

# *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](http://www.ukoln.ac.uk)

A centre of expertise in digital information management



THE SUNDAY TIMES  
**UNIVERSITY OF THE YEAR** 2011-12

# Running order.....

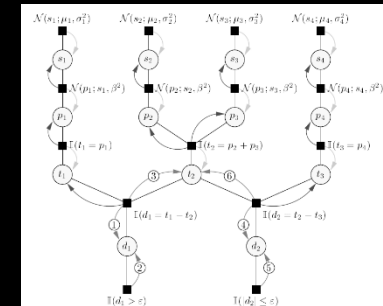
- What is data science?
- What does a data scientist do?

- Data scientist flavours
- Data scientist habitat



# What is *Data Science*?

A screenshot of a Bing search results page for the query "data scientist". The page shows 181,000,000 results. The top results include an advertisement for Altoros, a link to Indeed.com for "Data Scientist Jobs", and a news article from AOL titled "Data Scientist: The Hottest Job You Haven't Heard Of". The article snippet states: "Aug 10, 2011 · Data scientists are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities".



A screenshot of Microsoft Excel 2010 showing a Sales Analysis dashboard. The dashboard includes several charts and filters. The charts are: "PurchaseCount by Rating" (bar chart), "PurchaseCount by Hour" (line chart), "PurchaseCount by Country" (pie chart), and "PurchaseCount by Genre" (bar chart). The filters include Year Num, Quart, MonthNumOfYear, Category, Rating, Country, and Distributor. The dashboard is titled "Sales Analysis".

**Work with massive amounts of data**

**Self-service analysis delivered thru Excel 2010**



**information  
management**

2 part piece  
on BI &  
Data  
Science  
by  
Steve Miller  
2012

	<b>BI</b>	<b>Data Science</b>
<b>Content/Tools</b>	Decision Support System Lineage	Statistical Science Lineage
	Relational Database-Centric	Cloud-Centric, Massively Parallel, Other "Data Stores" (e.g. Cassandra, Hadoop)
	Data Warehouse	Data Platform
	Reporting/Dashboards Focus	Statistics/Experiments Focus
	OLAP	Machine Learning
	ETL	Data Munging/Conditioning
	Visualization	Visualization+Creative Design
	Big Proprietary + Open Source	Open Source + Small Proprietary
<b>Business</b>	IT-Owned	Analytics-Owned
	Technology/Business	Mathematics/Science
	Performance Management	Data Products
	Methodical	Inspirational
	Middle-Aged	Adolescent
	Division of Labor	Jack of All Trades
	Teams	One-Offs
	Short-to-Medium Sized Projects	Quicker Hits
	Precision	Speed
	More Governance	Less Governance
<b>Data</b>	Complete Data	Missing Data
	Quality Centric	Quantity Centric
	Absolute	Approximate
	More Internal Data	More External Data
	Structured Data	Structured + Unstructured Data
	Small-Medium-Large Data	Big Data



# What is Data *Science*? 2

- Science which is data intensive, data driven
- Science at web-scale
- Data as a commodity
- Data as infrastructure
- Data as research substrate
- Data as a science utility
- Data workflows, data tools, data publications



# Data : from Big to Broad (Jim Hendler)



BROAD data

Tetherless World Constellation

- 4<sup>th</sup> 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

McKinsey Global Institute



May 2011

Big data: The next frontier  
for innovation, competition,  
and productivity



Implications of  
“Big Data” and  
data science for  
organisations in  
all sectors

Predicts a  
shortage of  
190,000  
data scientists  
by 2019

# Forbes

## Big Data Needs Data Scientists, Or Quants, Or Excel Jockeys

### Data Scientist = Rock Star, Really?

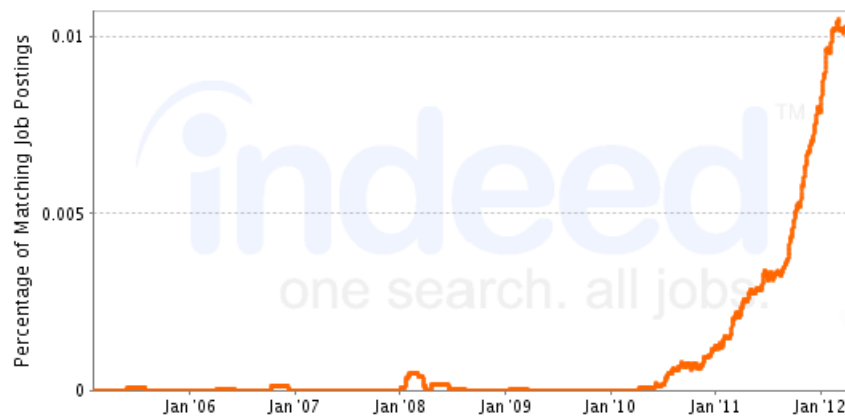
The Term "Data Scientist" is Still New

**CMS WIRE**

"Data Scientist" Jobs = Near Zero Until  
2010

Job Trends from Indeed.com

— "data scientist"

