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

UKOLN

DCC

UKOLN is supported by:

JISC

Microsoft Research

UNIVERSITY OF BATH

THE SUNDAY TIMES UNIVERSITY OF THE YEAR 2011-12

www.ukoln.ac.uk

A centre of expertise in digital information management

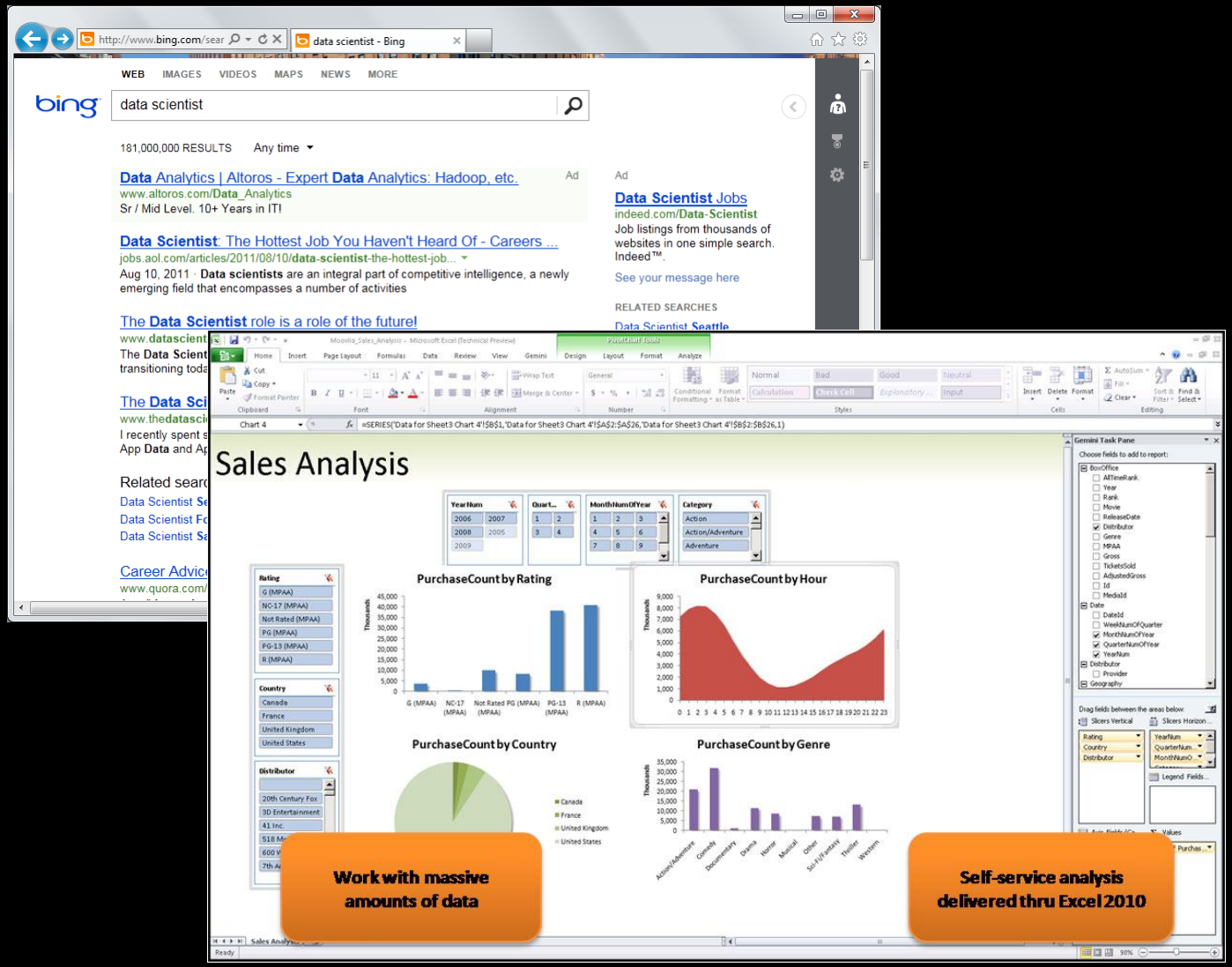
UKOLN

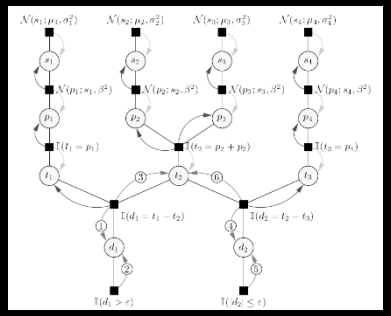
Running order…..

