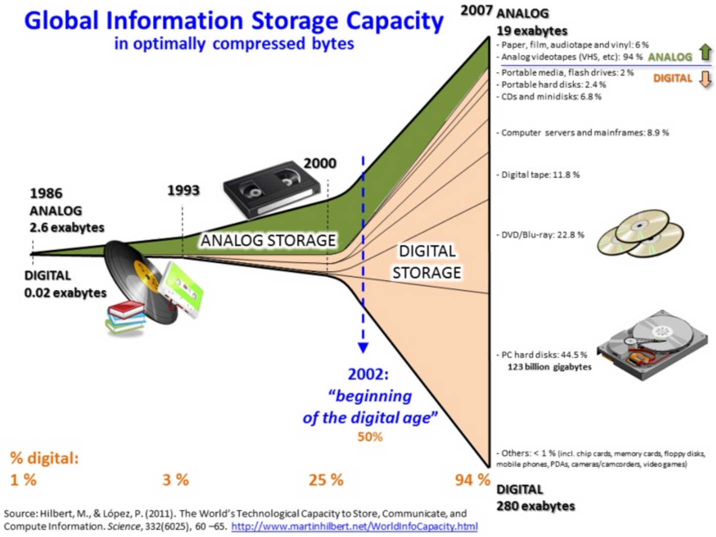
1. Ranking	6. Linear Regression
2. Online Learning	7. Sparsity
3. Recommendation Systems	8. Neural Networks
4. Classification	9. Decision Trees
5. Clustering	10. Solutions

Note. I have online notes for several of the background classes (some with more details than others, and some with associated videos). Relevant classes include:

CLASS	**Notes**	**Videos**
Calculus 1 (MATH 1910)	Online Notes	Online Videos
Calculus 2 (MATH 1920)	Online Notes	videos not available
Calculus 3 (MATH 2110)	Online Notes	videos not available
Linear Algebra (MATH 2010)	Online Notes	Online Videos
Mathematical Statistics 1 (STAT 4047/5047)	Online Notes	videos not available
Theory of Matrices (MATH 5090)	Online Notes	Online Videos

Simply click on an underlined link and you will be taken to the notes or videos (they are not password protected).

Note. The *amount* of data available has exploded during the first several years of the new millennium. Consider the following graph:



Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60 –65. http://www.martinhilbert.net/WorldInfoCapacity.html

From the Wikipedia page on Big Data (accessed 4/3/2021)

The increasing use of digital technology and the automatic collection of data (such as tracking the locations of cell phones, or internet searches) has certainly contributed to this volume of data (the *quality* of the data on the other hand…). For a few basic ideas, we turn to Wikipedia (not necessarily a top academic reference, but readily available, a source for more reliable sources, and seemingly appropriate for the discussion at hand).

Note. Chris Snijders, Uwe Matzat, and Ulf-Dietrich Reips describe "big data" as: "Big Data is a loosely defined term used to describe data sets so large and complex that they become awkward to work with using standard statistical software." (see their "Big Data": Big Gaps of Knowledge in the Field of Internet Science, *International Journal of Internet Science*, **7**(1), 1–5 (2012), available online [accessed 4/3/2021]). It seems part of the big data culture to describe the data with words starting with 'V'. Shikhman and Müller themselves mention Volume, Variety, Velocity (related to the high speed with which date is generated and analyzed), Validity (or "Veracity"), and Value; see page v. The idea of "validity" concerns the quality or accuracy of the data; as you know, traditional statistical tests are only useful in interpreting good data. In the Snijders et al paper (which is an editorial meant for "clarifying what type of questions and problems need input from the social and behavioral sciences"), a concern about applications of big data is stated as: "A crucial problem is that we do not know much about the underlying empirical micro-processes that lead to the emergence of these typical network characteristics of Big Data. Most of the underlying process models at the node level are inspired by mathematical ease of exposition, tractability or quite crude approximations of what could really be going on."

Note. The majority of examples presented by Shikhman and Müller are from the economic setting (brand loyalty in Section 1.3, credit investigation in Section 4.1, and capital asset pricing in Section 6.3). but they also include examples related to biology (DNA sequencing in Section 5.1 and nerve cells in Section 8.1), linguistics (semantic analysis in Section 3.3), sociology (community detection in Section 5.2), and engineering and computer science (quality control in Section 4.3 and spam filtering in Section 8.3).

Note. On a personal note, your humble instructor loves math but doesn't care much for numbers. I like statistics, but I'm not too keen on data! As a result, these notes are meant to take motivation from the "big data" or "predictive analytics" setting, but to maintain a high level of respect for the underlying mathematical and statistical ideas and to also maintain, as much as we can, mathematical rigor. So here we go...

Revised: 6/16/2021