

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



www.ukoln.ac.uk

A centre of expertise in digital information management

UKOLN is supported by:



Running order.....

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

- Data scientist flavours
- Data scientist habitat



What is *Data Science*?

181,000,000 RESULTS Any time ▾

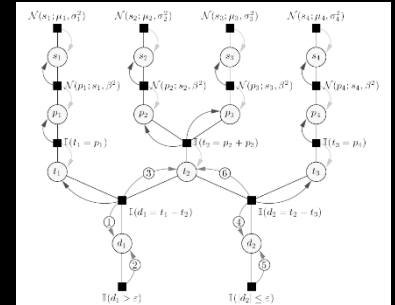
Data Analytics | Altoros - Expert Data Analytics: Hadoop, etc. Ad
www.althoros.com/Data_Analytics
Sr / Mid Level. 10+ Years in IT!

Data Scientist: The Hottest Job You Haven't Heard Of - Careers ...
jobs.aol.com/articles/2011/08/10/data-scientist-the-hottest-job...
Aug 10, 2011 · Data scientists are an integral part of competitive intelligence, a newly emerging field that encompasses a number of activities

The Data Scientist role is a role of the future!
www.datascientist.com
The Data Scientist is transitioning today

The Data Scientist
www.thedata...
I recently spent...
App Data and A...

Related search:
Data Scientist Se...
Data Scientist Fo...
Data Scientist Sa...
Career Advice
www.quora.com/...



Sales Analysis

PurchaseCount by Rating

Year Num	Chart L	Month Num of Year	Category
2006 2007	1 2	1 2 3	Action
2008 2005	3 4	4 5 6	Action/Adventure
2009	7 8 9	7 8 9	Adventure

PurchaseCount by Hour

PurchaseCount by Country

PurchaseCount by Genre

Work with massive amounts of data

Self-service analysis delivered thru Excel 2010

XBOX 360

BETTER WITH KINECT™ SENSOR

FORZA MOTORSPORT 4

3

TopGear

Microsoft game studios



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

	BI	Data Science
Content/Tools	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
Business	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
Data	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

Data : from Big to Broad (Jim Hendler)



BROAD data

Tetherless World Constellation

- 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

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

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

Data Scientist = Rock Star, Really?

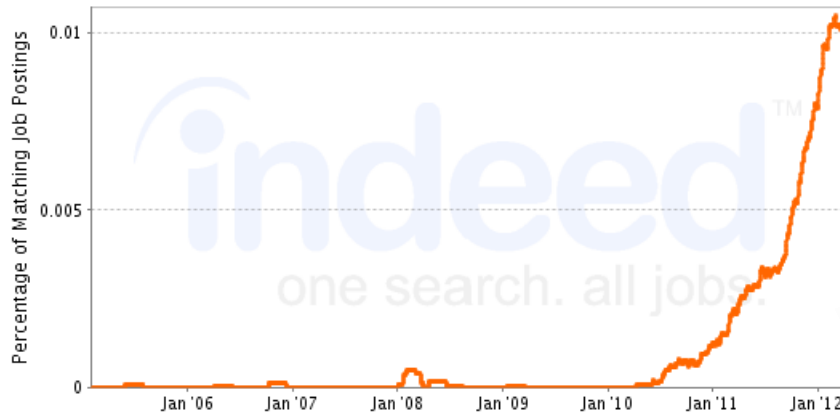
The Term “Data Scientist” is Still New

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 on the left, and a search bar is on the right. Below the logo is a navigation menu with links for News, Sport, Comment, Culture, Business, Money, Life & style, Travel, and Environment. The 'Datablog' link is highlighted. The main heading is 'DATA BLOG' in large orange letters, with the tagline 'Facts are sacred' below it. The 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'. To the right of the article are social sharing buttons for Facebook (348), Twitter (580), and Email. Below the article is a link to 'Another view: What data can and cannot do by Jonathan Gray'. The main image is a photograph of a handwritten diagram on a piece of paper. The diagram consists of a grid of vertical lines and horizontal lines. Above the grid is a horizontal line with five vertical tick marks. To the right of the grid is a circled 'G'. The text 'This IS A THIRD' is written to the left of the grid, with an arrow pointing to the top row. At the bottom of the grid, the text 'NOW FORM A BAND' is written. Below the image is the caption 'Page two of Sideburns, January 1977'. At the bottom of the page, there are two red circles and the text 'This is a chord... this is another... this is a third. NOW FORM A BAND'. On the right side of the page, there is a profile picture of Simon Rogers, his name, the date 'Thursday 24 May 2012 13:00 BST', the website 'guardian.co.uk', and a 'Jump to comments (8)' link. Below the profile picture are social media icons for Facebook, Twitter, and Google+, and a link to 'Article history'. At the bottom right, there is 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”***



