**DATA SCIENCE: STATISTICS OR COMPUTER SCIENCE?
IMPLICATIONS FOR STATISTICS EDUCATION**

Timothy J. Kyng, Ayse Bilgin, Busayasachee Puang-Ngern
Macquarie University, Australia

Abstract

Big Data / Data Science is a very important emerging area for statisticians. Software skills are increasingly important for statistical practitioners. Data science may be regarded by statisticians as a new name for statistical science but in industry and government the perception may be different. Recent advances in IT have enabled us to collect, store and easily access large amounts of data with modest cost. The capacity to analyse the data and use it for decision making has lagged behind. Software has been developed to filter, access and analyse data. Computer scientists and statisticians have been working separately, not jointly on this. This paper explores the implications of Big Data for statisticians' education and aims to identify what skills are needed and software packages to use as well as the gaps between the perceptions of practitioners and academics about these issues.

DATA SCIENCE: STATISTICS OR COMPUTER SCIENCE?
IMPLICATIONS FOR STATISTICS EDUCATION

- *Data Science is a very important emerging area for statisticians. Software skills are increasingly important for statistical practitioners.*
- *Data science may be regarded by statisticians as a new name for statistical science but in industry and government the perception may be different.*
- *This paper explores the implications of Data Science for statisticians' education and aims to identify what skills are needed and software packages to use as well as the gaps between the perceptions of practitioners and academics about these issues.*
- *We analysed recent job advertisements and conducted surveys of graduates in industry and academics to identify what are the important skills and the important software tools for working in DS in practice.*

DATA SCIENCE: IMPLICATIONS FOR THE STATISTICS DISCIPLINE
IS STATISTICS DEAD OR DYING?

- *Advances in IT: enabled us to collect, store, and easily access large amounts of data with modest cost. The capacity to analyse the data and use it for decision making has lagged behind. Software has been developed to filter, access and analyse data.*
- *Due to inadequate computer science education, many statisticians & actuaries are behind other professionals in the data analytics space.*
- *Most DS courses are very IT focused and business agnostic, volume of statistical theory and practice covered in these is low*
- *Data Scientists have skills which are in demand and which many statisticians lack. However the Data Scientists also lack many of the statistical skills which statisticians do have.*

DATA SCIENCE EDUCATION

- *lots of courses available - many introduced very recently*
- *8 of Australia's 38 universities have newly established DS postgrad degrees*
- *large variation in fees: from **free** (Coursera MOOC offered by Johns Hopkins University DS qualification / certificate) to **expensive** ($USD $60,000 Master of Information and Data Science at UC Berkeley)*
- *Professional societies are also moving (or have moved) to provide CPE courses in DS for their members: e.g. the French Actuarial Society has done this and the Australian Actuaries Institute is considering this.*

DATA SCIENCE EDUCATION – WHAT DOES IT COVER?

- *French Actuarial Society DS CPD 1 year part time course covers*

Python, R and data mining, Machine Learning, Parallel Computation, Data Manipulation and Visualization

- *Monash University Grad diploma 2 year part time course covers*

analytical theory, R and Python, big data processing tools such as Hadoop and Spark, data engineering and wrangling to visualisation and data management

DATA SCIENCE EDUCATION – WHAT DOES IT COVER?

- *Analysis of the content of many of the DS degrees shows that these degrees are very IT focussed and the volume of statistical theory and methodology covered is low.*
- *Many statistical methods are not covered at all or very briefly: e.g. extreme value theory, general insurance reserving methods, theory of statistical inference, theory of maximum likelihood estimation, linear models, generalised linear models, modelling of low frequency but high impact events (e.g. large losses in insurance, extreme events in finance)*
- *Consequently many types of statistical work couldn't be done by some of the DS graduates or practitioners*

SURVEY OF GRADUATES WORKING IN DS AND ACADEMICS IN RELEVANT DISCIPLINES

- *Created email list of graduates from 3 universities: Macquarie U, UWS, and Chulalongkorn U, who graduated between 2003 and 2014. Disciplines were: Stats, CS, Actuarial Science, IT, and Math. Used **snowball sampling** to reach the target population. We had 72 responses to our graduate survey (population size unknown).*
- *Created email list of academics from 39 Australian and 8 NZ universities' websites. Targeted Stats, CS, Actuarial Science, IS, IT, Math, and Marketing disciplines. From 163 university departments sampled, 62 university departments responded. The response rate for the university departments was 38%.*
- *Online questionnaires conducted via the Qualtrics Surveys website. Separate questionnaires for the 2 groups but the questions covered similar issues to facilitate comparison between the perceptions of academics and graduates working in industry. Some questions required selecting a rating from strongly disagree to strongly agree using a five point Likert scale for various statements or options.*