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

- Data scientist flavours
- Data scientist habitat

What is *Data* Science?





information management

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

What is Data ***Science***? 2

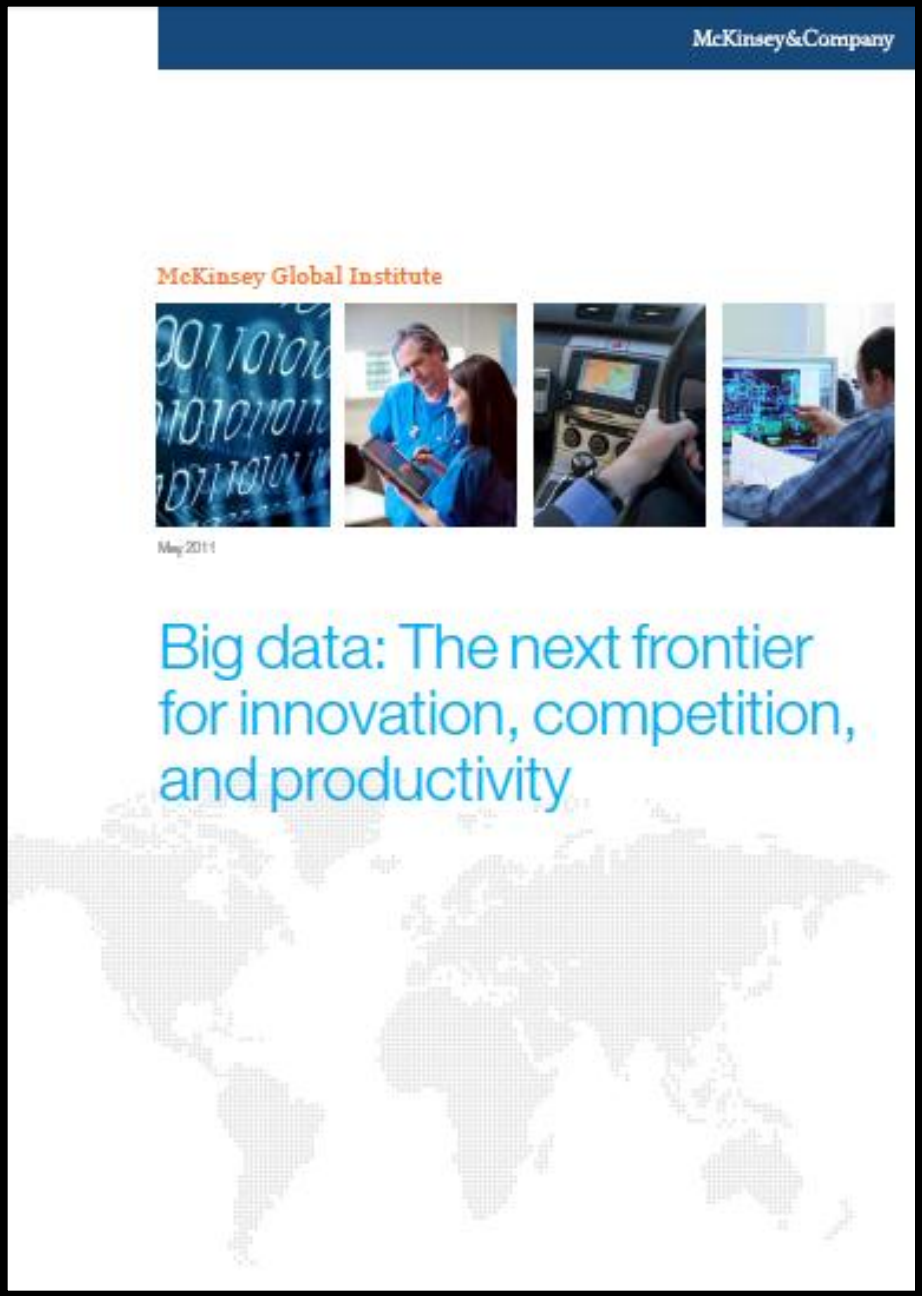
- Science which is data intensive, data driven
- Data as a commodity
- Data as infrastructure
- Data as research substrate
- Data as a science utility
- Data workflows, data tools, data publications
- “The Fourth Paradigm in Practice”

