MSc Data Science
at the University of Sheffield

Started in September 2014
Gianluca Demartini

• Lecturer in Data Science at the Information School since 2014
  – Ph.D. in Computer Science at U. Hannover, Germany
  – M.Sc. in Computer Science at U. Udine, Italy
• Worked at UC Berkeley (USA), Yahoo! Research (Spain), L3S Research Center (Germany), U. Fribourg (Switzerland)
• Teaching Social Computing in Fribourg
• Tutorials on Entity Search at ECIR 2012, on Crowdsourcing at ESWC 2013 and ISWC 2013
• Research Interests
  – Entity-centric Information Access, Semantic Search, Human Computation and Crowdsourcing
What is Data Science?

• An emerging field that seeks to discover and explore new ways of exploiting data to support decision-making for a range of domains and problems

• Data Science has the “potential to be a deep and profound research discipline impacting all areas of our lives” (O’Neil & Schutt, 2013)

• Most organisations, including health care, the public sector, local government, retail and manufacturing, are amassing vast amounts of heterogeneous data in real-time (i.e. “Big Data”)

• There is greater demand than ever before to manage, analyse and use data effectively, ethically and in a rigorous and scientific manner – need to seize the data opportunity
Seizing the data opportunity

A strategy for UK data capability

“In the information economy, the ability to handle and analyse data is essential for the UK’s competitive advantage and business transformation.”

The ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that’s going to be a hugely important skill in the next decades.” (Hal Varian)
"Data is the new product". Forrester: Top IT Predictions For 2015

Gianluca Demartini
Example Data Science problems

- **Supermarkets** identifying trends in shopper buying patterns *(e.g., people who bought X also bought Y)*
- **Universities** predicting whether students are likely to leave degree programmes before completion
- **Banks** wanting to identify fraudulent behaviour *(e.g., anomalous spending patterns)*
- **Financial organisations** predicting **stock market** changes based on analysis of events reported in the news
- **Organisations** characterising user groups and analysing **opinions from social media** sites for market intelligence
- **Local government** analysing and monitoring use of public resources
- ...
Data science skills

• “The UK requires a strong skills base, able to manage, analyse, interpret and communicate data, in order to extract insight and value” Seizing the Data Opportunity (HM Government, 2013)

• Data Science solutions involve knowledge and understanding of
  – Technologies (e.g., data warehousing, Hadoop)
  – Data modelling (e.g., representing and aggregating multiple datasets)
  – Data standards (e.g., open data and Linked Data)
  – Data analysis (e.g., data mining)
  – Communication (e.g., data visualisation, generate data reports)
  – Wider context (e.g., business processes, governance and ethics)
Type II Data Scientist is a “data-savvy manager” who is able to communicate with the Type I Data Scientist, is familiar with the technologies and processes involved, can manage cross-disciplinary teams, but also understands how to interpret and communicate results from analysing the data in the context of the organisation as a whole.

The Type II Data Scientist will understand the activities involved in the broader data lifecycle, such as data generation, adding metadata, data storage, data presentation, data access, the use and re-use of data, data governance and the strategic implications of data-driven decision making.

Source:
http://www.oralytics.com/2013/03/type-i-and-type-ii-data-scientists.html
Job opportunities

• A recent report by the e-Skills UK and SAS highlight that the demand for Big Data skills is expected to rise 92% between 2012-2017

• Organisations require ‘data savvy managers’ to manage and support **data-driven decision-making** built on three core skills
  – Data management (storage and linking), data analysis (including data literacy) and business and policy insight (context awareness)

• Potential job roles beyond IT and management include
  – ‘Big Data’ jobs, Business Analyst, Business Intelligence Analyst, Data Analyst, Data Architect, Data Mining Engineer, Data Scientist, Data Warehousing Manager
The Sheffield perspective

• Provide an Information Science perspective on the field rather than, e.g. Computer Science
  – Develop understanding of **Data Science in the wider business and societal contexts**, e.g. governance, security, data management, organisational decision making, literacy
  – Focus on **how people interact with data** and technology to perform tasks, solve problems and make decisions

• Utilise multiple research methods to analyse and interpret data (**quantitative and qualitative**)

• Experience with handling multiple forms of data
  – e.g. unstructured vs. structured; social media; multilingual and audiovisual

Gianluca Demartini
Programme team

http://www.sheffield.ac.uk/is/pgt/courses/data_science/dsteam

Paul Clough
(Programme Coordinator)

Gianluca Demartini
(Deputy Programme Coordinator)

Andrew Ball
(CEO, Peak Indicators)

Amy Balmain (Solutions Architect, Peak Indicators)
Target audience

• Home and overseas students with no requirement to have a background in Computer Science or Mathematics
  – Should appeal to a wider range of students
  – Differentiates this from many existing Data Science courses (i.e. more similar to iSchool programmes in the US)

• In the future we plan to target the professional market
  – e.g. through distance learning and/or part-time routes
Skills and other attributes

• Tools used by Data Scientists, such as **R and R Studio**, **SPSS**, **WEKA**, **Tableau**, **Oracle** and **Amazon AWS**

• Analyse **datasets of various sizes** to gain insights and extract information (and knowledge) following principled and structured methodologies

• Skills relevant to their potential future employment, together with an understanding of how to apply these **ethically**

• Acquire **research skills** relevant to their chosen field of work

• Develop **communication** and interpersonal **skills** that will complement their subject knowledge
Curriculum

• A candidate shall take (bold indicates new modules)
  – INF6027 Introduction to Data Science S1 15
  – INF6029 Data Analysis S1 15
  – INF6060 Information Retrieval: Search Engines and Digital Libraries S1 15
  – INF6028 Data Mining and Visualisation S2 15
  – INF6050 Database Design S2 15
  – INF6340 Research Methods and Dissertation Preparation S2 15

• (b) units to the value of forty five credits from the following
  – INF6110 Information Systems Modelling S1
  – INF6320 Information systems in Organisations S1 15
  – INF6022 Research Data Management S2 15
  – INF6040 Business Intelligence S2 15
  – INF6025 Information Governance S2 15
  – INF6024 Researching Social Media S2 15

• (c) INF6000 Dissertation S3 45

NB: datasets being compiled from data.gov.uk and from specific organisations as part of case studies (e.g. Peak Indicators)
Current status

• Full-time programme has started in September 2014.
  – 20+ students, ~50% home/overseas, 57% ladies 43% gentlemen
  – 58% never used a programming language
• Peak Indicators is closely involved in programme development (e.g. curriculum, advertising)
• External organisations contributed to the programme during the “Data Science industry day”
  – IBM, SY Police, Sheffield City Council
  – More to come: Barclays, PwC, UK Data Archive Centre, Findwise, Information Commissioner’s Office
• Marketing and curriculum development activities are now underway
  – 50+ applicants for Sept 2015
Highlights

• Both quantitative and qualitative methods for research and data analysis
  – Important in interpreting data accurately and gaining deep insights
• “Industry days”
• Gain hands-on experience with a range of tools (mainly open-source) that are commonly being used by Data Scientists
  – e.g. R, SPSS, Oracle, WEKA, Amazon AWS
  – Give students skills that are directly relevant to future employment and provide them a competitive advantage
• The programme is designed around the “type II” Data Scientist