SURVEY OF GRADUATES WORKING IN DS AND ACADEMICS IN RELEVANT DISCIPLINES

Both surveys had

- *2 questions about **generic skills / expertise** required for employment in the Big Data field,*
- *2 questions about **software tools and skills**;*
- *4 questions about **demographic information** on the respondents.*
- *The academics' survey included questions about their workplace, their experience in the Big Data area, and about degree programs and subjects offered.*
- *Graduate survey included questions about participants' education, workplace information and their opinions about the Big Data / Data Analytics roles*

SURVEY OF GRADUATES WORKING IN DS AND ACADEMICS IN RELEVANT DISCIPLINES

- *Sample size was small, and not randomly selected*
- *Detailed statistical analysis and statistical significance testing not performed, and not justified in the circumstances*
- *This is a descriptive study showing the situation (a snapshot) at the time the study was conducted*
- *Results presented as rankings of the various generic skills and of the various software tools. These ranks are indicative only, not measured with statistical precision*
- *Despite the deficiencies of the study / sample the results should still of interest to the statistics community*

SURVEY RESULTS – IMPORTANCE OF TYPES OF EXPERTISE

Expertise Area	graduates rating	academics rating	graduates ranking	academics ranking
Statistical Analysis	4.40	4.60	1	1
Data Mining	4.20	4.20	3	2
Statistical Learning	4.30	4.20	2	3
Programming	3.90	4.00	5	4
Mathematics	3.80	4.00	6	5
Machine Learning	3.80	3.70	7	6
Business Analysis	4.00	3.50	4	7
Artificial Intelligence	3.30	3.30	8	8
Marketing	3.30	3.10	9	9
Accounting	2.90	2.60	10	10

SURVEY RESULTS – IMPORTANCE OF TYPES OF EXPERTISE

- *Likert scale responses used to compute average rating for each type of expertise, then used to rank these*
- *We note that the overall rankings by the graduates and the academics are quite similar for the types of expertise*
- ***Statistical analysis** had the **highest ranking** for both groups*
- *The 2 groups agreed on which 3 (statistical analysis, data mining, machine learning) were the most important and which 3 (AI, marketing, accounting) were the least important*
- *"business analysis" ranked higher by graduates than by academics*

SURVEY RESULTS – IMPORTANCE OF SOFTWARE TOOLS

graduates		academics	
Software tool	rating	Software tool	rating
SQL	4.2	R	4.2
R	3.8	Python	3.9
SPSS Analytics	3.8	SQL	3.8
SPSS Modeler	3.8	Hadoop	3.7
Base SAS	3.8	MapReduce	3.6
Hadoop	3.7	Matlab	3.6
SAS Enterprise Miner	3.6	SPSS Analytics	3.5
Java	3.6	SAS Enterprise Miner	3.4
Oracle	3.6	Java	3.4
NoSQL	3.6	SPSS Modeler	3.3
Python	3.5	Base SAS	3.3
MapReduce	3.4	Oracle	3.3
Matlab	3.4	NoSQL	3.1
Hive	3.3	JaQL	3.1
VBA	3.3	Hive	3.0
JaQL	3.2	VBA	3.0
WinBUGS	3.0	WinBUGS	2.8

SURVEY RESULTS – IMPORTANCE OF SOFTWARE TOOLS

- SQL & R ranked highly by both groups (in top 3)
- winbugs ranked lowest by both groups and hive, vba, JaQL and winbugs the 4 lowest ranked by both groups
- Python ranked much lower by graduates than by academics
- MapReduce and Matlab ranked much lower by graduates than by academics
- SPSS analytics, SPSS modeler and BASE SAS ranked higher by graduates than by academics