Data : from Big to Broad (Jim Hendler)

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



http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

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

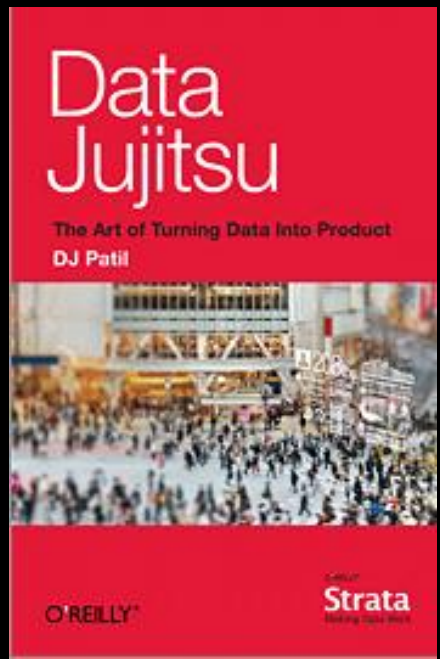
"Data Scientist" Jobs = Near Zero Until 2010

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"***

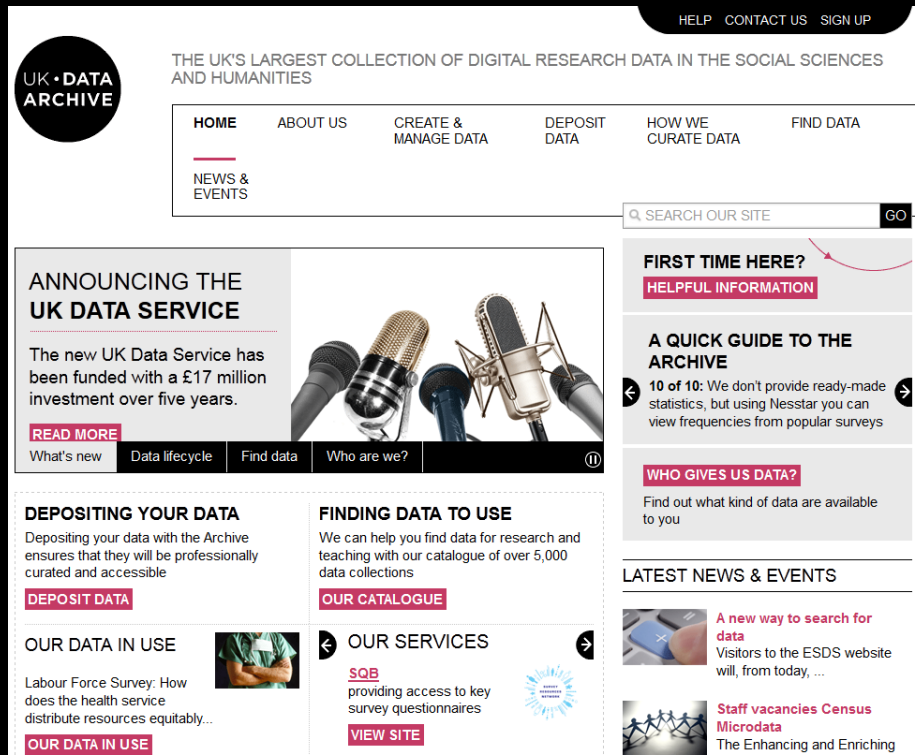
What does a data scientist do? 3

- 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"***

DMPonline
The DCC Data Management Planning Tool

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"***



- 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….

