STATISTICS, DATA SCIENCE AND MACHINE LEARNING[*]

Hedibert Freitas Lopes, Insper

June 25, 2017

Abstract

We will start with a few pointers to the recent raise of *data science* and its connections to statistics, machine learning and modern data analysis. In this handout all texts in red are hyperlinks to quite useful discussions on the interplay between statistics, computer science, machine learning, data science and (even) big data stuff.

[*] *Disclaimer:* These notes are by no means intended to be complete and/or original and simply reflect my rather limited view of the subject matters discussed in what follows. That being said, you are welcome to proceed on your own risk.

Contents

1 Introduction

1.1 Wikipedia's statistics, data mining, data science and machine learning

- Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.
- Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems.
- Data science is an interdisciplinary field about processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, which is a continuation of some of the data analysis fields such as statistics, machine learning, data mining, and predictive analytics.
- Machine learning is the subfield of computer science that "gives computers the ability to learn without being explicitly programmed." (Arthur Samuel, 1959)

1.2 American Statistical Association (ASA) - the role of statistics in data science

van Dyk, Fuentes, Jordan, Newton, Ray, Lang and Wickham (2015)[1] elaborated the American Statistical Association Statement on the Role of statistics in data science[2], where they say, amongst other things, that:

> *While there is not yet a consensus on what precisely constitutes data science, three professional communities, all within computer science and/or statistics, are emerging as foundational to data science: (i) Database Management enables transformation, conglomeration, and organization of data resources, (ii) Statistics and Machine Learning convert data into knowledge, and (iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis.*
>
> *At its most fundamental level, we view data science as a mutually beneficial collaboration among these three professional communities, complemented with significant interactions with numerous related disciplines. For data science to fully realize its potential requires maximum and multifaceted collaboration among these groups.*

[1] David van Dyk, Imperial College, Montse Fuentes, NCSU, Michael I. Jordan, UC Berkeley, Michael Newton, University of Wisconsin, Bonnie K. Ray, Pegged Software, Duncan Temple Lang, UC Davis, and Hadley Wickham, RStudio.

[2] http://magazine.amstat.org/blog/2015/10/01/asa-statement-on-the-role-of-statistics-in-data-science

1.3 Nancy Reid - Statistical science and data science: where do we go from here?

A summary of The 35th Fisher Memorial Lecture[3] appeared in the Royal Statistical Society's StatsLife page[4]. We quote a few important parts below:

> *Nancy Reid is a celebrated theoretical statistician; she has won numerous prizes including the Statistical Society of Canada Gold medal and the Royal Statistical Society's Guy medal in silver for her path-breaking and influential paper "Parameter Orthogonality and Approximate Conditional Inference", written jointly with Sir David Cox. She is director of the Canadian Statistical Sciences Institute, an Officer of the Order of Canada and is a past president of the the Statistical Society of Canada.*
>
> *In her lecture, Nancy recounted some of the perceptions she had encountered where 'statistical science' was unfavourably compared to 'big data'. The latter was associated with big machines and high-level computing whereas the former was being associated with small data and therefore less 'fun'. She acknowledged that there is currently a lot of hype surrounding 'big data', citing a 'hype curve' which showed that after the hype, big data was now being questioned with articles such as Tim Harford's 'Big Data – are we making a big mistake?' and Weapons of Math Destruction by Cathy O'Neil. There is also a growing suspicion of algorithms.*
>
> *Data science research, she said, could cover data collection and data quality, datasets with large 'n' and small 'p' as well as datasets with small 'n' and large 'p'. It could examine new types of data such as networks, graphs, digital text and images. It could include issues such as how to clean data, data management (ie, converting from raw to that which is 'analysable'), software programming, collaboration and project management.*
>
> *Nancy hopes that the area of data science will discover that the 'old core' is important, and that statistical scientists are often trying to solve a range of problems other than simply pattern recognition. Statisticians are often criticised for being too cautious, she said. However, lots of promises have been made around big data. Going back to her 'hype curve', Nancy concluded that the next big thing appears to be 'smart data'.*