RESULTS

- There is a higher level of agreement between the graduates and the academics regarding the generic skills needed than there is about the software tools.
- Statistical analysis ranked # 1 by both groups
- Python is a feature of DS education for most of the DS degrees we looked at but is rated a lot lower by our sample of graduates than by the academics
- SQL & R were rated highly by both groups. R is common in many statistics degrees but SQL is less so.
- Many of the graduates working in DS do not have degrees in DS. They probably learned to use the software tools on the job. DS degrees are relatively new.
- We have anecdotal evidence from talking to our recent graduates that there is high staff turnover in many of these jobs in the DS area

NEW DATA SCIENCE DEGREES IN AUSTRALIA

University of Melbourne	(Master of Business Analytics)
University of Technology Sydney	(Master of Data Science and Innovation)
Monash U	(Graduate Diploma in Data Science)
UWA	(Data Science at the undergraduate level)
Deakin University	(Master of Business Analytics)
Victoria University	(Master of Business Analytics)
RMIT	(Master of Analytics)
University of South Australia	(Master of Data Science)

FRENCH ACTUARIAL SOCIETY DATA SCIENCE CPD COURSE

Duration: 2 days per month over 1 year

Actuaries are insurance mathematicians / applied statisticians usually with good statistical background and programming skills. This course covers:

- A. Python and Computational Aspects
- B. R and data mining
- C. Algorithmic aspects of Machine Learning
- D. Probabilistic Aspects of Machine Learninig
- E. Parallel Computation
- F. Data Manipulation and Visualization
- G. Case Studies

MONASH UNIVERSITY (AUSTRALIA) GRADUATE DIPLOMA IN DATA SCIENCE

Fully online, 1.4 years duration, 8 units of study, fee = AUD $27,500.00

- Introduction to Data Science
- Modelling for Data Analysis
- Data exploration (retrieve relevant data from large, unorganized pools)
- Data Curation and Management (maintain data so it is available for reuse and preservation)
- Data analysis
- Data engineering (Data Infrastructure / Architecture: The data engineer gathers and collects the data, stores it, processes it, make it available for queries)
- Applied data analytics
- Advanced data wrangling (aka data cleaning)
- Business intelligence modelling

UC BERKELEY MASTER OF INFORMATION AND DATA SCIENCE DEGREE

Fee = $60,000 USD, 9 subjects , duration = 20 months

- Research Design and Application for Data and Analysis
- Exploring and Analyzing Data
- Storing and Retrieving Data
- Applied Machine Learning
- Visualizing and Communicating Data
- Field Experiments
- Legal, Policy & Ethical Considerations for Data Scientists
- Scaling Up! Really Big Data
- Synthetic Capstone Course

JOHNS HOPKINS UNIVERSITY FREE MOOC COURSES IN DATA SCIENCE.

Fee = $0 USD, 9 subjects , duration = 1 year

I. The Data Scientist's Toolbox (4 weeks 4 hpw)
II. R Programming (9 weeks 8 – 9 hpw)
III. Getting and Cleaning Data (4 weeks 4-9 hpw)
IV. Exploratory Data Analysis (4 weeks 4-9 hpw)
V. Reproducible Research (4 weeks 4-9 hpw)
VI. Statistical Inference (4 weeks 7 – 9 hpw)
VII. Regression Models (4 weeks 4-9 hpw)
VIII. Practical Machine Learning (4 weeks 4-9 hpw)
IX. Developing Data Products (4 weeks 4-9 hpw)
X. Data Science Capstone (8 weeks 5-9 hpw)

IMPLICATIONS FOR STATISTICIAN'S EDUCATION.

- In a mathematical statistics degree, collecting and managing data is typically not covered, same applies to data cleaning ("wrangling")
- Data science degrees are packed with content and heavily IT focussed. This is largely due to the large volume of data to be stored and the development of new hardware and software tools store, manage and analyse it.
- It would be difficult to include all of what's in a DS degree into a statistics degree without either increasing the volume of learning or sacrificing some of what we currently teach to make way for this new content.
- To remain competitive statistician's education needs to include more exposure to software tools. Tools such as SQL, Python, Machine Learning, Data Mining, Data Visualisation. A student wishing to have a career as a statistician needs to learn about data science but not necessarily to become a data scientist. Not all of the data that needs analysing is "big".
- Data science is a separate discipline but one statisticians cannot ignore

