KNIME as a Teaching Tool in Higher Education

Giuseppe Di Fatta
Associate Professor
G.DiFatta@reading.ac.uk
Data Science

- Data Science is an interdisciplinary subject that draws know-how and skills from a broad range of academic subject areas.
  - Computer Science
  - Statistics
  - Specific application domains

- More specifically in CS:
  - Advanced Computing paradigms, such as Cloud Computing, Parallel and Distributed Computing
  - Data Mining and Knowledge Discovery
  - Information retrieval and WWW
  - Visualisation

- And, most importantly, a data scientist should have an attitude to curiosity and discovery.
There will be a shortage of talent necessary for organizations to take advantage of big data. By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills.
“A New Breed”: the “data scientist” is a high-ranking professional with the training and curiosity to make discoveries in the world of big data.
There are 103,298 data scientists registered with Kaggle.

Connect with the world's top data scientists.

All you need is data and a question. Our data scientists will provide the answer.

“Efficient, fast, straight to the point. Minimal back & forth required.”
—Patrick Meier, iRevolution
A data scientist is “a hybrid computer scientist/software engineer/statistician”: demand/offer of specific HE programmes.
Data Science Degree Offer

- Degrees related to Data Science are on the rise.
- Mainly at PGT level.
KNIME at Reading

• UG level – BSc Computer Science
  – P2 module Java Programming
    · Eclipse/KNIME’s Metaprogramming example of Java Reflection
  – P3 module Data Mining Algorithms
    · Coursework based on KNIME
    – popular 1-year industrial placement

• Masters level
  – MSc Advanced Computer Science (NEW)
  – MRes Systems Engineering (Masters by Research)
    ➢ Opportunity for industrial projects and/or short industrial placement

• PhD in Computer Science
MRes Systems Engineering

- Masters by Research (MRes)
  - The MRes degree focuses on
    - a year-long research project (150 credits) and
    - three taught modules (30 credits) that are relevant to the research project.
  - The project can be on any topic within the diverse research interests of the School’s academic staff, with opportunities to carry out part of the project in an industrial or academic partner.
MSc Advanced Computer Science

- **MSc ACS: Master of Science in Advanced Computer Science**
  - 1-year taught postgraduate degree, intended for students who have already studied CS or a closely related subject as their first degree.
  - Teaching and learning methodologies: block-based lectures, assisted practical activities, seminars, self-directed research, student projects, student presentations.
  - Content is based on the research strengths of the School of Systems Engineering

- **Forging the next-generation computer scientists/data scientists**
  - For students who want to pursue a career in academia by continuing onto a PhD programme or in industrial R&D employment
  - For students looking for IT industry employment and own initiatives: it includes modules in Entrepreneurship, Social, Legal and Ethical aspects

http://www.reading.ac.uk/sse/pg-taught/sse-mscadvancedcomputerscience.aspx
Thames Valley

- Reading is at the heart of the Thames Valley, which is often referred to as the ‘Silicon Valley of Europe’. [http://www.thamesvalley.co.uk](http://www.thamesvalley.co.uk)
  - Home to 10 of the top 50 global organisations in the world and 13 of the world’s top 30 billion dollar brands.
  - a number of business sectors; a strong expertise in technologies and science.

- ICT companies involved in our UG placement programme:
### MSc ACS – Specifications

- **Start date:** October 2013

<table>
<thead>
<tr>
<th>Awarding Institution:</th>
<th>University of Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching Institution:</td>
<td>University of Reading</td>
</tr>
<tr>
<td>QAA subject Benchmarking group(s):</td>
<td>Computing</td>
</tr>
<tr>
<td>Faculty:</td>
<td>Faculty of Science</td>
</tr>
<tr>
<td>Programme length:</td>
<td>1 years</td>
</tr>
<tr>
<td>Programme Director:</td>
<td>Dr. Giuseppe Di Fatta (<a href="mailto:G.DiFatta@reading.ac.uk">G.DiFatta@reading.ac.uk</a>)</td>
</tr>
<tr>
<td>Programme Advisor:</td>
<td>Dr. Hong Wei (<a href="mailto:H.Wei@reading.ac.uk">H.Wei@reading.ac.uk</a>)</td>
</tr>
<tr>
<td>Expected Accreditation:</td>
<td>British Computer Society (BCS), the Chartered Institute for IT (*)</td>
</tr>
</tbody>
</table>

(*) A request for accreditation has been submitted to BCS: the application can be completed only at the end of the first intake and will be retrospectively valid.

- Full-time 12 months
- Part-time 24 months
- Flexible modular up to 5 years
  - *Fee calculated on a pro rata basis per credit (+10% admin charge)*
## MSc ACS – Modules

### Autumn Term

<table>
<thead>
<tr>
<th>Code</th>
<th>Module Title</th>
<th>Credits</th>
<th>Level</th>
<th>Compulsory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE4BC12</td>
<td>Brain-Computer Interface</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4TD12</td>
<td>Data Analytics and Mining</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4RS11</td>
<td>Research Studies</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4VR12</td>
<td>Virtual Reality</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SEMIP12</td>
<td>Image Processing</td>
<td>10</td>
<td>7</td>
<td>O</td>
</tr>
<tr>
<td>SE4AS12</td>
<td>Advanced Research Studies</td>
<td>10</td>
<td>7</td>
<td>O</td>
</tr>
<tr>
<td>MMM038</td>
<td>Practice of Entrepreneurship</td>
<td>20</td>
<td>7</td>
<td>O</td>
</tr>
<tr>
<td>SE3SL11</td>
<td>Social, Legal and Ethical Aspects in Engineering</td>
<td>10</td>
<td>6</td>
<td>O</td>
</tr>
</tbody>
</table>