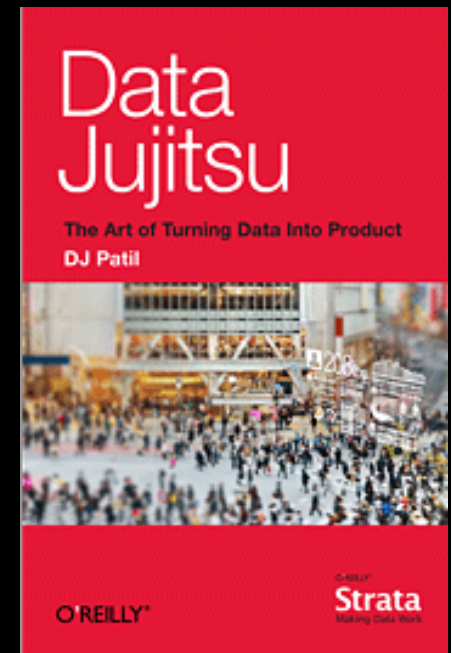


Is There a Shortage of Data Scientists?



# 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? 2

- Creates visualisations of complex data
- Produces the Guardian newspaper Data Blog
- Data journalist variant
- ***“creates stories from data”***

The screenshot shows the Guardian Data Blog interface. At the top, the Guardian logo is visible, along with navigation links for News, Sport, Comment, Culture, Business, Money, Life & style, Travel, and Environment. The 'Datablog' section is highlighted. The main article title is 'Anyone can do it. Data journalism is the new punk' by Simon Rogers. The article text begins with 'Can anyone be a data journalist? Simon Rogers on what we can learn from a 1977 diagram'. A social sharing bar shows 348 shares, 580 tweets, and 47 likes. A link to 'Another view: What data can and cannot do by Jonathan Gray' is provided. The article features a hand-drawn diagram on a yellowed piece of paper. The diagram consists of a grid of vertical lines with horizontal lines at the top and bottom. An arrow points to the third vertical line from the left, with the text 'This IS A THIRD' written next to it. Below the grid, the text 'NOW FORM A BAND' is written in a bold, underlined font. The diagram is labeled 'Page two of Sideburns, January 1977'. A small video player is embedded below the diagram, showing the same hand-drawn diagram. The article is posted by Simon Rogers on Thursday, 24 May 2012, at 13:00 BST. The article has 8 comments. The page also includes a 'Media' section with a link to 'Data journalism - Open journalism'.

# What does a data scientist do? 3

- 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”***

**DIMP***online*  
The  D | C | C Data Management Planning Tool



because good research needs good data

# What does a data scientist do? 4

- 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”*

The screenshot shows the UK Data Archive website. At the top, it states "THE UK'S LARGEST COLLECTION OF DIGITAL RESEARCH DATA IN THE SOCIAL SCIENCES AND HUMANITIES". The navigation menu includes: HOME, ABOUT US, CREATE & MANAGE DATA, DEPOSIT DATA, HOW WE CURATE DATA, FIND DATA, NEWS & EVENTS. There is a search bar with the text "SEARCH OUR SITE" and a "GO" button. The main content area features a large banner for "ANNOUNCING THE UK DATA SERVICE" with a microphone image and a "READ MORE" button. Below this are sections for "DEPOSITING YOUR DATA", "FINDING DATA TO USE", "OUR DATA IN USE", and "OUR SERVICES". On the right side, there are three boxes: "FIRST TIME HERE? HELPFUL INFORMATION", "A QUICK GUIDE TO THE ARCHIVE" (with a "10 of 10" indicator), and "WHO GIVES US DATA?". At the bottom right, there is a "LATEST NEWS & EVENTS" section with two items: "A new way to search for data" and "Staff vacancies Census Microdata".



- Leadership & co-ordination
- Data Strategy, Roadmaps, Planning
- Data Policy
- Legal and ethical (Fol, Data Protection)
- Advocacy (data informatics)
- Data repositories
- Data storage
- Data analysis
- Data visualisation
- Data mining
- Data modelling
- Data licensing
- Training....