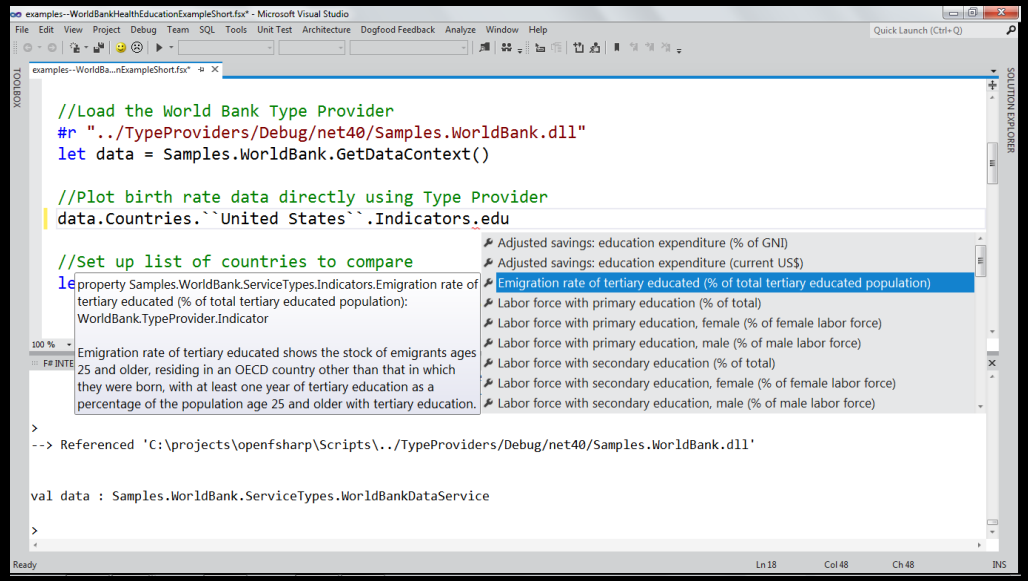
Data Scientist roles

- ***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

Jer Thorp: Hope / Crisis, NYT Word Frequency

Data engineer

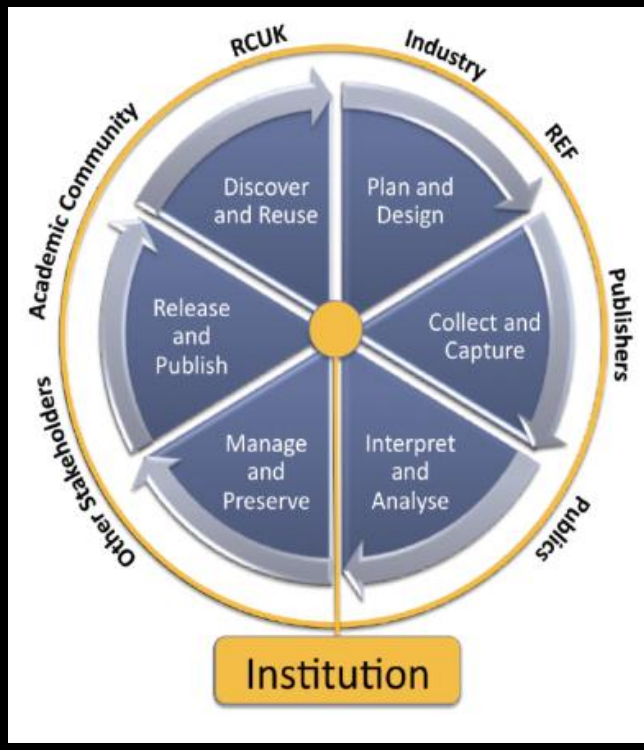
- Focus on software development, coding, programming, tools
- Customises methods and tools for end-users
- Code-focussed
  - R
  - SAS
  - SQL/NoSQL
  - Hadoop
  - F#



http://preview.tryfsharp.org

Institutional data scientist

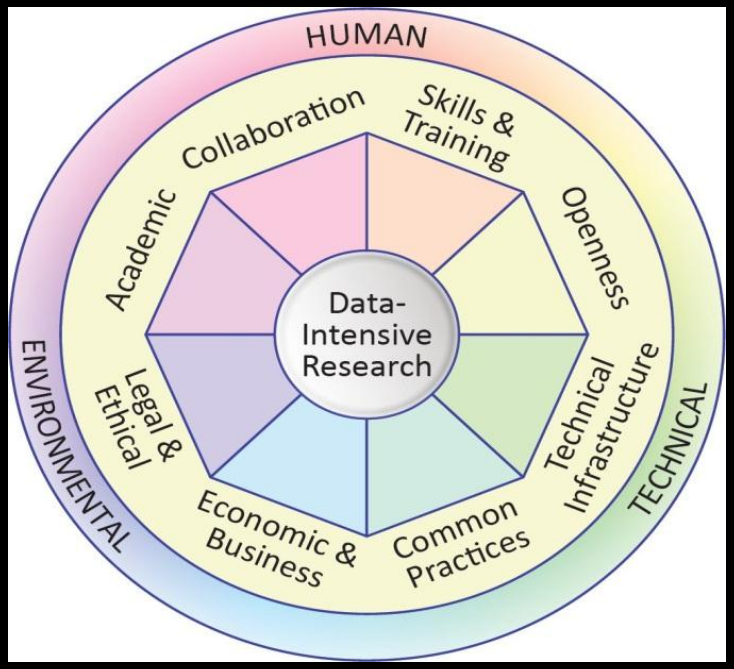
- **Co-ordination and Collaboration**
  - Liaison / subject librarians
  - Repository manager
  - IT/Computing Services
  - Research Support & Development Office
  - Doctoral Training Centres
  - Researchers
- **Advocacy**
- **Training**



