BIG DATA SKILLS:

CHALLENGES FOR THE UNIVERSITY WORLD

CREATING A NEW GENERATION OF DATA SCIENTISTS

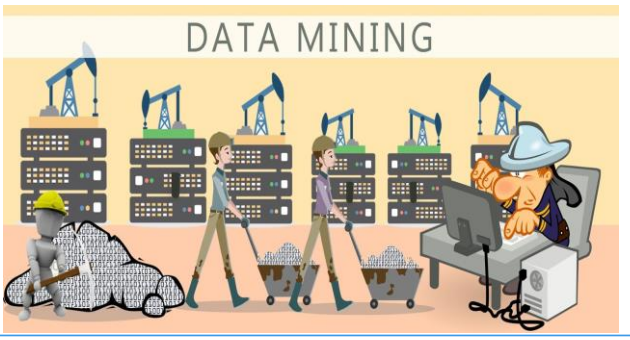
Massimiliano Marcellino
Bocconi University

CES 2017 Seminar on the new generation of statisticians and data scientists
20 June 2017, Geneva

Università Commerciale Luigi Bocconi

INTRODUCTION

- About a century ago, oil was a new commodity that generated massive changes in the world economy. Today we also have a "new" commodity: DATA!
- Data were with us for millennia but now people recognize that they are a key economic ingredient, like oil. What happened?



Università Commerciale Luigi Bocconi

THE DATA REVOLUTION…

- Advances in technology have permitted:
  - Improved storage, access and handling of traditional data, collected for example by public administration, banks, insurance companies, utilities, etc.
  - Easier and cheaper access and handling of data produced by hi-tech devices, such as satellites, medical devices, electronic instruments in planes and ships, etc.
  - Mass access to computers and the internet, with the associated digitalization of many activities, from dating to shopping, which permits to store the associated data
  - Mass introduction of connected mobile devices (phones, cameras, watches, etc.) and digital connection of an increasing number of things, from appliances to cars and all sorts of industrial machineries

THE DATA REVOLUTION…

- Data are now much more abundant, ubiquitous, and by far more valuable than just a few years ago
- But data should be “refined” to be useful and valuable, just like oil…

Bocconi

STATISTICS FOR THE DATA REVOLUTION

- The increased value added of data is related to advances in the statistical tools for (big) data analysis, which can be grouped into two main areas: Data Science and Business Analytics.
- Data Science is the practice of using automated methods to analyze massive amounts of data and to extract knowledge from them.
- Business Analytics is the practice of transforming data into (business) insights to allow for better decision-making.
- The two areas are typically kept separated, but the largest gains likely come from their merger, for example in the context of statistical institutes.

Bocconi

THE DATA REVOLUTION & STATISTICAL INSTITUTES

- How to collect, process and disseminate data after the (on-going) data revolution?
- How to compete (or cooperate) with private “data refineries”?
- Need of both Data Science (data collection & processing) and Business Analytics (data dissemination, on-demand statistics, extraction and communication of the relevant signal, new areas (social, environment, etc) and new needs (improved precision, timeliness, coverage, frequency,…) in old areas (economics, business, etc), …)
- Are there enough competencies in statistical institutes at the moment? What is missing? Who could provide the missing competencies?
- Stronger integration with the research community, universities in particular

Università Commerciale Luigi Bocconi

Bocconi

CHALLENGES FOR THE UNIVERSITY WORLD

- I would like to consider a few issues:
  - Is there sufficient offer of academic degrees in Business Analytics (BA) and in Data Science (DS)?
  - What is (or should be) taught in these degrees?
  - How long do they last and are they full-time or part-time?
  - What is the difference with respect to standard degrees in Statistics and related areas?

Bocconi

ACADEMIC DEGREES IN BA AND DS

- An internet search reveals at least 26 Master of Science (MSc) degrees in Business Analytics (BA) and 11 in Data Science (DS), offered by good-top level universities around the world.
- Overall, there are many more, for example MastersPortal.eu reports 251 Masters in Data Science & Big Data ...
- … and Bocconi just started an undergrad program in Econ and Computer Science, with the associated MSc in BA&DS starting next academic year!
- Average duration varies between 9m and 3y, median duration is 12m.
- Full-time attendance is generally required (but there is also plenty of shorter on-line courses, though difficult to assess their quality).