# Family of 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



Jer Thorp: Hope / Crisis, NYT Word Frequency

New York Times Data Artist in Residence, Jer Thorp Joins Stellar Cast of Speakers at TEDxVancouver 2011

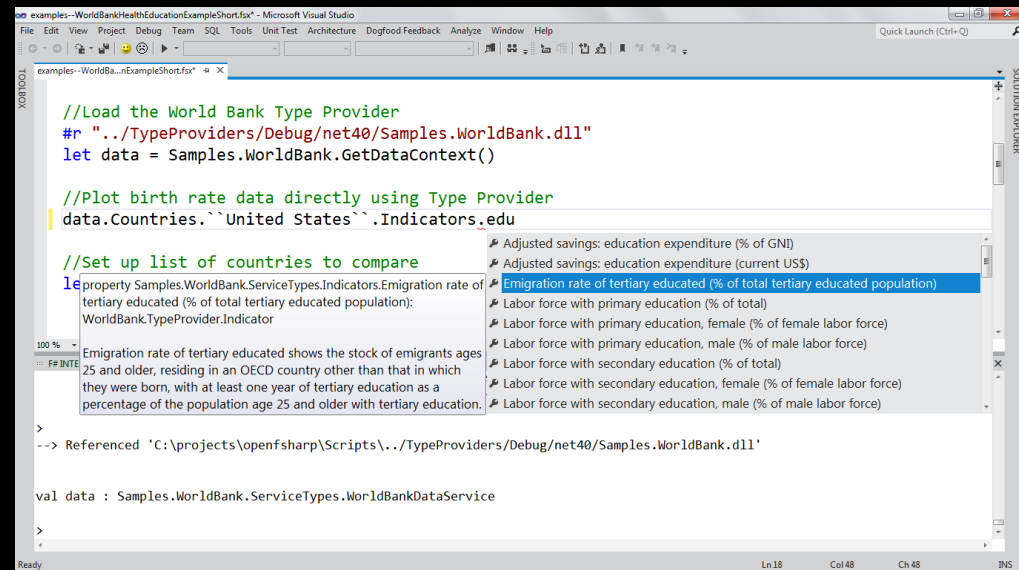
Posted by TEDxVancouver Team on October 17th, 2011 · No Comments

The New York Times



# Data engineer

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



```
examples--WorldBankHealthEducationExampleShort.fsx* - Microsoft Visual Studio
File Edit View Project Debug Team SQL Tools Unit Test Architecture Dogfood Feedback Analyze Window Help
examples--WorldBankHealthEducationExampleShort.fsx*
//Load the World Bank Type Provider
#r "../TypeProviders/Debug/net40/Samples.WorldBank.d11"
let data = Samples.WorldBank.GetDataContext()

//Plot birth rate data directly using Type Provider
data.Countries.`United States`.Indicators.edu

//Set up list of countries to compare
let property Samples.WorldBank.ServiceTypes.Indicators.Emigration rate of
tertiary educated (% of total tertiary educated population):
WorldBank.TypeProvider.Indicator
Emigration rate of tertiary educated shows the stock of emigrants ages
25 and older, residing in an OECD country other than that in which
they were born, with at least one year of tertiary education as a
percentage of the population age 25 and older with tertiary education.
Adjusted savings: education expenditure (% of GNI)
Adjusted savings: education expenditure (current US$)
Emigration rate of tertiary educated (% of total tertiary educated population)
Labor force with primary education (% of total)
Labor force with primary education, female (% of female labor force)
Labor force with primary education, male (% of male labor force)
Labor force with secondary education (% of total)
Labor force with secondary education, female (% of female labor force)
Labor force with secondary education, male (% of male labor force)

--> Referenced 'C:\projects\openfsharp\Scripts\..\TypeProviders/Debug/net40/Samples.WorldBank.d11'

val data : Samples.WorldBank.ServiceTypes.WorldBankDataService
```

