The Library (Big) Data scientist

IFLA/ALA webinar:
“Big Data: new roles and opportunities for new librarians”

June 15th 2016

IFLA Big Data Special Interest Group (SIG)
Wouter Klapwijk, Stellenbosch University, SIG convenor
bigdata@infoserv.inist.fr
IFLA Big Data SIG

• Proposed at WLIC 2014, Lyon
• Petitioned for at WLIC 2015, Cape Town
• Endorsed by the IFLA Professional Committee in December 2015
• SIG sponsor: IT Section

• Objectives:
  1. Provide a **focus point** for developing ideas regarding Big Data as it affects libraries
  2. Provide a platform within IFLA to assess and develop the **avenues of response** from IFLA to this developing area
Deconstructing “Big Data” and “data science”

This talk is based on a number of everyday questions:

1. What does “data science” mean?
   ✧ is it only happening in Tech places like Facebook and Google?

2. What are Data Scientists?
   ✧ can Librarians also be Data Scientists?

3. Is data science the science of Big Data?
   ✧ what is the relationship between Big Data and data science?

4. Exactly what is “Big Data” anyway?
   ✧ just how big is Big? or is Big relative? is library data Big?
A set of fundamental \textit{principles} that guide the extraction of \textit{knowledge} from data

The “civil engineering of data”: turning \textit{data} into \textit{data products}

\textbf{Goal of data science:}

\begin{itemize}
  \item to \textbf{improve decision-making}, for the betterment of organizations and society at large
\end{itemize}
“data mining”

✓ helps accomplish data science goals via technologies that incorporate data science principles

✓ but ... its techniques are much more extensive than the set of principles comprising data science

“data warehousing”

✓ a facilitating technology for “data mining”

✓ but ... not always included as part of “data mining”
Relation to other computing concepts

“data processing”
✓ is more general than data science
✓ there is processing involved in all aspects of computing

“Big Data technologies”
✓ are often used for data processing in support of data mining techniques
✓ ... and other data science activities
Science or Craft?

- The term *data science* has existed for over 30 years – it is a field onto itself.

- Foundation rests in century old practices of Statistics, Mathematics, and since mid-20th century, also Computer Sciences.

- It is not just a rebranding of Statistics and Machine Learning in the context of the Tech industry.

- Much of the field development is happening in Industry, and not in Academia.
<table>
<thead>
<tr>
<th>Domain</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internet</td>
<td>Recommendation systems (Amazon = books; Facebook = friends)</td>
</tr>
<tr>
<td>Finance</td>
<td>Credit ratings</td>
</tr>
<tr>
<td>Education</td>
<td>Personalized learning and assessment</td>
</tr>
<tr>
<td>Government</td>
<td>Policies based on data</td>
</tr>
</tbody>
</table>
Practicing data science

What do Data Scientist do?
The Data Scientist

Two aspects to consider:

1. understand what they DO in business

2. understand which SKILLS they must possess
What do they DO?

1. They ask questions
   - Probe, being curious

2. They (try to) solve problems
   - Analytical thinking, making new discoveries

3. They cultivate (new) soft skills
   - Communicating and visualizing data
What do they DO?

1. They ask questions
   - Probe, being curious

2. They (try to) solve problems
   - Analytical thinking, making new discoveries

3. They cultivate (new) soft skills
   - Communicating and visualizing data

How much of the above do you as librarian do?
<table>
<thead>
<tr>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is it in our pedigree to continuously ask questions?</td>
</tr>
<tr>
<td>Do we have the taste and mindset for analytical thinking?</td>
</tr>
<tr>
<td>Do we only do ad hoc analysis, or do we prefer an ongoing conversation with data?</td>
</tr>
<tr>
<td>Is there enough of a Business Analyst or Social Scientist in us?</td>
</tr>
</tbody>
</table>
Which SKILLS do they need?

- Hard skills
- Domain-specific skills
- Soft skills

Data Scientist
Which SKILLS do they need?