Università Commerciale Luigi Bocconi

Bocconi

SELECTED MSC IN BUSINESS ANALYTICS (1/2)

University	Business	Economics	CS	Statistics	School/faculty	Program	Duration
Arizona State	55.6%	0.0%	44.4%	0.0%	School of Business / Information Systems	Business Analytics	9 months FT
Case Western Reserve Ohio	90.0%	0.0%	0.0%	10.0%		MSM-Business Analytics	11 months FT
Drexel U	33.0%	0.0%	33.0%	33.0%	LeBow College of Business	Business Analytics	12 months FT
ESSEC	25.0%	0.0%	25%	50.0%	Information Systems, Decision Sciences, and Statistics	Data Sciences & Business Analytics	9 months FT
George Washington U	0.0%	0.0%	12.5%	87.5%	Business School / Decision Sciences	Business Analytics	12-16 months FT
IESEG	60.7%	0.0%	14.3%	17.9%	School of Management	Big Data Analytics for Business	1 year / 3 terms
Imperial College of London	20.0%	10.0%	30.0%	40.0%	Business School	Business Analytics	12 months FT
Indiana U	50.0%	16.7%	33.3%	0.0%	Kelly School of Business	Business Analytics	15 months FT
Michigan State U	59.3%	0.0%	22.2%	18.5%	Eli Broad College of Business	Business Analytics	12 months FT

Bocconi

SELECTED MSC IN BUSINESS ANALYTICS (2/2)

University	Business	Economics	CS	Statistics	School/faculty	Program	Duration
MIT	22.2%	0.0%	44%	0.0%	Sloan School of Management	Business Analytics	12 months FT
NUS	28.6%	0.0%	28.6%	42.8%	School of Computing and Business School	Business Analytics	12 months FT
NYU	53.3%	6.7%	23.3%	16.7%	School of Business Information, Operations & Management Sciences	Business Analytics	12 months PT
Politecnico di Milano	45.5%	0.0%	49.1%	5.5%	Graduate School of Business	Master in Business Analytics and Big Data	12 months FT
Purdue	91.7%	8.3%	0.0%	0.0%	Krannert School of Management	Business Analytics and Information Management	11 months FT
Rensselaer Polytechnic Institute	100%	0.0%	0.0%	0.0%	School of Management/	Business Analytics	9 months FT
Southern Methodist U	9.1%	0.0%	14.8%	18.5%	Cox School of Business	Business Analytics	12 months FT
Syracuse Un	50.0%	0.0%	16.7%	33.3%	Business School	Business Analytics	18 months FT
UC San Diego	100%	0.0%	0.0%	0.0%	Rady School of Management	Business Analytics	12 months FT