1.4 Michael Jordan - Machine learning versus statistics

Source: reddit machine learning blog[5]

The blog starts with a short summary of Jordan's impressive vitae:

> *Michael I. Jordan is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley.*

[3] https://www.statslife.org.uk/events/eventdetail/699/11/data-science-the-view-from-the-mathematical-sciences
[4] https://www.statslife.org.uk/features/3072-event-report-the-35th-fisher-memorial-lecture-by-nancy-reid
[5] https://www.reddit.com/r/MachineLearning/comments/2fxi6v/ama_michael_i_jordan

He received his Masters in Mathematics from Arizona State University, and earned his PhD in Cognitive Science in 1985 from the University of California, San Diego. He was a professor at MIT from 1988 to 1998.

His research interests bridge the computational, statistical, cognitive and biological sciences, and have focused in recent years on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in distributed computing systems, natural language processing, signal processing and statistical genetics.

Prof. Jordan is a member of the National Academy of Sciences, a member of the National Academy of Engineering and a member of the American Academy of Arts and Sciences. He is a Fellow of the American Association for the Advancement of Science. He has been named a Neyman Lecturer and a Medallion Lecturer by the Institute of Mathematical Statistics. He received the David E. Rumelhart Prize in 2015 and the ACM/AAAI Allen Newell Award in 2009. He is a Fellow of the AAAI, ACM, ASA, CSS, IEEE, IMS, ISBA and SIAM.

Highlights of Jordan's interview:

Throughout the eighties and nineties, it was striking how many times people working within the "ML community" realized that their ideas had had a lengthy pre-history in statistics.

Decision trees, nearest neighbor, logistic regression, kernels, PCA, canonical correlation, graphical models, K-means and discriminant analysis come to mind, and also many general methodological principles (e.g., method of moments, Bayesian inference methods of all kinds, M estimation, bootstrap, cross-validation, EM, ROC, and stochastic gradient descent), and many many theoretical tools (large deviations, concentrations, empirical processes, Bernstein-von Mises, U statistics, etc).

When Leo Breiman developed random forests, was he being a statistician or a machine learner?

When my colleagues and I developed latent Dirichlet allocation, were we being statisticians or machine learners?

Are the SVM and boosting machine learning while logistic regression is statistics, even though they're solving essentially the same optimization problems?

I think the ML community has been exceedingly creative at taking existing ideas across many fields, and mixing and matching them to solve problems in emerging problem domains, and I think that the community has excelled at making creative use of new computing architectures. I would view all of this as the proto emergence of an engineering counterpart to the more purely theoretical investigations that have classically taken place within statistics and optimization.

But one shouldn't definitely not equate statistics or optimization with theory and machine learning with applications.

The "statistics community" has also been very applied, it's just that for historical reasons their collaborations have tended to focus on science, medicine and policy rather than engineering.

The emergence of the "ML community" has helped to enlargen the scope of "applied statistical inference". It has begun to break down some barriers between engineering thinking (e.g., computer systems thinking) and inferential thinking. And of course it has engendered new theoretical questions.

1.5 Additional reading material

Discussions about the interplay between statistics, data sciences, machine learning, big data, etc.

John Tukey (1962) The future of data analysis

David Hand (2013) Data mining: statistics and more?

Marie Davidian (2013) Aren't we data science?

Hal Varian (2014) Big data: new tricks for econometrics

Einav and Levin (2014) Economics in the age of big data

Athey and Imbens (2015) Lectures on machine learning

David Donoho (2015) 50 years of data science

Peter Diggle (2015) Statistics: a data science for the 21st century

van Dyk *et al.* (2015) Role of statistics in data science

Francis Diebold (2016) Machine learning versus econometrics

Uchicago (2016) Machine learning: what's in it for economics?

Coveney, Dougherty, Highfield (2016) Big data need big theory too

Franke *et al.* (2016) Statistical Inference, Learning and Models in Big Data

AMSTAT NEWS is the magazine of the ASA, the largest and oldest scientific association.

Davidian (1 jul 2013) Aren't we data science?