### Spring Term

<table>
<thead>
<tr>
<th>Code</th>
<th>Module Title</th>
<th>Credits</th>
<th>Level</th>
<th>Compulsory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE4BD13</td>
<td>Big Data Analytics</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4CC12</td>
<td>Cloud Computing</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4GP12</td>
<td>GPU Computing</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4SI11</td>
<td>Swarm Intelligence &amp; Artificial Life</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4VI11</td>
<td>Visual Intelligence</td>
<td>10</td>
<td>7</td>
<td>C</td>
</tr>
<tr>
<td>SE4MD12</td>
<td>Manipulator Dynamics &amp; Haptics</td>
<td>10</td>
<td>7</td>
<td>O</td>
</tr>
<tr>
<td>SE4MI12</td>
<td>Medical Image and Signal Processing</td>
<td>10</td>
<td>7</td>
<td>O</td>
</tr>
<tr>
<td>SE4NN12</td>
<td>Advanced Neural Networks</td>
<td>10</td>
<td>7</td>
<td>O</td>
</tr>
</tbody>
</table>

### Summer Term

<table>
<thead>
<tr>
<th>Code</th>
<th>Module Title</th>
<th>Credits</th>
<th>Level</th>
<th>Compulsory/Optional</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEMPR12</td>
<td>MSc Project</td>
<td>60</td>
<td>7</td>
<td>C</td>
</tr>
</tbody>
</table>
Advanced Computing Paradigms

Parallel and Distributed Computing

- **GPU Computing**
  - Desktop Supercomputing

- **Cloud Computing**
  - IaaS, PaaS, SaaS
  - Big Data processing (MapReduce, Hadoop)
  - Large-scale Systems Architectures
Data Analytics and Mining

4\textsuperscript{th} scientific paradigm: data-intensive/data-driven scientific discovery

- Data-centric Computing
  - Predictive/Advanced Analytics, Data Mining
- Data Workflows
  - Data Access, Transformation, Analytics and Mining
  - Visualisation and Exploitation

G. Di Fatta
Big Data Analytics

- Big Data Analytics principles, techniques and challenges
  - Volume, Velocity, Variety and Veracity
  - Parallel data mining techniques and tools
  - Analysis of fast streaming real time data
  - Recommender systems
  - Unstructured data analysis

- Module “Data Analytics and Mining”:

- Module “Big Data Analytics”:

- Module “Cloud Computing”:

- Module “GPU Computing”:
Student Activities with KNIME

• Some examples of activities carried out by students as part of their studies:
  – Developing new KNIME nodes for specific data mining algorithms
    • Module coursework
    • Final Year Project
  – Developing a new KNIME node for a company during industrial placement year (covered by NDA).
  – Personal extra curriculum activities (e.g., Elance jobs)
K–Means Clustering

- Partitional Clustering approach
  - Clusters are disjoint subsets of the input data
  - Each cluster is associated with a centroid (centre of mass)
  - Iterative greedy approach

New KNIME nodes implemented by students (PhD, MSc, BSc):

- brute force K–Means with rich initialisation options
- optimised (faster) algorithm (deterministically equivalent to brute force K–Means)
  - KD–Tree K–Means
  - BSP–Tree K–Means
- Spherical K–Means (for text data)
Multidimensional Scaling (MDS)

- Multidimensional Scaling (MDS) algorithms produce a lower dimensional representation of high dimensional data such that the between-object distances are preserved as much as possible.
  - Often used for visualisation of high dimensional data

- MDS implementations in KNIME:
  - “PCA” is the classical linear approach (aka classical MDS, Torgerson scaling)
  - KNIME node “MDS” is based on Sammon mapping (a non-linear approach)

- New KNIME node implemented by an UG student:
  - Landmark MDS (LMDS) is based on the Nyström approximation
    - In classical MDS complexity of computing eigenvectors $O(m^3)$
    - LMDS uses a subset of $q$ points, called ‘landmarks’ ($q \ll m$)
Self–Organising Maps (SOM)

- Self–Organizing Maps (SOM) are a particular type of artificial neural networks.
- Unsupervised training produces a low-dimensional (typically 2D, a map) representation of the input data. Training is aimed at preserving the topological properties of the input space.

- New KNIME node implemented by an UG student (FYP):
  - SOM with UMatrix representation
Segmented Linear Regression

- New KNIME node for **Segmented Linear Regression**: segments separated by a breakpoint can be useful to quantify an abrupt change of a response function.

  - Application: degradation of tyre grip performance during F1 races.
Freelance Work in Data Science – Example

- A freelance work carried out by a PhD student with KNIME
  - work offered on Elance
  - Input: a set of aerial photographs of a site, taken in 36 different spectral bands.
  - Task: image segmentation to identify different environment and building areas.
  - Solution: feature generation (PCA) + clustering using KNIME
Conclusions

- A Data Scientist is “a hybrid computer scientist/software engineer/statistician“.
  - University degrees need to integrate more practical activities to frontal lectures, in general.
  - In particular, Data Science requires lots of hand-on activities with real-world data and case studies.

- Training in Data Science is an actual investment.

- KNIME provides a user-friendly interface and has demonstrated to be an excellent tool to introduce Data Analytics concepts and Data Mining algorithms
  - It supports the integration of practical activities in University courses.
Questions?

The University of Reading is ranked among the top 1% of the world’s universities, according to the Times Higher Education (THE) World University Rankings 2012.