D | C | C

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, there is a navigation menu with links for HOME, ABOUT US, CREATE & MANAGE DATA, DEPOSIT DATA, HOW WE CURATE DATA, and FIND DATA. Below the navigation is a search bar and a section titled 'ANNOUNCING THE UK DATA SERVICE' with a sub-headline 'The new UK Data Service has been funded with a £17 million investment over five years.' To the right of this section is a 'FIRST TIME HERE? HELPFUL INFORMATION' box. Below the main announcement are sections for 'DEPOSITING YOUR DATA' and 'FINDING DATA TO USE'. At the bottom, there are sections for 'OUR DATA IN USE' and 'OUR SERVICES'. The website is clean and professional, with a focus on providing information about data services.

- Leadership & co-ordination
- Strategy and planning
- 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....



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



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

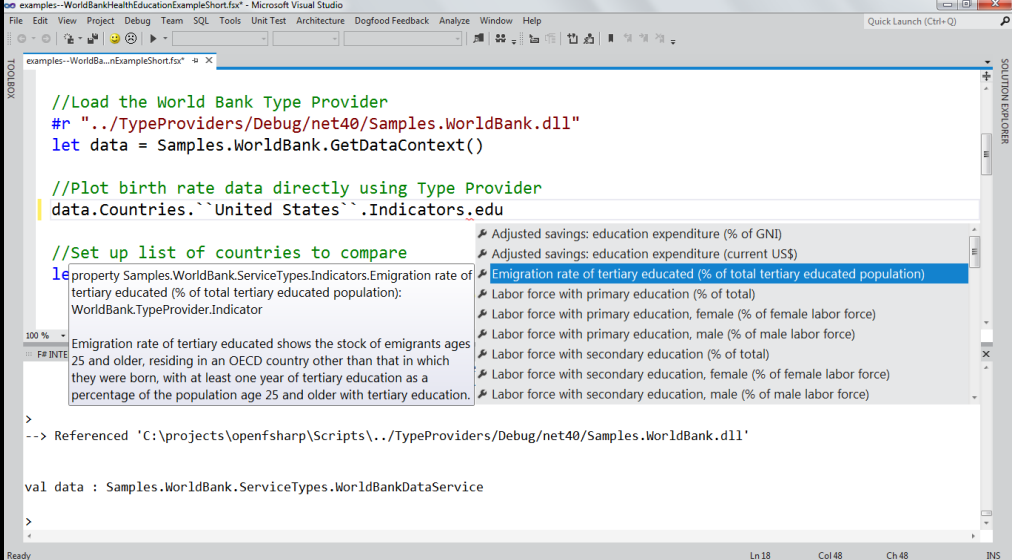
Posted by TEDxVancouver Team on October 17th, 2011 · No Comments

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 Quick Launch (Ctrl+Q)
examples--WorldBankHealthEducationExampleShort.fsx
//Load the World Bank Type Provider
#r "../TypeProviders/Debug/net40/Samples.WorldBank.dll"
let data = Samples.WorldBank.GetDataContext()

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

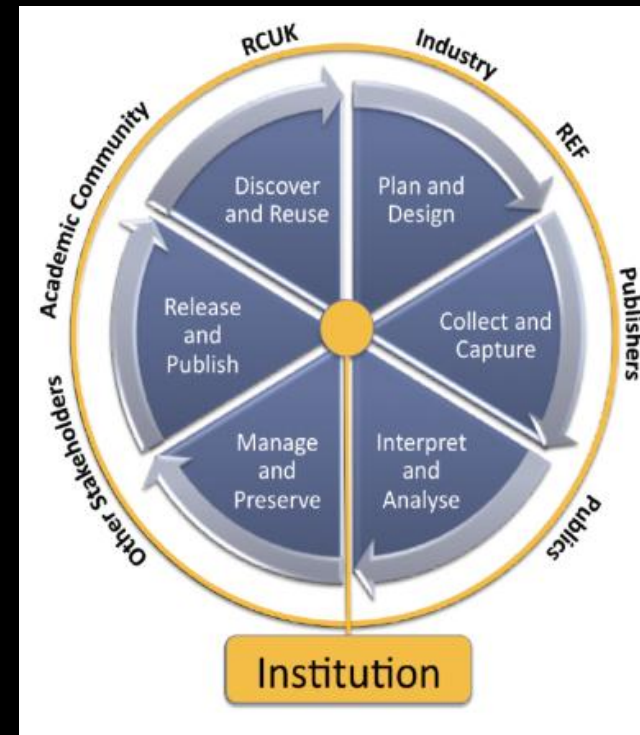
//Set up list of countries to compare
let property Samples.WorldBank.ServiceTypes.Indicators.Emigration rate of
tertiary educated (% of total tertiary educated population):
WorldBankTypeProvider.Indicator
Adjusted savings: education expenditure (% of GNI)
Adjusted savings: education expenditure (current 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)
100 % --
FR INTE 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.
>
--> Referenced 'C:\projects\openfsharp\Scripts\..\TypeProviders/Debug/net40/Samples.WorldBank.dll'

