What is a Data Scientist? (…Data Scientists in the Wild…)

Dr Liz Lyon,
Associate Director, Digital Curation Centre,
Director, UKOLN, University of Bath, UK
Dr Kenji Takeda,
Microsoft Research Connections

Microsoft eScience Workshop, Chicago, October 2012

This work is licensed under a Creative Commons Licence
Attribution-ShareAlike 2.0

UKOLN is supported by:

www.ukoln.ac.uk

A centre of expertise in digital information management
Running order…..

• What is data science?
• What does a data scientist do?

• Data scientist flavours
• Data scientist habitat
What is **Data Science**?
<table>
<thead>
<tr>
<th>Content/Tools</th>
<th>BI</th>
<th>Data Science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Support System Lineage</td>
<td>Statistical Science Lineage</td>
<td></td>
</tr>
<tr>
<td>Relational Database-Centric</td>
<td>Cloud-Centric, Massively Parallel, Other “Data Stores” (e.g. Cassandra, Hadoop)</td>
<td></td>
</tr>
<tr>
<td>Data Warehouse</td>
<td>Data Platform</td>
<td></td>
</tr>
<tr>
<td>Reporting/Dashboards Focus</td>
<td>Statistics/Experiments Focus</td>
<td></td>
</tr>
<tr>
<td>OLAP</td>
<td>Machine Learning</td>
<td></td>
</tr>
<tr>
<td>ETL</td>
<td>Data Munging/Conditioning</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>Visualization + Creative Design</td>
<td></td>
</tr>
<tr>
<td>Big Proprietary + Open Source</td>
<td>Open Source + Small Proprietary</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Business</th>
<th>IT-Owned</th>
<th>Analytics-Owned</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technology/Business</td>
<td>Mathematics/Science</td>
<td></td>
</tr>
<tr>
<td>Performance Management</td>
<td>Data Products</td>
<td></td>
</tr>
<tr>
<td>Methodical</td>
<td>Inspirational</td>
<td></td>
</tr>
<tr>
<td>Middle-Aged</td>
<td>Adolescent</td>
<td></td>
</tr>
<tr>
<td>Division of Labor</td>
<td>Jack of All Trades</td>
<td></td>
</tr>
<tr>
<td>Teams</td>
<td>One-Offs</td>
<td></td>
</tr>
<tr>
<td>Short-to-Medium-Sized Projects</td>
<td>Quicker Hits</td>
<td></td>
</tr>
<tr>
<td>Precision</td>
<td>Speed</td>
<td></td>
</tr>
<tr>
<td>More Governance</td>
<td>Less Governance</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data</th>
<th>Complete Data</th>
<th>Missing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quality Centric</td>
<td>Quantity Centric</td>
<td></td>
</tr>
<tr>
<td>Absolute</td>
<td>Approximate</td>
<td></td>
</tr>
<tr>
<td>More Internal Data</td>
<td>More External Data</td>
<td></td>
</tr>
<tr>
<td>Structured Data</td>
<td>Structured + Unstructured Data</td>
<td></td>
</tr>
<tr>
<td>Small-Medium-Large Data</td>
<td>Big Data</td>
<td></td>
</tr>
</tbody>
</table>
What is Data Science? 2

- Science which is data intensive, data driven
- Data as a commodity
- Data as infrastructure
- Data as research substrate
- Data as a science utility
- Data workflows, data tools, data publications
- “The Fourth Paradigm in Practice”
Data : from Big to Broad
(Jim Hendler)

4th context: Broad Data
- The huge amount of freely available, but widely varied, Open Data on the World Wide Web (Structured and Semi-structured)
  - Example: The extended Facebook OGP graph (the part outside Facebook’s datasets)
  - Example: The growing linked open data cloud of freely available RDF linked data
  - Example: More than 710,000 datasets that are available on the Web free from governments around the world
Implications of “Big Data” and data science for organisations in all sectors

Predicts a shortage of 190,000 data scientists by 2019

http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation
Big Data Needs Data Scientists, Or Quants, Or Excel Jockeys

Data Scientist = Rock Star, Really?
The Term “Data Scientist” is Still New

“Data Scientist” Jobs = Near Zero Until 2010

Is There a Shortage of Data Scientists?
What does a data scientist do? 1

- Understands problems tackled with a data-centric approach
- Understands data-centric analysis
- Tackles problems using `<Data+Analytics>` lens
- Data mashing, munging, manipulation
  - Data analytics for business advantage
    - Data jujitsu
  - “turns data into product”
What does a data scientist do?

- Creates visualisations of complex data
- Produces the Guardian newspaper Data Blog
- Data journalist variant
- “creates stories from data”
What does a data scientist do?

- Creates data management plans
- Uses standards for data description, schema
- Uses persistent identifiers for datasets
- Manages data access through embargos
- Applies appropriate data licenses
- Facilitates data citation
- “gets credit for their data”
What does a data scientist do?

- Acts as a data steward
- Deposit data in an appropriate repository
- Curate, annotate, cleanse, redact
- Facilitates data preservation & archiving for long term use
- Data forensics
- Data archaeology
- “adds value to data”
• Leadership & co-ordination
• Strategy and planning
• Policy
• Legal and ethical (FoI, Data Protection)
• Advocacy (data informatics)
• Data repositories
• Data storage
• Data analysis
• Data visualisation
• Data mining
• Data modelling
• Data licensing
• Training....
Data Scientist roles

- **data engineer** - focus on software development, coding, programming, tools
- **data analyst** – focus on business/scientific analytics and statistics e.g. R, SAS, Excel to support researchers and modellers, business
- **data librarian** – focus on advocacy, research data management / informatics in a university / institute
- **data steward** – focus on long term digital preservation, repositories, archives, data centres
- **data journalist** – focus on telling stories and news
Data engineer

- Focus on software development, coding, programming, tools
- Customises methods and tools for end-users
- Code-focussed
  - R
  - SAS
  - SQL/NoSQL
  - Hadoop
  - F#

http://preview.tryfsharp.org
Institutional data scientist

- Co-ordination and Collaboration
  - Liaison / subject librarians
  - Repository manager
  - IT/Computing Services
  - Research Support & Development Office
  - Doctoral Training Centres
  - Researchers
- Advocacy
- Training

Liz Lyon, Informatics Transform, IJDC Current Issue, 2012
Understanding the data science habitat: PI, institution, funder

Community Capability Model Framework

http://communitymodel.sharepoint.com/
CCMF 8 Capability Factors

- HUMAN
- Skills & Training
- Openness
- Technical Infrastructure
- ENVIRONMENTAL
- Environmental
- Legal & Ethical
- Economic & Business
- Academic
- Collaboration
- Common Practices

Data-Intensive Research
CCMF supporting data science

- Intelligence-gathering
- Decision-making
- Planning
- Investment
- Capacity
- Capability
- Knowledge transfer
CCMF Team

- UKOLN: Liz Lyon, Alex Ball, Monica Duke, Michael Day, Manjula Patel, Michelle Smith

- Microsoft: Kenji Takeda, Alex Wade

CCMF White Paper

http://communitymodel.sharepoint.com/Documents/CCMDIRWhitepaper-v1-0.pdf
Infrastructure, Intelligence, Innovation: driving the Data Science agenda
8th International Digital Curation Conference, Amsterdam, 14-16 January 2013
Thank you.

CCMF Resources download from
http://communitymodel.sharepoint.com/Pages/default.aspx

Slides at
http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html

Informatics Transform paper at