Bartlett (1 oct 2013) We are data science

Matloff (1 nov 2014) Statistics losing ground to computer science

van Dyk *et al.* (1 oct 2015) Role of statistics in data science

Jones (1 nov 2015) The identity of statistics in data science

Priestley (1 jan 2016) Data science: the evolution or the extinction of statistics?

See also Press (28 may 2013) A very short history of data science

2 Data Science

Table 1 list many of the US degrees in data sciences, data analytics, business analytics and variations. Below are two examples of a "data science iniciative" and a master's program in "analytics".

2.1 The Top 10 In-Demand Data Science Skills

Source: Crowdflower's 2016 Data Scientist report[6]

> *Data Science, as a field, is still evolving. Which is to say that what's a best practice today might be replaced by a better practice tomorrow. We looked at nearly 4,000 data science job postings on LinkedIn to find out what skills organizations wanted from their new hires. We ran those job postings through the CrowdFlower platform and had our contributors mark which skills showed up in which jobs.*
>
> *What data scientists spend the most time doing?*
> *What's the least enjoyable part of data science?*
>
> - *Building training sets: 3% & 10%*
> - *Cleaning and organizing data: 60% & 57%*
> - *Collecting data sets: 19% & 21%*
> - *Mining data for patterns: 9% & 3%*
> - *Refining algorithms: 4% & 4%*
> - *Other: 5% & 5%*
>
> *Here are the top 10 in-demand skills for data scientists (& % of jobs with skill):*
>
> - *SQL - 56%*
> - *Hadoop - 49%*
> - *Python - 39%*
> - *Java - 36%*
> - *R - 32%*
> - *Hive - 31%*
> - *Mapreduce - 22%*
> - *NoSQL - 18%*
> - *Pig - 16%*
> - *SAS - 16%*

2.2 Rice Data Science Iniciative

> *As part of Rice University's recently announced $150 million investment in research excellence, Rice intends to hire multiple faculty members whose research focus is in data science. We seek two distinct kinds of specialists: 1) data scientists who can make fundamental contributions to the theory, algorithms, and systems aspects of data*