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**

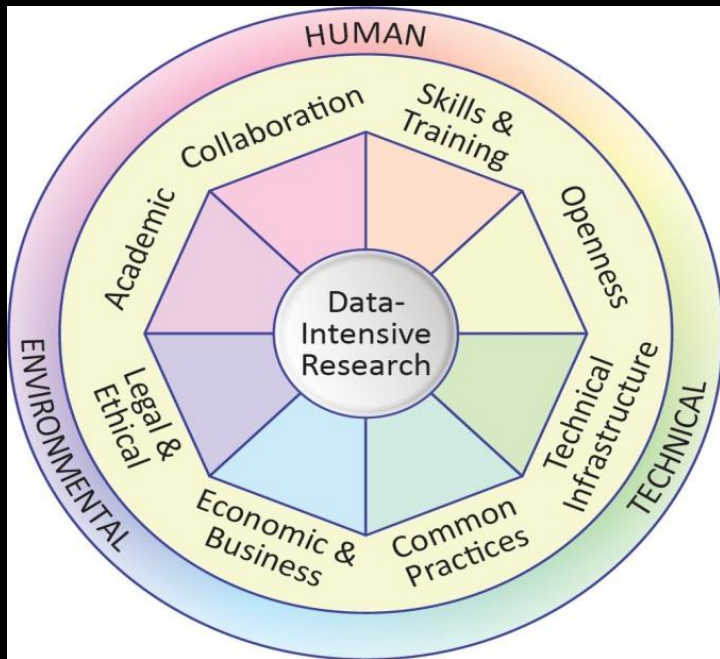


Liz Lyon, Informatics Transform, IJDC Current Issue, 2012

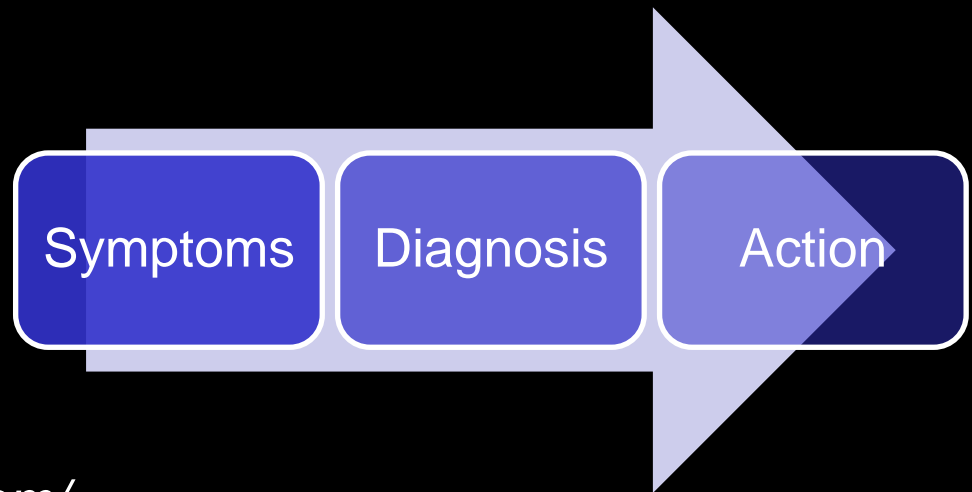
Research360

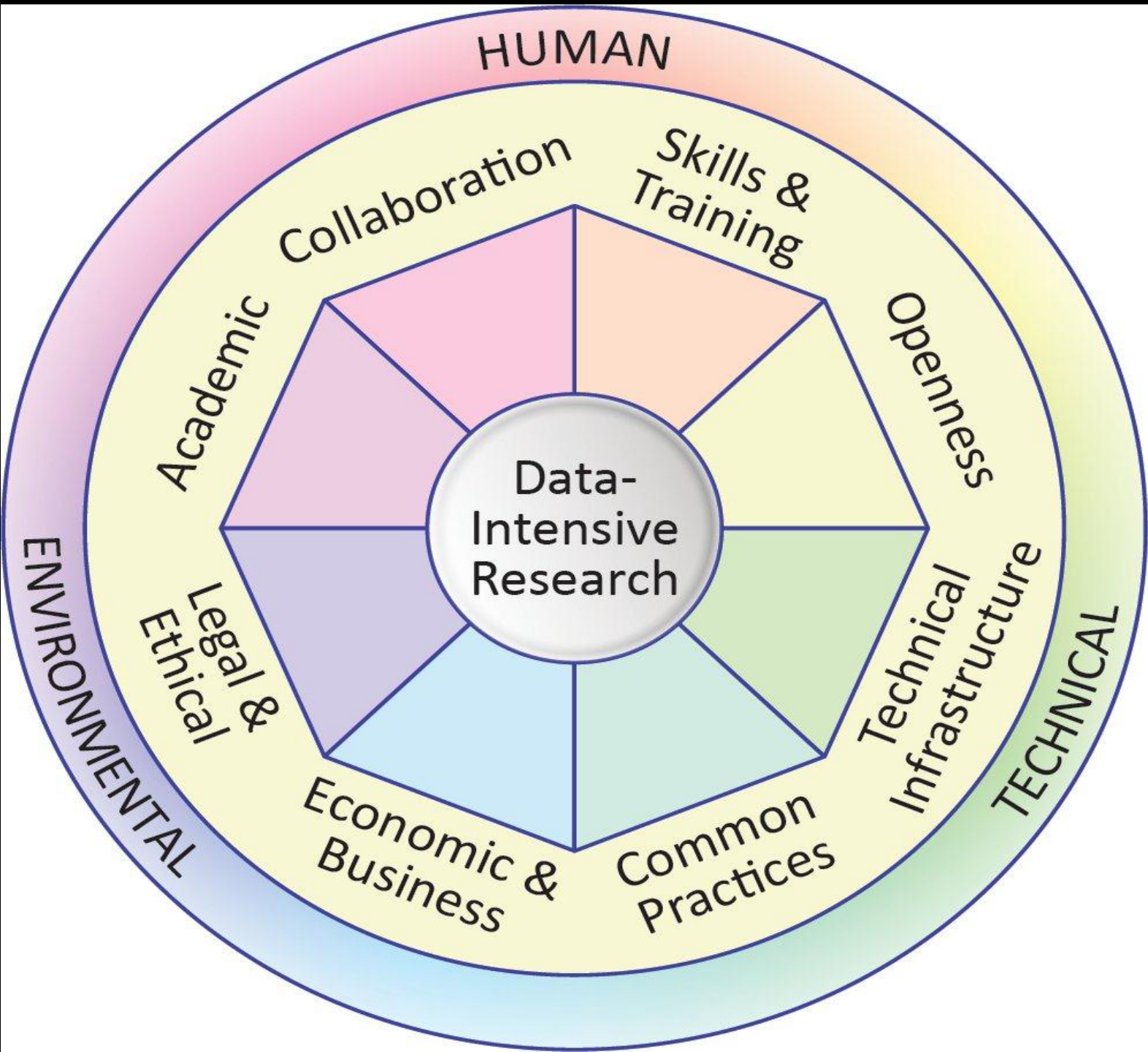
Managing data across the institutional research lifecycle

Understanding the data science habitat : PI, institution, funder



Community Capability Model Framework

