HDRUK
Health Data Research UK

UNIVERSITY OF LEEDS

A Roadmap to Data Maturity: A practical guide for CIOs to maximise the valuable use of data

Monica Jones
Chief Data Officer – University of Leeds
Associate Director & National Strategic Lead - HDRUK

HDRUK
Health Data Research UK

Our mission & vision

HDR UK's mission is to unite the UK's health data to enable discoveries that improve people's lives

Our 20-year vision is for large scale data and advanced analytics to benefit every patient interaction, clinical trial, biomedical discovery and enhance public health.

Core funders 2023-28

UKRI Medical Research Council

NIHR | National Institute for Health and Care Research

British Heart Foundation

CANCER RESEARCH UK

UKRI Economic and Social Research Council

UKRI Engineering and Physical Sciences Research Council

Ymchwil Iechyd a Gofal Cymru
Health and Care Research Wales

CHIEF SCIENTIST OFFICE

HSC Public Health Agency
Research and Development

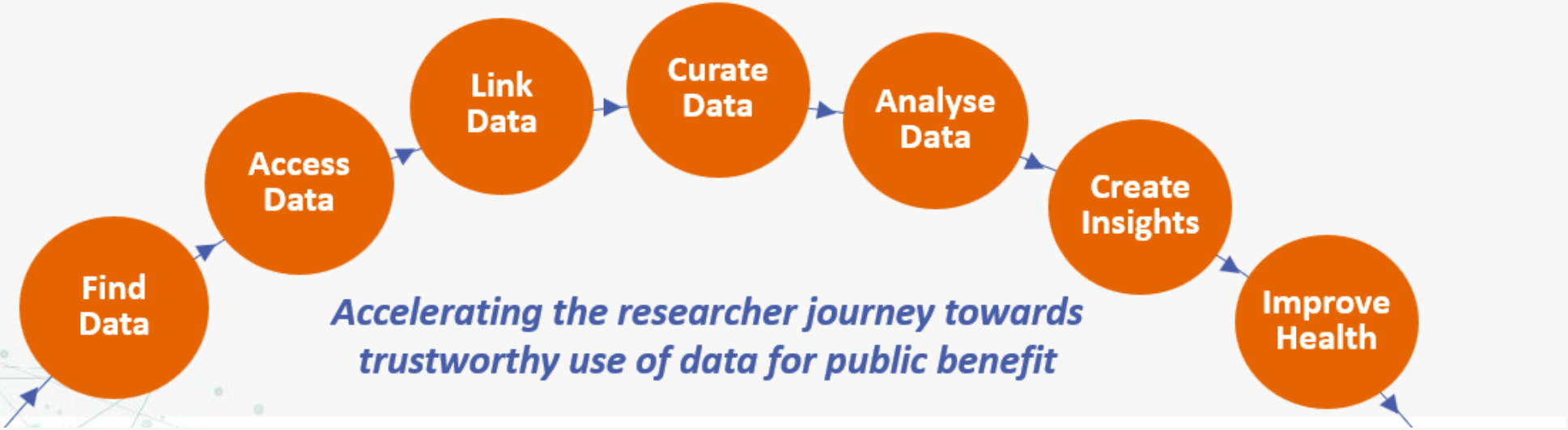
What are we about?

HDRUK
Health Data Research UK

Improve health and boost UK science by making it easier for researchers to find, access and use diverse, high quality data

Provide leadership to fix difficult technical problems, by creating innovative solutions needed for researchers to use large-scale data safely and securely

Accelerate & streamline health data science by developing open collaborations that connect data, people and organisations across the UK and internationally



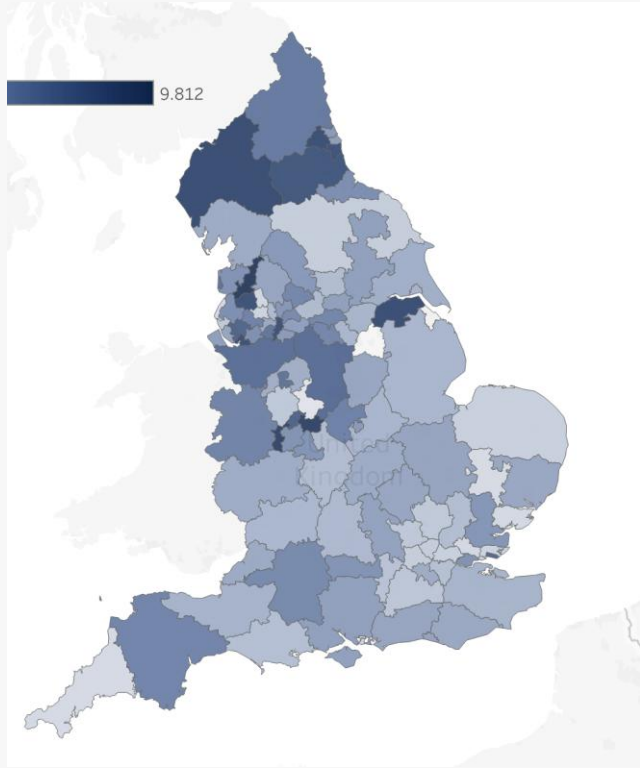
The impact of COVID-19 on cancer, using open standards such as SNOMED CT, in both local and national cancer data

DATA-CAN
The Health Data Research Hub for Cancer

HDRUK
Health Data Research UK

Covid-19 was shown to **impact cancer referrals** and have a significant **impact on projected excess deaths**. [1]

Access to the National SDE holding over **12 billion rows** of data and more than **50 million patients** in **England**