[6]http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

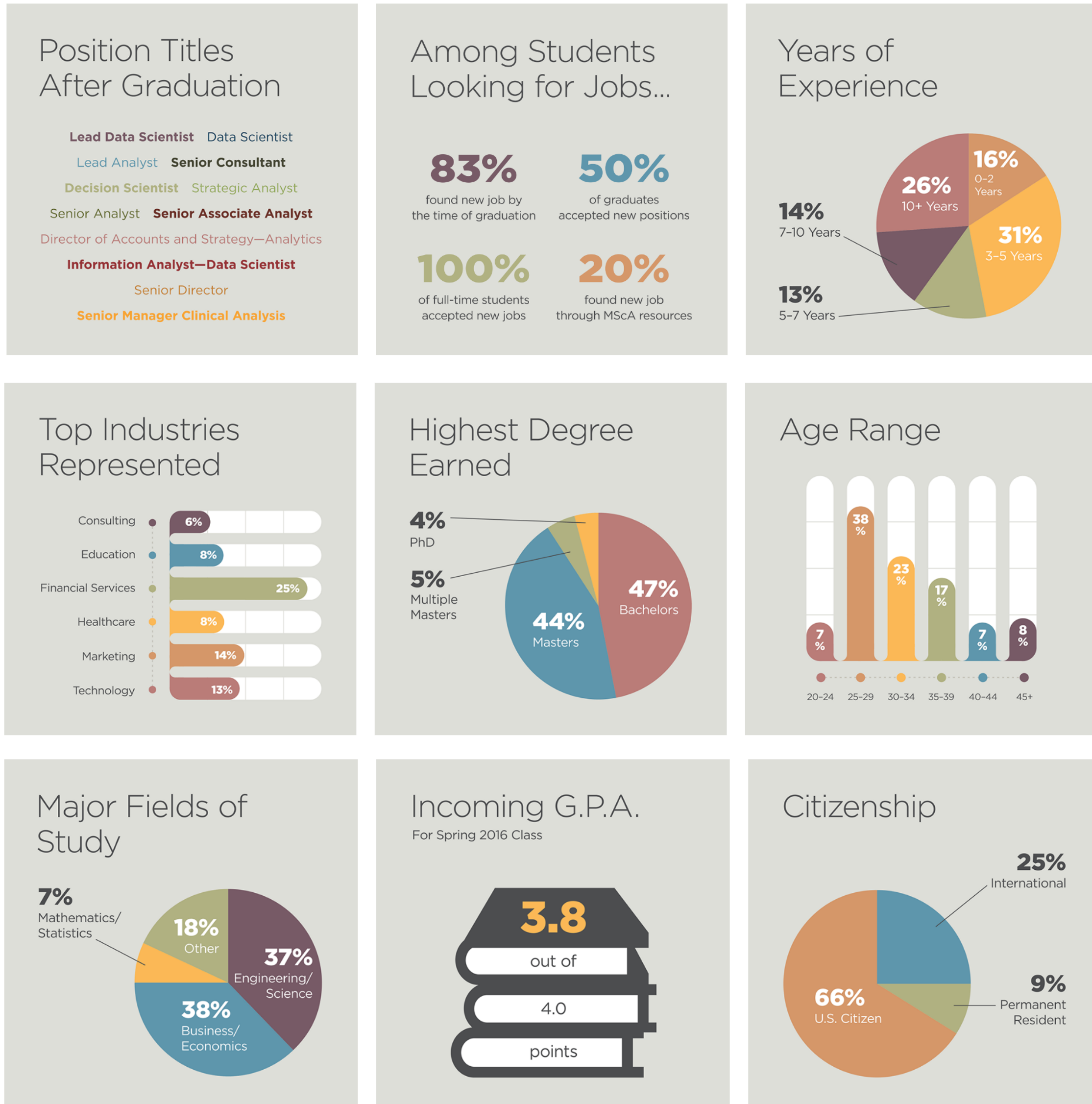
science, and 2) scholars who work on emerging applications of data science to problems in health and medicine, urban analytics, and high-volume or high-velocity data-intensive science. Highly motivated and qualified candidates with research experience in areas such as statistical inference and machine learning are encouraged to apply. A Ph.D. in Statistics, Computer Science, or related fields is required.

2.3 Chicago Booth's M.Sc. in Analytics

- Prerequisite courses
  Linear Algebra and Matrix Analysis
  Programming for Analytics

- 7 foundation courses
  Introduction to Statistical Concepts
  Research Design for Business Applications
  Statistical Analysis
  Database System Design and Implementation
  Time Series Analysis and Forecasting
  Data Mining Principles
  Machine Learning and Predictive Analytics
  Linear and Nonlinear Models for Business Application

- 3 electives
  Financial Analytics
  Marketing Analytics
  Credit and Insurance Risk Analytics
  Real Time Analytics
  Data Visualization Techniques
  Health Analytics
  Linear Algebra and Matrix Analysis
  Big Data and Text Analytics
  Optimization and Simulation Methods for Analytics
  Bayesian Methods
  Digital Marketing Analytics in Theory and Practice



2.4 Columbia University's M.Sc. in Data Science

The Master of Science in Data Science allows students to apply data science techniques to their field of interest, building on four foundational courses offered in our Certification of Professional Achievement in Data Sciences program. Our students have the opportunity to conduct original research, included in a capstone project, and interact with our industry partners and faculty. Students may also choose an elective track focused on entrepreneurship or a subject area covered by one of our seven centers.

Candidates for the Master of Science in Data Science are required to complete a minimum of 30 credits, including 21 credits of required/core courses and 9 credits of electives. This program may be pursued part-time or full-time.

