Modern Bayesian Statistics
Part I: Statistics, Data Science, Machine Learning

Hedibert Freitas Lopes[1]
hedibert.org

13^a aMostra de Estatística
IME-USP, October 2018

[1] Professor of Statistics and Econometrics at Insper, São Paulo.

Outline

Careercast's 10 best jobs of 2017[2]

Rank	Profession	Median Salary	Projected 7-year growth
1	Statistician	$80,110	34%
2	Medical Services Manager	$94,500	17%
3	Operations Research Analyst	$79,200	30%
4	Information Security Analyst	$90,120	18%
5	Data Scientist	$111,267	16%
6	University Professor	$72,416	15%
7	Mathematician	$111,298	22%
8	Software Engineer	$100,690	17%
9	Occupational Therapist	$81,910	29%
10	Speech Pathologist	$73,250	23%

[2] `http://www.careercast.com/jobs-rated/best-jobs-2017`

Careercast's 10 best jobs of 2018[3]

Rank	Profession	Median Salary	Projected 7-year growth
1	Genetic Counselor	$74,120	29%
2	Mathematician	$81,950	33%
3	University Professor	$75,430	15%
4	Occupational Therapist	$81,910	24%
5	Statistician	$84,060	33%
6	Medical Services Manager	$96,540	20%
7	Data Scientist	$111,840	19%
8	Information Security Analyst	$92,600	28%
9	Operations Research Analyst	$79,200	27%
10	Actuary	$100,610	22%

[3] `https://www.careercast.com/jobs-rated/2018-best-jobs`

Greater and Lesser Statistics[4]

Greater statistics can be defined . . . as everything related to learning from data, from the first planning or collection to the last presentation or report.

Lesser statistics is the body of specifically statistical methodology that has evolved within the profession – roughly, statistics as defined by texts, journals, and doctoral dissertations.

Greater statistics tend to be inclusive, eclectic with respect to methodology, closely associated with other disciplines, and practiced by many outside of academia and often outside professional statistics.

Lesser statistics tends to be exclusive, oriented to mathematical techniques, less frequently collaborative with other disciplines, and primarily practiced by members of university departments of statistics.

[4] Chambers (1993) Greater or lesser statistics: A choice for future research. *Statistics and Computing*, 3(4), 182-184.

Data science vs. statistics: two cultures?[5]

[W]e define data science as the union of six areas of greater data science, based on Donoho (2017) 50 years of data science. Journal of Computational and Graphical Statistics, *26(4), 745-766:*

1. *Data gathering, preparation, and exploration.*
2. *Data representation and transformation.*
3. *Computing with data.*
4. *Data modeling.*
5. *Data visualization and presentation.*

We take the position that data science is a reaction to the narrow understanding of lesser statistics; simply put, data science has come to mean a broader view of statistics.

[5] Carmichael and Marron (2018) *Japanese Journal of Statistics and Data Science*, 1, 117-138. `https://doi.org/10.1007/s42081-018-0009-3`

Statistics 101

One can be forgiven . . . for mistaking statistics as a set of recipes.

Too many people interact with statistics exclusively via a standard Statistics 101 type class which may in fact treat statistics as a handful of formulas to memorize and steps to follow.

While we believe the material taught in these courses is vital to doing science, it is perhaps time to rethink such introductory classes and teach data before (or concurrently with) teaching statistics.

Some principal components of data science

Prediction vs. inference - do vs. understand - engineering vs. science

- Engineering is the business of creating a thing that does something. Science is the business of understanding how something works.
- Predictive modeling is one of the main drivers of artificial intelligence (AI). Modern AI systems are typically based on deep learning and are extremely data hungry

Empirically vs. theoretically driven

- Data science is exploratory data analysis gone mad. – Neil Lawrence
- "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete" it was argued that EDA will replace the scientific method.
  We disagree.
  This article is an extreme example of the broader attitude that correlation, and fancy models applied to large data sets, can replace causal inference and the careful, time intensive scientific method.

Some principal components of data science

Problem first vs. hammer looking for a nail

- Both research approaches are valid and productive, however the balance in academic statistics may have shifted too far to the former (hammer) approach.
- Data science is focused on problem solving and it is this problem solving which makes data analysis useful to other disciplines.

The 80/20 rule (maybe could even by the 90/10 rule)
The basic idea is that the first reasonable thing you can do to a set of data often is 80% of the way to the optimal solution. Everything after that is working on getting the last 20%.