<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**



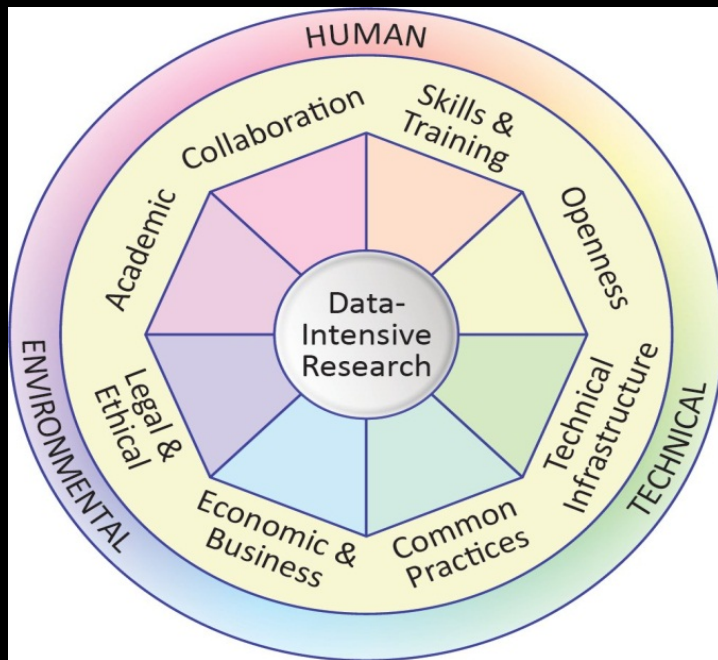
Research360

Managing data across the institutional research lifecycle

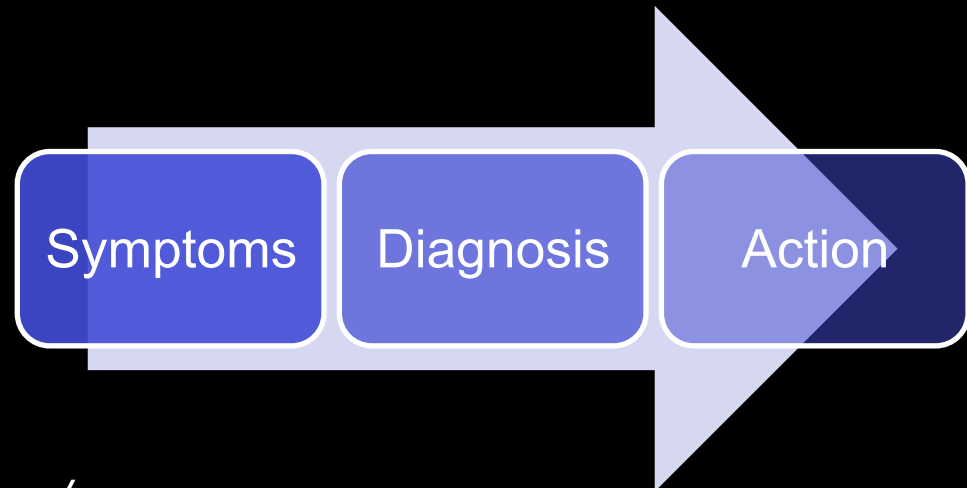
*Liz Lyon, Informatics Transform,  
IJDC Current Issue, 2012*