Required/core courses:

- Probability theory
- Algorithms for data science
- Statistical inference and modeling
- Computer systems for data science
- Machine learning for data science
- Exploratory data analysis and visualization

2.5 University of Texas at Austin's M.Sc. in Business Analytics

The MSBA curriculum is designed to let students complete their 36 credit hours and receive their STEM Certified MSBA degree in just 10 months. Our program's coursework is made up of three main components:

- Foundation: Ensure proficiency in the technical skills needed in the fast-paced field of Business Analytics
- Application: Go beyond theories and concepts to see you how your new skills are currently being applied
- Demonstration: Go out of the classroom and work with real companies on their current projects, giving you real world experience

Required/core courses:

- Introduction to Predictive Modeling
- Data Analytics Programming
- Advanced Predictive Modeling
- Text Mining
- Decision Analysis
- Learning Structures and Time Series
- Stochastic Control and Optimization

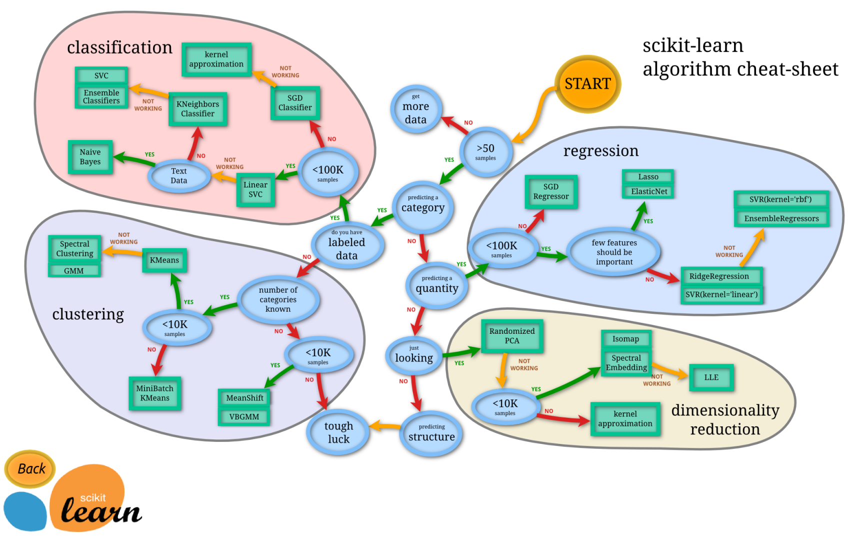
3 Machine learning

Source: Analytics Vidhya[7]

A typical course on ML contains most of the following topics:

- Linear regression
- Logistic regression
- Decision tree
- Support vector machines (SVM)
- Naive Bayes
- K nearest neighbours (KNN)
- K-means
- Random forest
- Dimensionality reduction algorithms
- Gradient boost & adaboost

A little bit of fun!



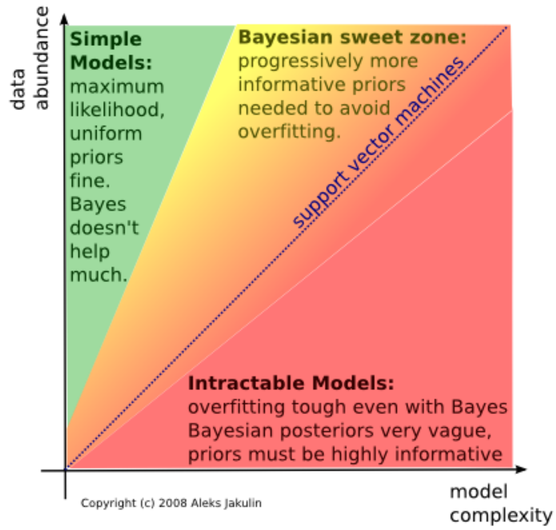
Glossary

Machine learning	Statistics
network, graphs	model
weights	parameters
learning	fitting
generalization	test set performance
supervised learning	regression/classification
unsupervised learning	density estimation, clustering
large grant = \$1,000,000	large grant= \$50,000
nice place to have a meeting: Snowbird, Utah, French Alps	nice place to have a meeting: Las Vegas in August

1

[7] https://www.analyticsvidhya.com/blog/2015/08/common-machine-learning-algorithms

3.1 Model complexity vs data abundance



Source: Aleks Jakulin (2008)

4 Statistical Inference, Learning and Models in Big Data

Source: Statistical Inference, Learning and Models in Big Data[8]

Paper's authors: Franke, Beate; Plante, Jean-François; Roscher, Ribana; Lee, Annie; Smyth, Cathal; Hatefi, Armin; Chen, Fuqi; Gil, Einat; Schwing, Alexander; Selvitella, Alessandro; Hoffman, Michael M.; Grosse, Roger; Hendricks, Dietrich and Reid, Nancy.