Università Commerciale Luigi Bocconi

Bocconi

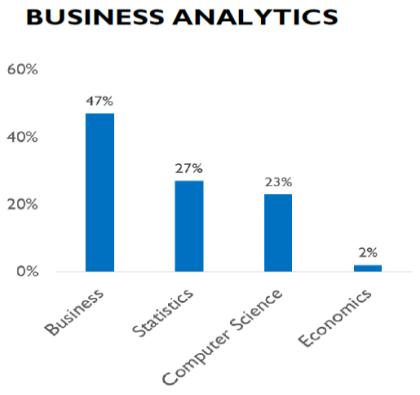
SELECTED MSC IN DATA SCIENCE

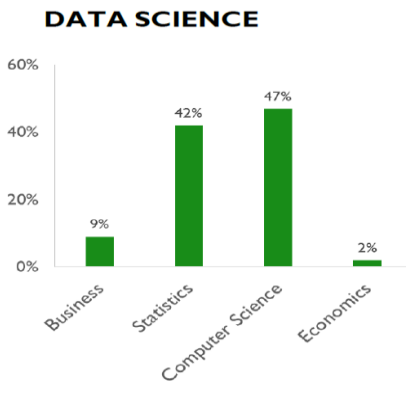
University	Business	Economics	CS	Statistics	School/ Faculty	Program	Duration
Barcelona	9.1%	9.1%	27.3%	54.5%	Graduate School of Economics	MSc Data Science	9 months FT
Columbia University	0.0%	0.0%	43%	43%	School of Engineering and Applied Science / Data Science Institute	MSC Data Science	1 year FT
NYU	0.0%	0.0%	18.8%	81.2%	Center for Data Science	MSC Data Science	2 years FT
Politecnico di Torino	0.0%	0.0%	25.0%	75.0%	School of Sciences /Mathematics, Computer Science, Science economiche-social e matematiche-statistiche	M.Sc. in Stochastics and Data Science	2 years FT
Pompeu Fabra	27.3%	9.1%	27.3%	36.4%	Economics and Business	MSc in Data Science	9 months FT
Techinische Universitat Dortmund	0.0%	0.0%	50.0%	50.0%	Statistics /Statistics; Information Technology, and Mathematics	MSc Data Science	2 years FT
UC Berkeley	0.0%	0.0%	40.0%	60.0%	School of Information	Master of Information and Data Science	20 months FT
UCL	0.0%	0.0%	33.3%	66.7%	Department of Computer Science	MSc Data Science	1 year FT
Università di Bologna	50.0%	0.0%	37.5%	12.5%	Business School	MSc Data Science	1 year FT
Universite Paris-Saclay	0.0%	0.0%	45.5%	54.5%	ENSAE ParisTech, Ecole Polytechnique, Télécom-ParisTech, Université Paris Sud	MSc Data Science	1 year FT
University of Amsterdam	16.7%	0.0%	66.7%	16.7%	Faculty of Science	MSc Data Science	1 year FT

Bocconi

COMPULSORY COURSES IN BA AND DS

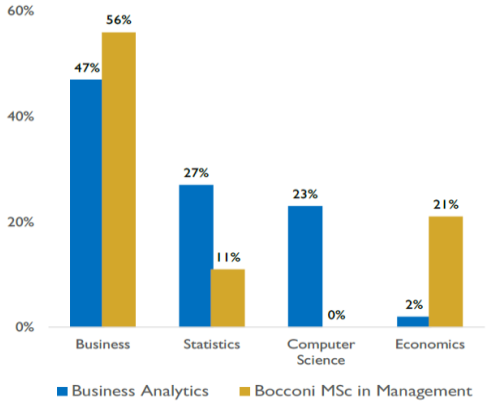
- The compulsory courses offered in each MSc can be classified into four main scientific areas: Business, Statistics, Computer Science, Economics

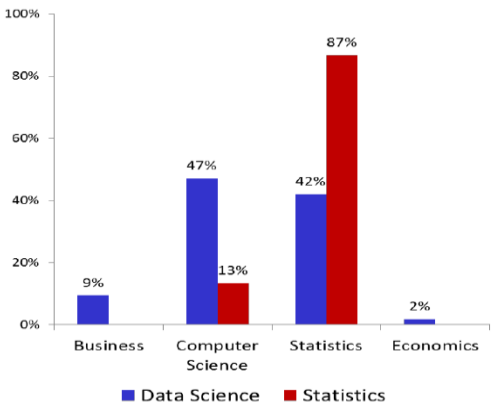




COMPULSORY COURSES IN BA AND DS

- Let us now consider the "average" differences of MScs in BA with respect to standard MScs in Management, and of MScs in DS wrt Statistics





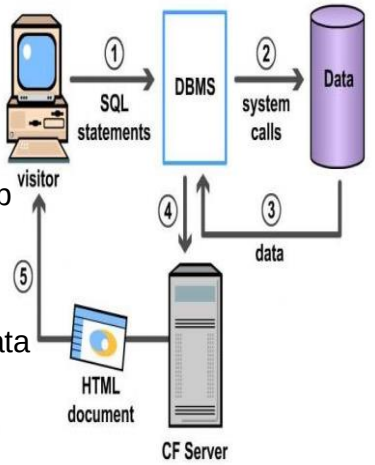
WHICH COMPULSORY COURSES?

- Some core competencies are often common to BA and DS degrees, such as:
  - Database Management
  - Statistics and Probability
  - Machine Learning
- Other competencies are more specific, with BA degrees offering more business/economics/finance courses and DS degrees more computing/statistics
- Both core and elective courses are typically tailored to the new big data features, such as size, structure or lack of it, sparsity, etc. Let us look at some examples.