Master in Data Science: 2007-2011

University	Degree	Credit	Established
North Carolina State University	Analytics	30	2007
University of Tennessee at Knoxville	Business Analytics	39	2010
Saint Joseph's University	Business Intelligence and Analytics	30	2010
Louisiana State University at Baton Rouge	Analytics	39	2011
University of Cincinnati	Business Analytics	35	2011
Northwestern University	Predictive Analytics	11	2011

Master in Data Science: 2012

University	Degree	Credit	Established
Northwestern University	Analytics	11	2012
University of San Francisco	Analytics	35	2012
Drexel University	Business Analytics	45	2012
Fordham University	Business Analytics	30	2012
University of Michigan at Dearborn	Business Analytics	30	2012
Stevens Institute of Technology	Business Intelligence and Analytics	36	2012

Master in Data Science: 2013

University	Degree	Credit	Established
Harrisburg University of Science and Technology	Analytics	36	2013
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

Master in Data Science: 2014

University	Degree	Credit	Established
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
Bentley University	Business Analytics	30	2014
Indiana University	Business Analytics	30	2014
Quinnipiac University	Business Analytics	33	2014
Southern Methodist University	Business Analytics	33	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. John's 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

Master in Data Science: 2015

University	Degree	Credit	Established
Capella University	Analytics	48	2015
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 Wisconsin's Extension	Data Science	36	2015
University of North Carolina at Charlotte	Data Science	33	2015
Penn State Great Valley	Data Analytics	30	2015

Ciência de dados no Brasil: Formação executiva

1. FGV: Formação executiva de machine learning
   Carga Horária: 64h
2. FIAP: Big Data Science: Machine Learning e Data Mining
3. FIA: MBA Analytics em Big Data
   Carga Horária: 600 horas
4. IGTI: MBA em Ciência dos Dados & Big Data
   Carga Horária: 370 horas
5. Unisul: MBA em Engenharia e Ciência dos Dados
   Carga Horária: 375 horas
6. PUC-Minas: Ciência dos Dados e Big Data
   Carga Horária: 432 horas

Ciência de dados no Brasil: Pós-Graduação

1. Einstein (Especialização): Data science e informática para área da saúde
   Carga Horária: 420 horas
2. IESB (Especialização): Ciência dos Dados
   Carga Horária: 400 horas
3. UNIFACCAMP (Lato Sensu): Mineração e Ciência dos Dados
   Carga Horária: 392 horas
4. Faculdades Integradas de Bauru: Data science com ênfase machine learning
   Carga Horária: 360 horas
5. São Carlos: Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria
6. UFPR (Especialização): Data Science e Big Data
   Carga Horária: 390 horas
7. Newton Paiva (Especialização): Ciência dos Dados e Big Analytics
   Carga Horária: 360 horas
8. Centro Universitário Faria de Brito (Especialização): Especialização em Ciência dos Dados
   Carga Horária: 427 horas
9. Uni7 (Especialização): Especialização em Ciência de Dados com Big Data, BI e Data Analytics
   Carga Horária: 406 horas
10. UniChristus (Especialização): Ciência dos Dados e Inteligência de Negócios (Big Data e BI)
    Carga Horária: 405 horas
11. UFBA (Especialização): Especialização em Ciência de Dados e Big Data
    Carga Horária: 476 horas
12. UFRGS (Especialização): Big Data & Data Science
    Carga Horária: 360 horas
13. UNISINOS (Especialização/EAD): Big Data, Data Science & Data Analytics
    Carga Horária: 360 horas
14. Poli-PE (Especialização): Ciência dos Dados e Analystics Carga Horária: 360 horas