DATA SCIENCE: STATISTICS OR COMPUTER SCIENCE? IMPLICATIONS FOR STATISTICS EDUCATION

Timothy J. Kyng, Ayse Bilgin, Busayasachee Puang-Ngern
Macquarie University, Australia

Abstract

Big Data / Data Science is a very important emerging area for statisticians. Software skills are increasingly important for statistical practitioners. Data science may be regarded by statisticians as a new name for statistical science but in industry and government the perception may be different. Recent advances in IT have enabled us to collect, store and easily access large amounts of data with modest cost. The capacity to analyse the data and use it for decision making has lagged behind. Software has been developed to filter, access and analyse data. Computer scientists and statisticians have been working separately, not jointly on this. This paper explores the implications of Big Data for statisticians' education and aims to identify what skills are needed and software packages to use as well as the gaps between the perceptions of practitioners and academics about these issues.

DATA SCIENCE: STATISTICS OR COMPUTER SCIENCE? IMPLICATIONS FOR STATISTICS EDUCATION

- *Data Science is a very important emerging area for statisticians. Software skills are increasingly important for statistical practitioners.*
- *Data science may be regarded by statisticians as a new name for statistical science but in industry and government the perception may be different.*
- *This paper explores the implications of Data Science for statisticians' education and aims to identify what skills are needed and software packages to use as well as the gaps between the perceptions of practitioners and academics about these issues.*
- *We analysed recent job advertisements and conducted surveys of graduates in industry and academics to identify what are the important skills and the important software tools for working in DS in practice.*

DATA SCIENCE: IMPLICATIONS FOR THE STATISTICS DISCIPLINE
IS STATISTICS DEAD OR DYING?

- *Advances in IT: enabled us to collect, store, and easily access large amounts of data with modest cost. The capacity to analyse the data and use it for decision making has lagged behind. Software has been developed to filter, access and analyse data.*
- *Due to inadequate computer science education, many statisticians & actuaries are behind other professionals in the data analytics space.*
- *Most DS courses are very IT focused and business agnostic, volume of statistical theory and practice covered in these is low*
- *Data Scientists have skills which are in demand and which many statisticians lack. However the Data Scientists also lack many of the statistical skills which statisticians do have.*

DATA SCIENCE EDUCATION

- *lots of courses available - many introduced very recently*
- *8 of Australia's 38 universities have newly established DS postgrad degrees*
- *large variation in fees: from **free** (Coursera MOOC offered by Johns Hopkins University DS qualification / certificate) to **expensive** ($USD $60,000 Master of Information and Data Science at UC Berkeley)*
- *Professional societies are also moving (or have moved) to provide CPE courses in DS for their members: e.g. the French Actuarial Society has done this and the Australian Actuaries Institute is considering this.*

DATA SCIENCE EDUCATION – WHAT DOES IT COVER?

- *French Actuarial Society DS CPD 1 year part time course covers*

Python, R and data mining, Machine Learning, Parallel Computation, Data Manipulation and Visualization

- *Monash University Grad diploma 2 year part time course covers*

analytical theory, R and Python, big data processing tools such as Hadoop and Spark, data engineering and wrangling to visualisation and data management

DATA SCIENCE EDUCATION – WHAT DOES IT COVER?

- *Analysis of the content of many of the DS degrees shows that these degrees are very IT focussed and the volume of statistical theory and methodology covered is low.*
- *Many statistical methods are not covered at all or very briefly: e.g. extreme value theory, general insurance reserving methods, theory of statistical inference, theory of maximum likelihood estimation, linear models, generalised linear models, modelling of low frequency but high impact events (e.g. large losses in insurance, extreme events in finance)*
- *Consequently many types of statistical work couldn't be done by some of the DS graduates or practitioners*

SURVEY OF GRADUATES WORKING IN DS AND ACADEMICS IN RELEVANT DISCIPLINES

- *Created email list of graduates from 3 universities: Macquarie U, UWS, and Chulalongkorn U, who graduated between 2003 and 2014. Disciplines were: Stats, CS, Actuarial Science, IT, and Math. Used **snowball sampling** to reach the target population. We had 72 responses to our graduate survey (population size unknown).*
- *Created email list of academics from 39 Australian and 8 NZ universities' websites. Targeted Stats, CS, Actuarial Science, IS, IT, Math, and Marketing disciplines. From 163 university departments sampled, 62 university departments responded. The response rate for the university departments was 38%.*
- *Online questionnaires conducted via the Qualtrics Surveys website. Separate questionnaires for the 2 groups but the questions covered similar issues to facilitate comparison between the perceptions of academics and graduates working in industry. Some questions required selecting a rating from strongly disagree to strongly agree using a five point Likert scale for various statements or options.*