Paper's abstract: The need for new methods to deal with big data is a common theme in most scientific fields, although its definition tends to vary with the context. Statistical ideas are an essential part of this, and as a partial response, a thematic program on statistical inference, learning, and models in big data was held in 2015 in Canada, under the general direction of the Canadian Statistical Sciences Institute, with major funding from, and most activities located at, the Fields Institute for Research in Mathematical Sciences. This paper gives an overview of the topics covered, describing challenges and strategies that seem common to many different areas of application, and including some examples of applications to make these challenges and strategies more concrete.

> *Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them.*
>
> *Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy.*

[8]https://arxiv.org/pdf/1509.02900v2.pdf

The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

There is little doubt that the quantities of data now available are indeed large, but that's not the most relevant characteristic of this new data ecosystem.

4.1 Big data need big theory too

No matter their 'depth' and the sophistication of data-driven methods, such as artificial neural nets, in the end they merely fit curves to existing data.

Not only do these methods invariably require far larger quantities of data than anticipated by big data aficionados in order to produce statistically reliable results,

but they can also fail in circumstances beyond the range of the data used to train them because they are not designed to model the structural characteristics of the underlying system.

4.2 Big data's four Vs

The difficulty of transforming big data into knowledge is related to its complexity, the essence of which is broadly captured by the 'Four Vs': volume, velocity, variety and veracity (ethics and privacy)

Strategies for Big Data Analysis

- Data wrangling (data carpentry or data cleaning)
- Visualisation
- Reducing dimensionality
- Sparsity and regularisation
- Optimisation
- Measuring distance
- Representation learning (feature learning or deep learning)
- Sequential learning
- Multi-disciplinarity

5 Ten Simple Rules for Effective Statistical Practice

- Ten Simple Rules for Effective Statistical Practice[9]
  By Kass, Caffo, Davidian, Meng, Yu and Reid (2016)
  - Rule 1. Statistical Methods Should Enable Data to Answer Scientific Questions
  - Rule 2. Signals Always Come with Noise

[9] http://www.utstat.utoronto.ca/reid/research/journal.pcbi.1004961.pdf

– Rule 3. Plan Ahead, Really Ahead
– Rule 4. Worry about Data Quality
– Rule 5. Statistical Analysis Is More Than a Set of Computations
– Rule 6. Keep it Simple
– Rule 7. Provide Assessments of Variability
– Rule 8. Check Your Assumptions
– Rule 9. When Possible, Replicate!
– Rule 10. Make Your Analysis Reproducible (Nature special issue[10])
– Rule 0. Treat statistics as a science, not a recipe

6 Bayesian statistical learning

Bayesian Statistical Learning: Readings in Statistics and Econometrics[11]

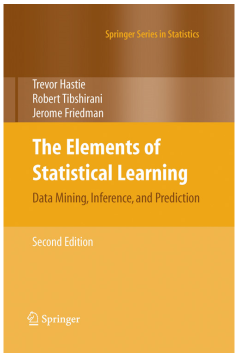
Organizer: Hedibert Freitas Lopes and Paulo Marques

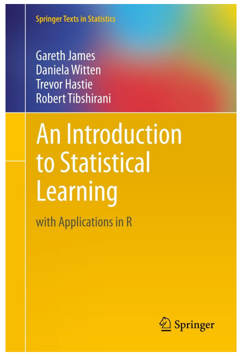
In this Readings in Statistics and Econometrics we will study and discuss, through a series of well established papers, the broad topic of Statistical Learning with an emphasis on its natural Bayesian solutions. We will give lectures discussing traditional Statistical Learning techniques, alternated with seminars given by the participants on papers presenting Bayesian counterparts to the techniques discussed in the lectures. Below is the outline of the 10 meetings (5 lectures and 9 seminars):