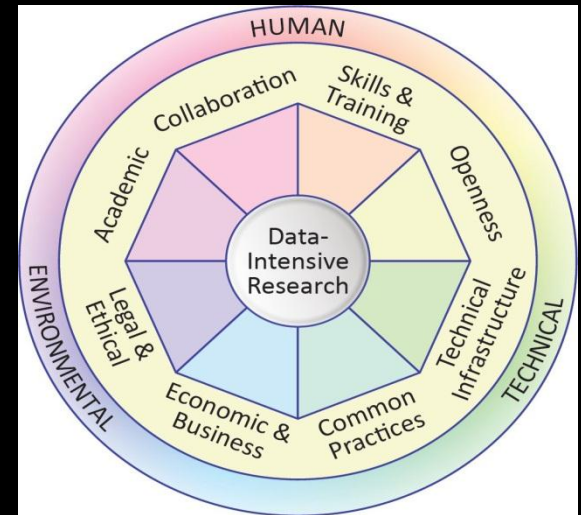




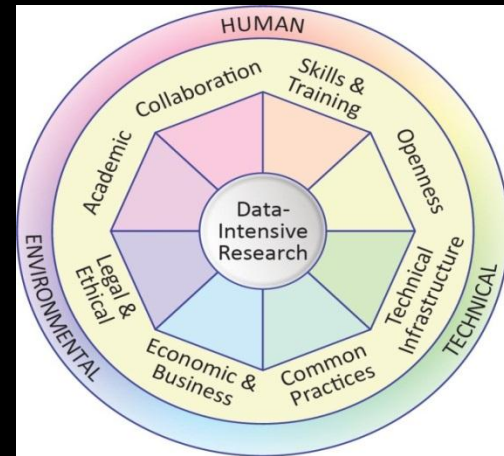
CCMF 8 Capability Factors

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*

8th International Digital Curation Conference,
Amsterdam, 14-16 January 2013

Thank you.



ResearchConnections



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

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