SURVEY OF GRADUATES WORKING IN DS AND ACADEMICS IN RELEVANT DISCIPLINES

Both surveys had

- *2 questions about **generic skills / expertise** required for employment in the Big Data field,*
- *2 questions about **software tools and skills**;*
- *4 questions about **demographic information** on the respondents.*
- *The academics' survey included questions about their workplace, their experience in the Big Data area, and about degree programs and subjects offered.*
- *Graduate survey included questions about participants' education, workplace information and their opinions about the Big Data / Data Analytics roles*

SURVEY OF GRADUATES WORKING IN DS AND ACADEMICS IN RELEVANT DISCIPLINES

- *Sample size was small, and not randomly selected*
- *Detailed statistical analysis and statistical significance testing not performed, and not justified in the circumstances*
- *This is a descriptive study showing the situation (a snapshot) at the time the study was conducted*
- *Results presented as rankings of the various generic skills and of the various software tools. These ranks are indicative only, not measured with statistical precision*
- *Despite the deficiencies of the study / sample the results should still of interest to the statistics community*

SURVEY RESULTS – IMPORTANCE OF TYPES OF EXPERTISE

Expertise Area	graduates rating	academics rating	graduates ranking	academics ranking
Statistical Analysis	4.40	4.60	1	1
Data Mining	4.20	4.20	3	2
Statistical Learning	4.30	4.20	2	3
Programming	3.90	4.00	5	4
Mathematics	3.80	4.00	6	5
Machine Learning	3.80	3.70	7	6
Business Analysis	4.00	3.50	4	7
Artificial Intelligence	3.30	3.30	8	8
Marketing	3.30	3.10	9	9
Accounting	2.90	2.60	10	10

SURVEY RESULTS – IMPORTANCE OF TYPES OF EXPERTISE

- *Likert scale responses used to compute average rating for each type of expertise, then used to rank these*
- *We note that the overall rankings by the graduates and the academics are quite similar for the types of expertise*
- ***Statistical analysis*** *had the* ***highest ranking*** *for both groups*
- *The 2 groups agreed on which 3 (statistical analysis, data mining, machine learning) were the most important and which 3 (AI, marketing, accounting) were the least important*
- *"business analysis" ranked higher by graduates than by academics*

SURVEY RESULTS – IMPORTANCE OF SOFTWARE TOOLS

graduates		academics	
Software tool	rating	Software tool	rating
SQL	4.2	R	4.2
R	3.8	Python	3.9
SPSS Analytics	3.8	SQL	3.8
SPSS Modeler	3.8	Hadoop	3.7
Base SAS	3.8	MapReduce	3.6
Hadoop	3.7	Matlab	3.6
SAS Enterprise Miner	3.6	SPSS Analytics	3.5
Java	3.6	SAS Enterprise Miner	3.4
Oracle	3.6	Java	3.4
NoSQL	3.6	SPSS Modeler	3.3
Python	3.5	Base SAS	3.3
MapReduce	3.4	Oracle	3.3
Matlab	3.4	NoSQL	3.1
Hive	3.3	JaQL	3.1
VBA	3.3	Hive	3.0
JaQL	3.2	VBA	3.0
WinBUGS	3.0	WinBUGS	2.8

SURVEY RESULTS – IMPORTANCE OF SOFTWARE TOOLS

- SQL & R ranked highly by both groups (in top 3)
- winbugs ranked lowest by both groups and hive, vba, JaQL and winbugs the 4 lowest ranked by both groups
- Python ranked much lower by graduates than by academics
- MapReduce and Matlab ranked much lower by graduates than by academics
- SPSS analytics, SPSS modeler and BASE SAS ranked higher by graduates than by academics