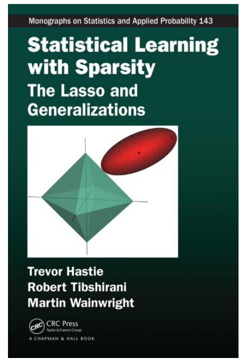
- Januaryt 29th: Lecture 1 - k-nearest neighors (k-NN)
- February 5th
  Seminar 1: Holmes and Adams (2002,2003)
  Seminar 2: Cucala, Marin, Robert and Titterington (2009)
- February 12th: Lecture 2 - LASSO regularization
- February 19th
  Seminar 3: Griffin and Brown (2010,2012,2014)
  Seminar 4: Polson, Scott and Windle (2014)
- February 26th: Lecture 3 - random forests
- March 4th
  Seminar 5: Chipman, George and McCulloch, (2010)
- March 11th: Lecture 4 - supporting vector machines
- March 18th
  Seminar 6: Tipping (2001)
  Seminar 7: Polson and Scott (2011)
- April 1st: Lecture 5 - k-means clustering
- April 8th
  Seminar 8: Dirichlet Processes
  Seminar 9: Kulis and Jordan (2012)

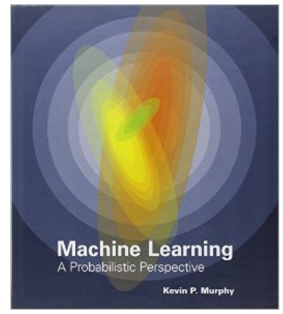
[10]http://www.nature.com/news/reproducibility-1.17552
[11]http://hedibert.org/current-teaching/#tab-statistical-learning