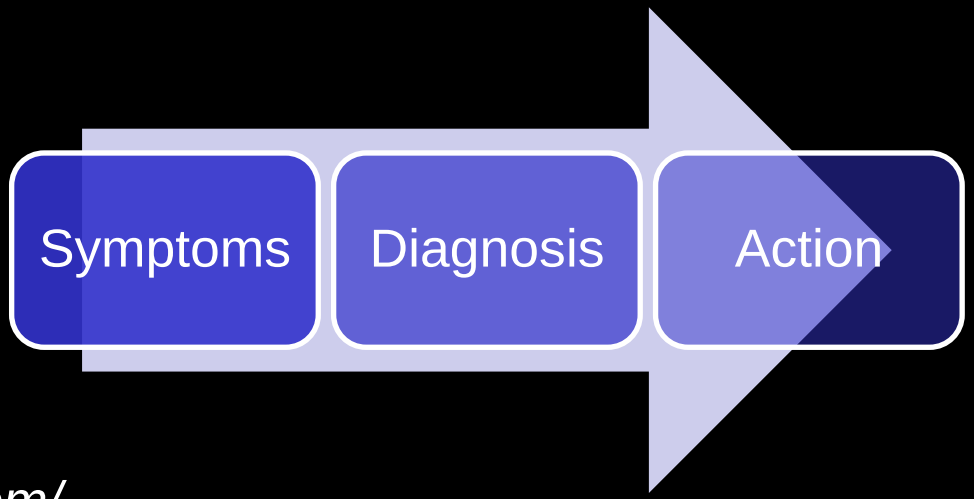
Liz Lyon, Informatics Transform, IJDC Current Issue, 2012