Bocconi

DATABASE MANAGEMENT

- Courses in database management typically cover topics such as:
  - Database design
  - Current definitive DB languages (SQL)
  - Data quality, missing values, transformations
  - Data scraping, cleansing and de-duping
  - Data transfer
  - Big Data, Linked Data and the Semantic Web
  - Web scraping and APIs
  - In-memory data management
  - Hadoop and the MapReduce framework
  - Structured, Unstructured, Semi-structured data
  - Data Visualization and Manipulation
  - Descriptive, Predictive and Prescriptive data
  - Use of Data tools



Bocconi

STATISTICS AND PROBABILITY

- Courses in statistics and probability typically cover topics such as:
  - Overview of elementary tools from statistics and probability
  - Basics of statistical inference (estimation and hypothesis testing)
  - Applied regression analysis
  - Variable selection (lasso, ridge, etc.)
  - Resampling techniques (MC, bootstrap, MCMC, etc.)
  - Beyond linearity: smoothing splines and generalized additive models
  - Tree-based models (regression and classification trees, bagging, random forests, boosting, etc.)
  - Unsupervised learning (principal component analysis, clustering, etc)

MACHINE LEARNING

- Courses in machine learning typically cover topics such as:
  - Models for discrete data and Gaussian models
  - Bayesian statistics
  - Classification: logistic and probit regression
  - Nearest neighbour, Bayesian and linear classifiers
  - Directed graphical models
  - Mixture models
  - High-dimensional data and sparse linear models
  - Variational inference
  - Monte Carlo methods
  - State space models
  - Kernel methods for classification and regression
  - Gaussian and Dirichlet processes
  - Neural networks

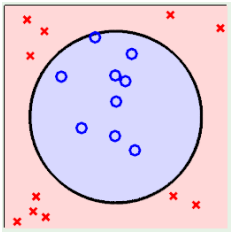
IS IT ENOUGH?

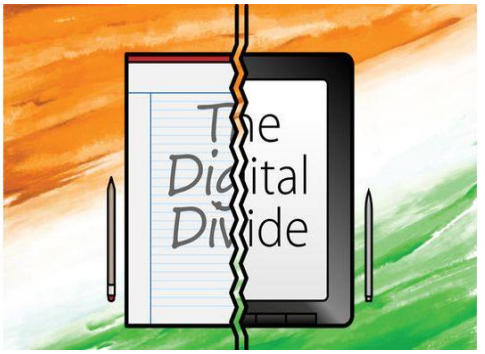
- For the moment, a data analytics talent gap is clearly evident. McKinsey Global Institute projects a shortage of 190,000 data scientists by 2018
- Business analytics expertise is scarce - ranked second in a Computer World survey on the most difficult skills to find.
- There are also unaccounted “network effects”: by collecting more data, a firm improves its products/services, which attracts more customers, which generates more data, which requires more analysts…
- On the other hand, people with a background in business, economics, statistics, math, engineering, etc. can be relatively easily self-trained

ARE THERE ANY CHALLENGES?

❖ It is important to also provide a critical assessment of the usefulness of big data and reliability of associated statistical methods. For example:

- Private ownership and provision of data and software
- Digital “divide”
- Digital “bias”
- Continuous evolution
- Data snooping
- Interaction with traditional indicators





CONCLUSIONS

❖ Big Data are precious, even more so when used in combination with proper statistical methods to analyze and extract knowledge from them, and when the extracted knowledge is transformed into (business) insights to allow for better decision-making.

❖ Big data pose big challenges for statistical institutes: How to collect, process and disseminate data after the (on-going) data revolution? How to compete (or cooperate) with private “data refineries”? Are there enough competencies in statistical institutes at the moment? Where and how to get the missing competencies?

❖ Stronger integration with the research community, universities in particular

Bocconi

CONCLUSIONS

- Big Data also pose big challenges for the university world, as they require a set of specific competencies that were partly lacking from traditional degrees in statistics and related areas. To tackle this issue, universities now offer a variety of degrees in data science and business analytics, but continuous monitoring and updating is needed.
- Academic research is also challenged, as Big Data have a set of features related to their size, structure and scope that make traditional statistical and econometric methods unsuited, as well as some of the methods directly imported from computer science, which are primarily meant for structured i.i.d. data.
- We are investigating some of these fascinating issues in a joint project with Eurostat, and we look forward to sharing the results with you.