RESULTS

- There is a higher level of agreement between the graduates and the academics regarding the generic skills needed than there is about the software tools.
- Statistical analysis ranked # 1 by both groups
- Python is a feature of DS education for most of the DS degrees we looked at but is rated a lot lower by our sample of graduates than by the academics
- SQL & R were rated highly by both groups. R is common in many statistics degrees but SQL is less so.
- Many of the graduates working in DS do not have degrees in DS. They probably learned to use the software tools on the job. DS degrees are relatively new.
- We have anecdotal evidence from talking to our recent graduates that there is high staff turnover in many of these jobs in the DS area

NEW DATA SCIENCE DEGREES IN AUSTRALIA

University of Melbourne	(Master of Business Analytics)
University of Technology Sydney	(Master of Data Science and Innovation)
Monash U	(Graduate Diploma in Data Science)
UWA	(Data Science at the undergraduate level)
Deakin University	(Master of Business Analytics)
Victoria University	(Master of Business Analytics)
RMIT	(Master of Analytics)
University of South Australia	(Master of Data Science)

FRENCH ACTUARIAL SOCIETY DATA SCIENCE CPD COURSE

Duration: 2 days per month over 1 year

Actuaries are insurance mathematicians / applied statisticians usually with good statistical background and programming skills. This course covers:

- A. Python and Computational Aspects
- B. R and data mining
- C. Algorithmic aspects of Machine Learning
- D. Probabilistic Aspects of Machine Learninig
- E. Parallel Computation
- F. Data Manipulation and Visualization
- G. Case Studies

MONASH UNIVERSITY (AUSTRALIA) GRADUATE DIPLOMA IN DATA SCIENCE

Fully online, 1.4 years duration, 8 units of study, fee = AUD $27,500.00

- Introduction to Data Science
- Modelling for Data Analysis
- Data exploration (retrieve relevant data from large, unorganized pools)
- Data Curation and Management (maintain data so it is available for reuse and preservation)
- Data analysis
- Data engineering (Data Infrastructure / Architecture: The data engineer gathers and collects the data, stores it, processes it, make it available for queries)
- Applied data analytics
- Advanced data wrangling (aka data cleaning)
- Business intelligence modelling

UC BERKELEY MASTER OF INFORMATION AND DATA SCIENCE DEGREE

Fee = $60,000 USD, 9 subjects , duration = 20 months

- Research Design and Application for Data and Analysis
- Exploring and Analyzing Data
- Storing and Retrieving Data
- Applied Machine Learning
- Visualizing and Communicating Data
- Field Experiments
- Legal, Policy & Ethical Considerations for Data Scientists
- Scaling Up! Really Big Data
- Synthetic Capstone Course

JOHNS HOPKINS UNIVERSITY FREE MOOC COURSES IN DATA SCIENCE.

Fee = $0 USD, 9 subjects , duration = 1 year

I. The Data Scientist's Toolbox (4 weeks 4 hpw)
II. R Programming (9 weeks 8 – 9 hpw)
III. Getting and Cleaning Data (4 weeks 4-9 hpw)
IV. Exploratory Data Analysis (4 weeks 4-9 hpw)
V. Reproducible Research (4 weeks 4-9 hpw)
VI. Statistical Inference (4 weeks 7 – 9 hpw)
VII. Regression Models (4 weeks 4-9 hpw)
VIII. Practical Machine Learning (4 weeks 4-9 hpw)
IX. Developing Data Products (4 weeks 4-9 hpw)
X. Data Science Capstone (8 weeks 5-9 hpw)

IMPLICATIONS FOR STATISTICIAN'S EDUCATION.

- In a mathematical statistics degree, collecting and managing data is typically not covered, same applies to data cleaning ("wrangling")
- Data science degrees are packed with content and heavily IT focussed. This is largely due to the large volume of data to be stored and the development of new hardware and software tools store, manage and analyse it.
- It would be difficult to include all of what's in a DS degree into a statistics degree without either increasing the volume of learning or sacrificing some of what we currently teach to make way for this new content.
- To remain competitive statistician's education needs to include more exposure to software tools. Tools such as SQL, Python, Machine Learning, Data Mining, Data Visualisation. A student wishing to have a career as a statistician needs to learn about data science but not necessarily to become a data scientist. Not all of the data that needs analysing is "big".
- Data science is a separate discipline but one statisticians cannot ignore