Statistics, data sciences, machine learning, big data

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

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

ASA Statement on the Role of Statistics in Data Science

"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,

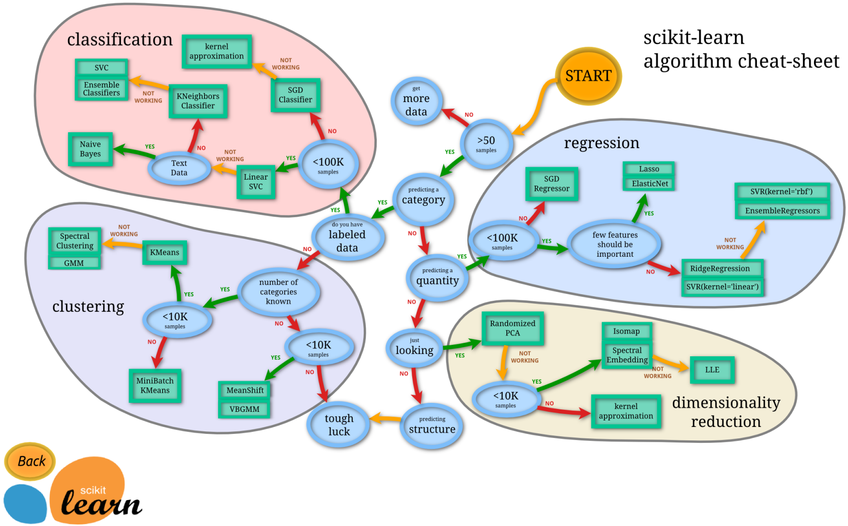
(ii) Statistics and Machine Learning convert data into knowledge, and

(iii) Distributed and Parallel Systems provide the computational infrastructure to carry out data analysis."

Machine learning

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

Source: Analytics Vidhya


scikit-learn
algorithm cheat-sheet
START
classification
regression
clustering
dimensionality
reduction
get
more
data
>50
samples
predicting a
category
do you have
labeled
data
predicting a
quantity
just
looking
predicting
structure
tough
luck
<100K
samples
SGD
Classifier
kernel
approximation
KNeighbors
Classifier
SVC
Ensemble
Classifiers
Naive
Bayes
Text
Data
Linear
SVC
Spectral
Clustering
GMM
KMeans
<10K
samples
MiniBatch
KMeans
number of
categories
known
MeanShift
VBGMM
SGD
Regressor
Lasso
ElasticNet
few features
should be
important
SVR(kernel='rbf')
EnsembleRegressors
RidgeRegression
SVR(kernel='linear')
Randomized
PCA
Isomap
Spectral
Embedding
LLE
YES
NO
NOT
WORKING
Back
scikit
learn

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

Michael Jordan on ML vs Statistics

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 neighboor, 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).

Source: reddit machine learning blog

Michael Jordan (more)

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.

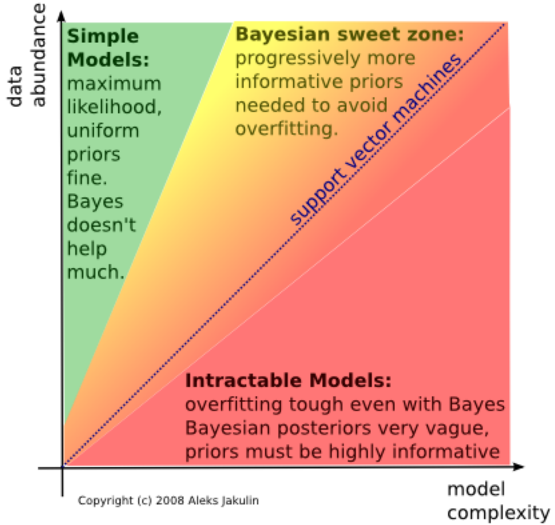
Michael Jordan (a bit more)

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.

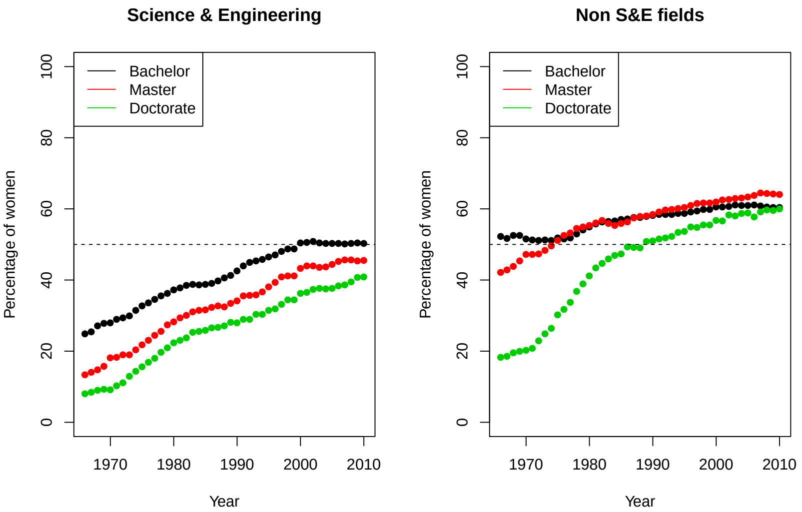
The emergence of the "ML community" has helped to enlargen the scope of "applied statistical inference". It has begun to break down some barriers between engineering thinking (e.g., computer systems thinking) and inferential thinking. And of course it has engendered new theoretical questions.