Research360
Managing data across the institutional research lifecycle

Understanding the data science habitat : PI, institution, funder

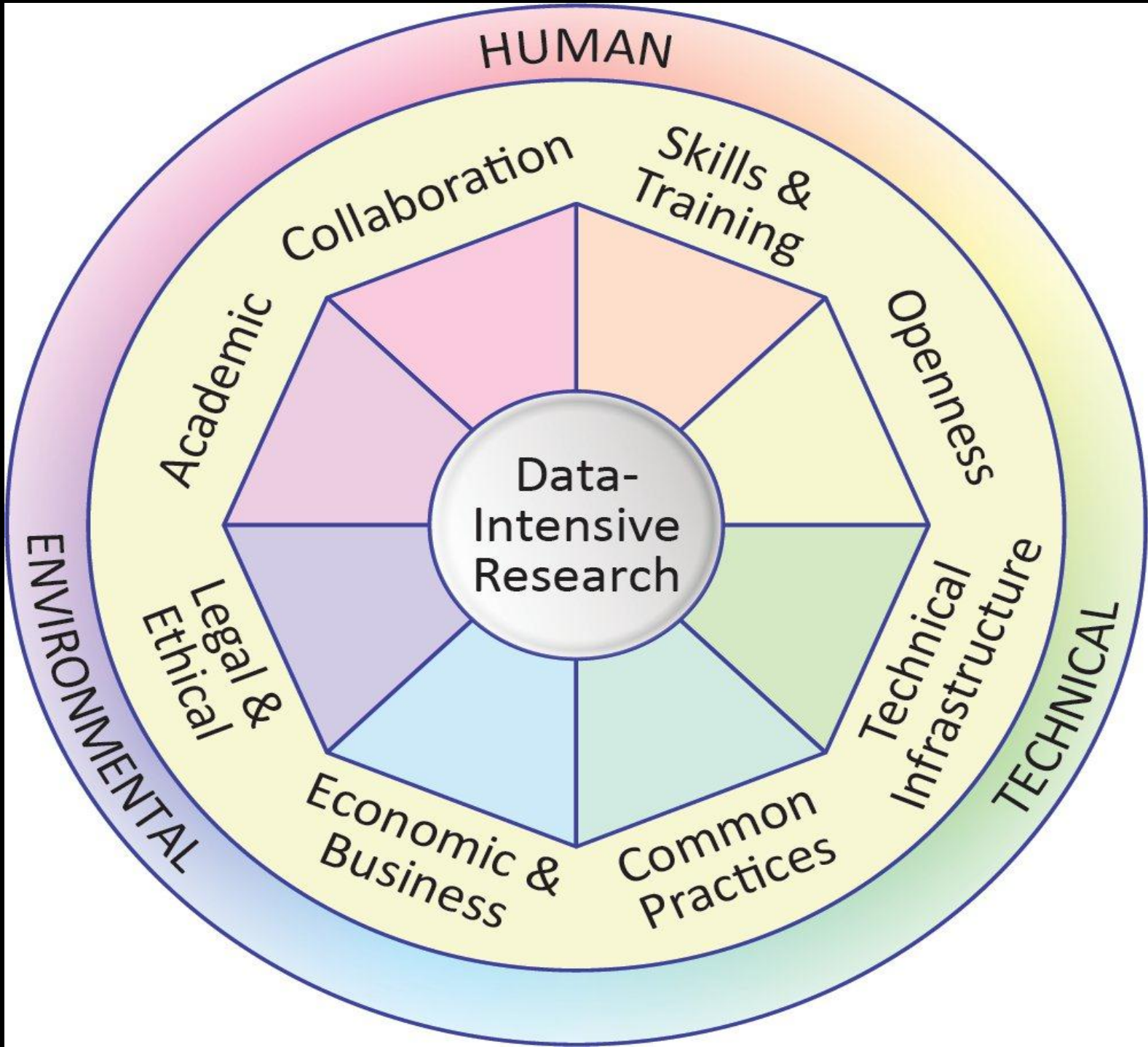


Community Capability Model Framework



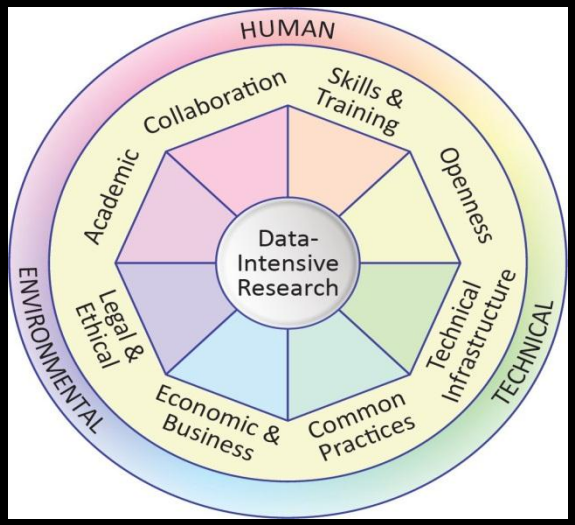
http://communitymodel.sharepoint.com/

CCMF 8 Capability Factors



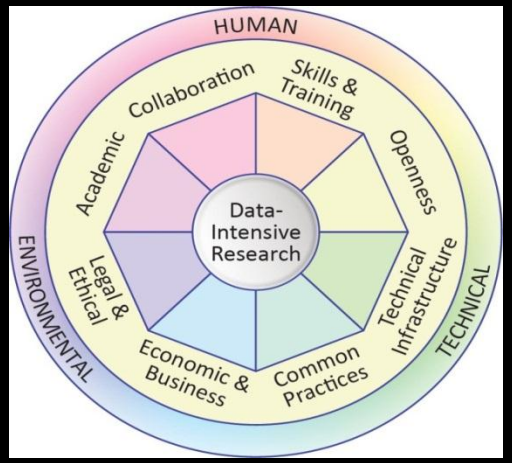
CCMF supporting data science

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



http://communitymodel.sharepoint.com/

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

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