# Understanding the data science habitat : PI, institution, funder

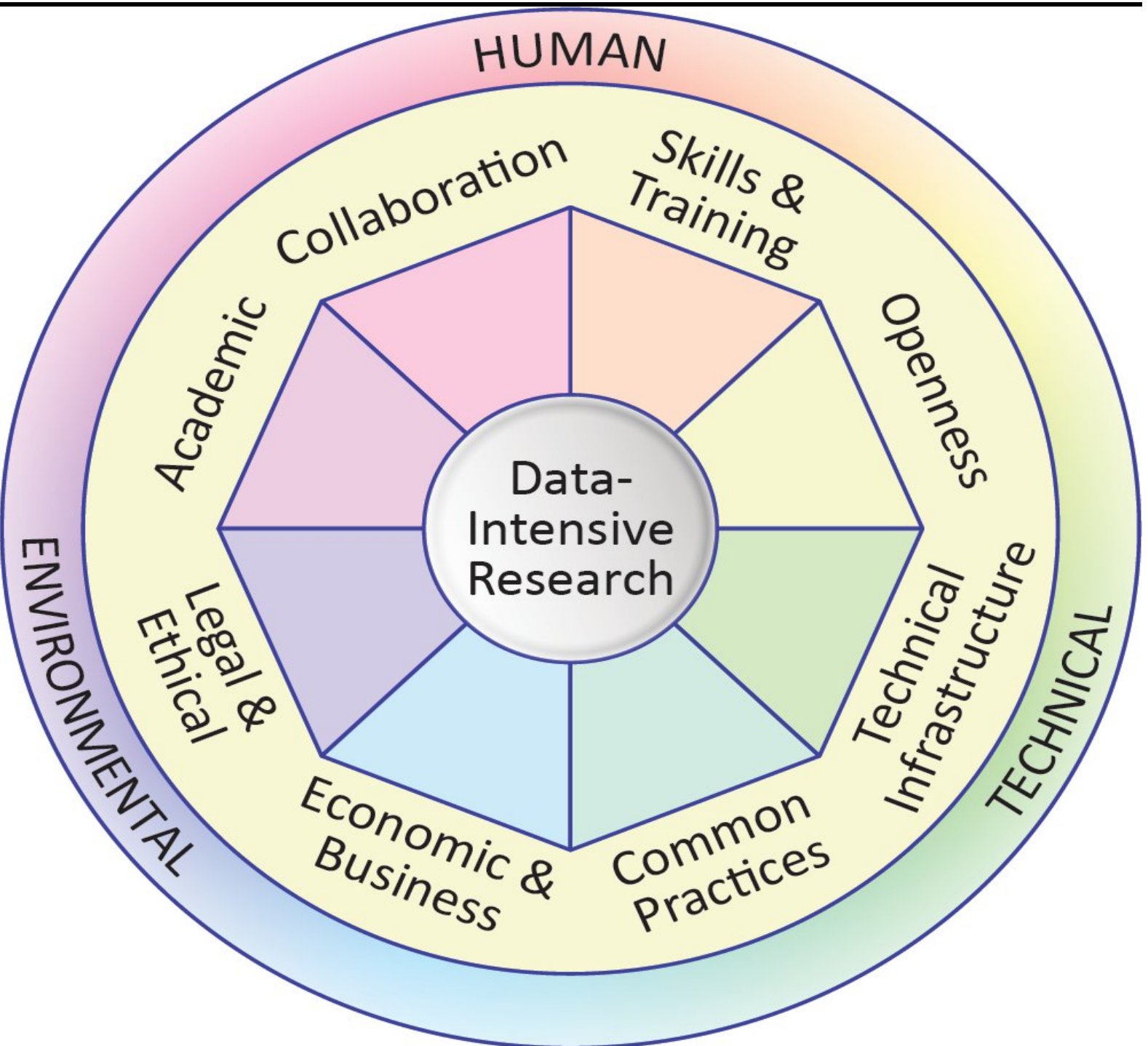


## Community Capability Model Framework



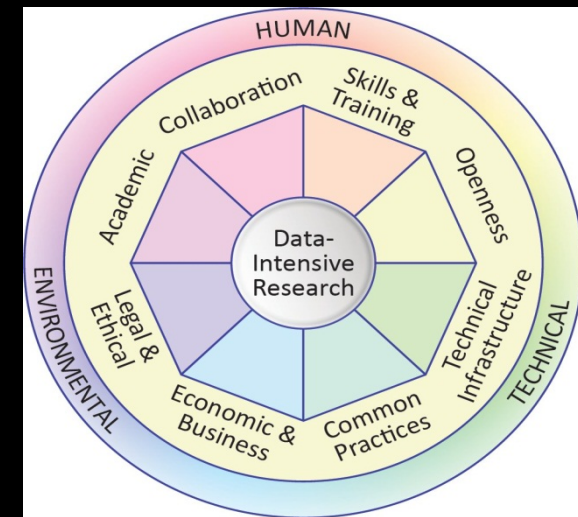
<http://communitymodel.sharepoint.com/>

# CCMF 8 Capability Factors

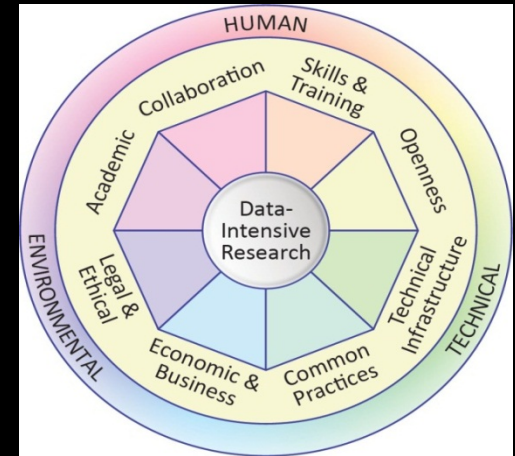


# CCMF supporting data science

- 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*

8<sup>th</sup> International Digital Curation Conference,  
Amsterdam, 14-16 January 2013

# Thank you.

*CCMF Resources download from*  
<http://communitymodel.sharepoint.com>

*Slides at*

<http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html>

*Informatics Transform paper at*

<http://www.ijdc.net/index.php/ijdc/article/view/210/279>