Data Scientist

- **Hard skills**
  - Linear algebra
  - Statistics
  - Artificial Intelligence
  - Machine Learning

- **Domain-specific skills**
  - Understand the business, e.g. librarianship

- **Soft skills**

- **Communication**
Data Scientists are team members

Different skills are embedded across **multi-disciplinary** team members
It is important to **understand** data science even if you never intend to do it yourself.
Practicing data science

What does the craft of data science look like?
3 disciplinary areas

SOURCES
- DATA

ANALYTICS
- COLLECT
- CLEAN
- INTEGRATE
- PROCESS

VISUALIZATION
- COMMUNICATE
Each disciplinary area requires different skills
Database (1960 - )
- First integrated datastore (Bachman), 1963
- Relational data model (Codd), 1970
- SQL (Boyce & Chamberlain), 1970+

Data Warehouse (1975 - )
- First commercial RDBMS (Oracle), 1979
- DB2 (IBM), 1983
- First KDD workshop, 1989
- First KDD data mining conference (Fayyaad, Shapiro), 1995

Big Data (2005 - )
- NoSQL (Evans), 2009
Incomplete Data Taxonomy: some data are neither just big nor just small
Small data

- The term denotes the *opposite* of Big Data
- Data usually housed in databases and data warehouses
- Usually structured, qualitative and indexical in nature
- **Examples**: Library data, Research Data (RDM)
- Research data = primary data
SOURCES

Big data

• Datasets that are too large for traditional data processing and storage systems (3V’s, 4V’s, 5V’s)

• Classified into 3 classes of “datafication”:
  1. Directed data (e.g. surveillance data)
  2. Automated data (e.g. device generated data)
  3. Volunteered data (e.g. social networks data)
ANALYTICS

Database (1960 - )
- First integrated datastore (Bachman), 1963
- Relational data model (Codd), 1970
- SQL (Boyce & Chamerlain), 1970+

Data Warehouse (1975 - )
- First commercial RDBMS (Oracle), 1979
- DB2 (IBM), 1983
- First KDD workshop, 1989
- First KDD data mining conference (Fayyad, Shapiro), 1995

Big Data (2005 - )
- NoSQL (Evans), 2009

Business Intelligence (BI)

Data Science
ANALYTICS

There are 4 broad classes of analytics (often used in combination):

1. Data mining and pattern recognition
   - AI – Machine Learning – Data Mining

2. Data visualization and visual analytics
   - App development

3. Statistical analysis
   - Statistical techniques and principles (regression, etc.)

4. Prediction, simulation, and optimization
   - Algorithms
Practicing data science

In Libraries
1. Data at Scale

Value and Insight can be extracted from small data by scaling them up into larger datasets, for reuse through digital data infrastructures.
2. Analyzing Exhaust data

**Exhaust data** = produced as a by-product of the main function of a device or system

Most exhaust data is *transient* in nature – it is never examined and simply discarded!

**Example**: log of a self-checkout unit
Example of analyzing Exhaust data

Structured and Unstructured data

- VISITS
- PATRONS
- LOANS
- LOCATIONS
- DIGITIZED BOOKS

Actionable Insights

- Better forecasts for future library planning
- Better usage of systems and resources
- Productivity gain with better decision-making
Examples of Analysis and Visualization

Library analytics toolkit – Harvard University: 
https://osc.hul.harvard.edu/liblab/projects/library-analytics-toolkit

Text analytics – Google Books Ngram Viewer: 
https://books.google.com/ngrams

Open Source implementation – Bookworm: 
http://bookworm.culturomics.org
In summary

The fundamental principle of data science is that data, and the capability to extract useful knowledge from it, should be regarded as a key strategic asset.

Libraries must learn to start thinking data-analytically. Do we only use gut and intuition, or also data and rigor, in our decision-making?
In summary

You can apply the same principles, tools and techniques for small data than you would for big data

“...the tools of data science are as appropriate for gigabyte as they are for petabyte scale datasets...”

(https://datascience.berkeley.edu/about/what-is-data-science/)
In summary

Challenges for librarians:

- There is a shortage of Big Data talent
- The Big Data SIG is attempting to understand and frame Big Data problems

Opportunities for librarians:

- Grow your data analytical skills
- Attend online courses: Khan Academy, Coursera, Software Carpentry, digital books
- There are free software tools: R, (SQL Server 2016 includes R), Python, app visualization tools
Thank you

wklap@sun.ac.za