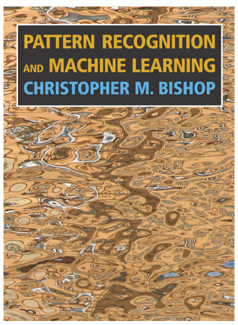
7 A few textbooks

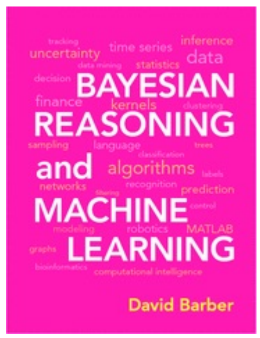


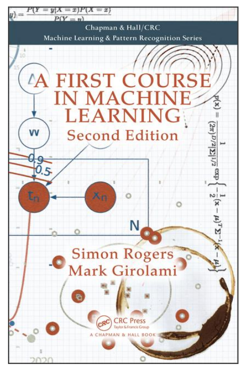


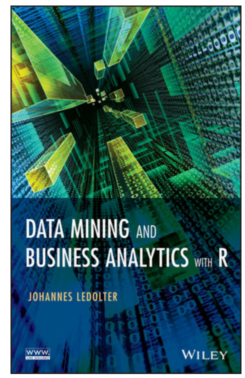


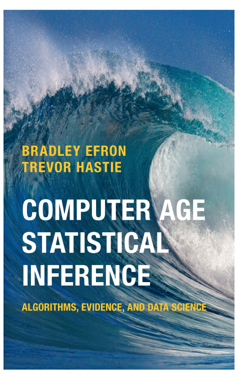


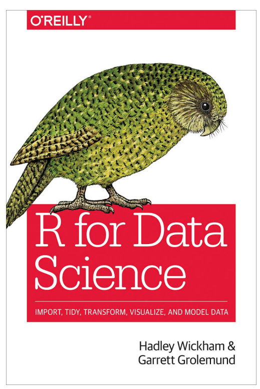


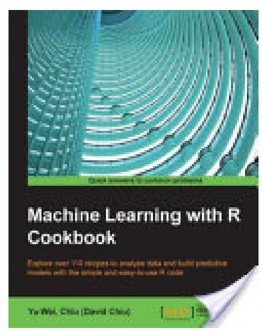


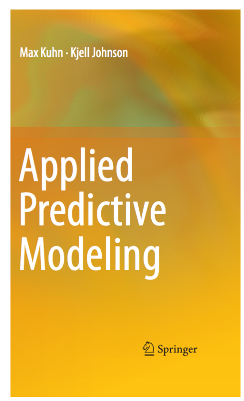












University	Degree	Credit	Established
Texas A&M University	Analytics	36	2013
Southern Methodist University	Applied Statistics and Data Analytics	36	2013
Arizona State University	Business Analytics	30	2013
Benedictine University	Business Analytics	64	2013
George Washington University	Business Analytics	33	2013
Michigan State University	Business Analytics	30	2013
New York University	Business Analytics	14	2013
Rensselaer Polytechnic Institute	Business Analytics	30	2013
University of Texas at Austin	Business Analytics	36	2013
Carnegie Mellon University	Computational Data Science	9	2013
Washington University in St. Louis	Customer Analytics	30	2013
Pace University	Customer Intelligence and Analytics	36	2013
City University of New York	Data Analytics	36	2013
Southern New Hampshire University	Data Analytics	12	2013
University of Maryland	Data Analytics	39	2013
Illinois Institute of Technology	Data Science	34	2013
New York University	Data Science	36	2013
Bowling Green State University	Analytics	33	2014
Dakota State University	Analytics	30	2014
Georgia Institute of Technology	Analytics	36	2014
Georgia State University	Analytics	32	2014
University of Chicago	Analytics	11	2014
Villanova University	Analytics	33	2014
Saint Louis University	Applied Analytics	36	2014
Maryville University	Applied Statistics and Data Analytics	36	2014
Indiana University	Business Analytics	30	2014
University of Colorado Denver	Business Analytics	30	2014
University of Denver	Business Analytics	58	2014
University of Miami	Business Analytics	16	2014
University of Minnesota	Business Analytics	45	2014
University of Rochester	Business Analytics	41	2014
University of Southern California	Business Analytics	27	2014
University of Texas at Dallas	Business Analytics	36	2014
Creighton University	Business Intelligence and Analytics	33	2014
St. Johns University	Data Mining and Predictive Analytics	30	2014
Elmhurst College	Data Science	30	2014
South Dakota State University	Data Science	30	2014
University of St. Thomas	Data Science	36	2014
University of Virginia	Data Science	11	2014
West Virginia University	Data Science	30	2014
Worcester Polytechnic Institute	Data Science	33	2014
Johns Hopkins University	Government Analytics	12	2014
University of California at Berkeley	Information and Data Science	27	2014
Philadelphia University	Modeling, Simulation and Data Analytics	30	2014
University of Arkansas	Statistics and Analytics	30	2014
Brandeis University	Strategic Analytics	30	2014
University of California, San Diego	Data Science and Engineering	38	2014
Georgetown University	Analytics	30	2015
University of New Hampshire	Analytics	36	2015
University of the Pacific	Analytics	30	2015
American University	Analytics—Online	33	2015
Valparaiso University	Analytics and Modeling	36	2015
College of William&Mary	Business Analytics	30	2015
Fairfield University	Business Analytics	30	2015
Iowa State University	Business Analytics	30	2015
Mercer University	Business Analytics	30	2015
Northeastern University	Business Analytics	30	2015
University of Dallas	Business Analytics	30	2015
University of Iowa	Business Analytics	30	2015
University of Notre Dame	Business Analytics	30	2015
University of Texas at Arlington	Business Analytics	36	2015
Xavier University	Customer Analytics	30	2015
Clarkson University	Data Analytics	33	2015
Slippery Rock University	Data Analytics	33	2015
Columbia University	Data Science	30	2015
Indiana University Bloomington	Data Science	30	2015
Southern Methodist University	Data Science	31	2015
University of Rochester	Data Science	30	2015
University of WisconsinExtension	Data Science	36	2015
University of North Carolina at Charlotte	Data Science	33	2015

Table 1: New degrees in Data Sciences, Data Analytics, Business Analytics, and related ones.