Model complexity vs data abundance



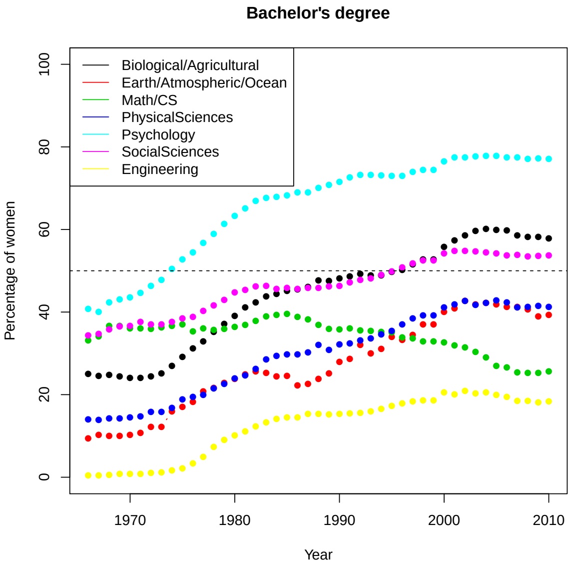
Source: Aleks Jakulin (2008)

Women in Science & Engineering[6]

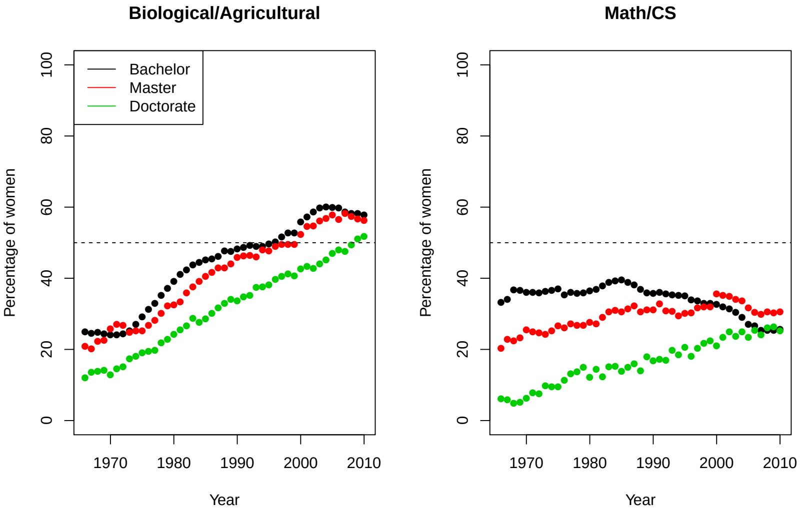


[6] https://www.nsf.gov/statistics/nsf13327

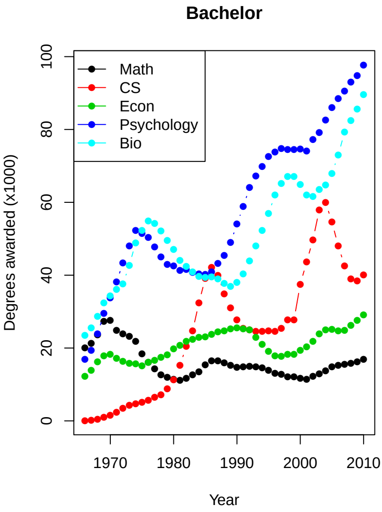
Bachelor's degrees awarded to women

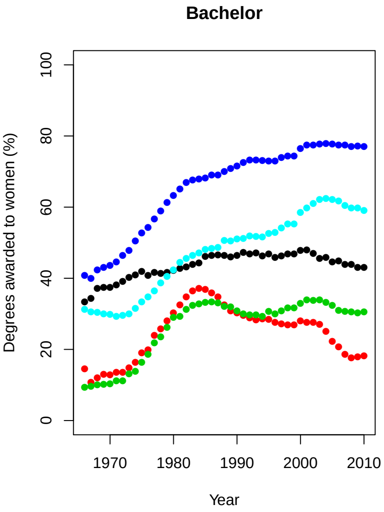


Bachelor's, Master's and Doctorate's degrees



Degrees awarded in several fields





Percentage of degrees awarded to women in several fields