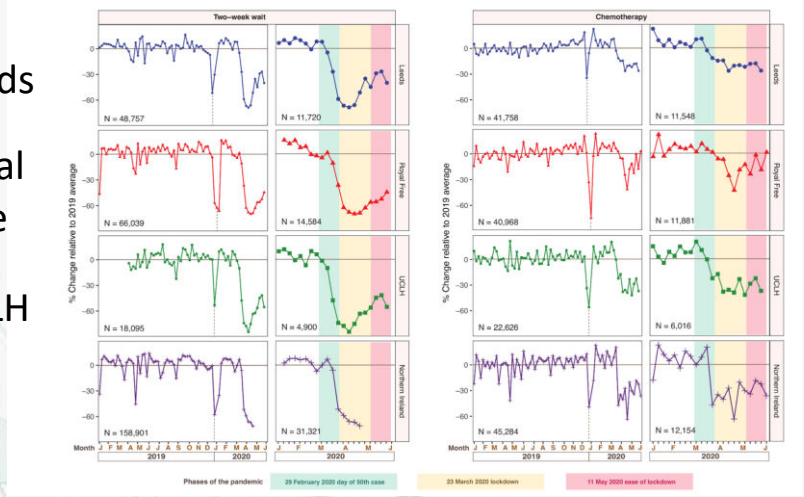


Leeds

Royal Free

UCLH

N.I.



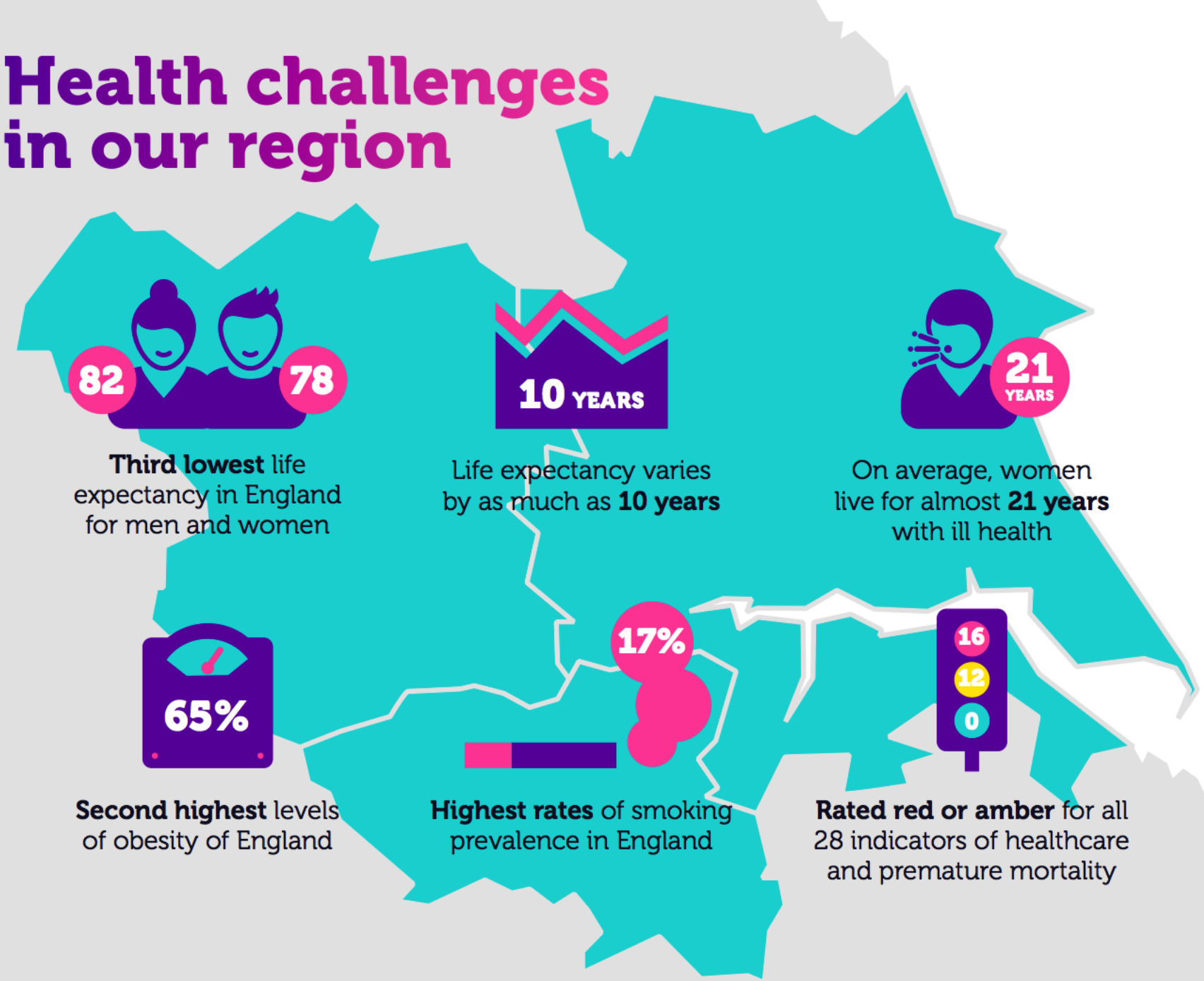
databricks

DELTA LAKE

Lai AG, Pasea L, Banerjee A, Jones M, et al. Estimated impact of the COVID-19 pandemic on cancer services and excess 1-year mortality in people with cancer and multimorbidity: near real-time data on cancer care, cancer deaths and a population-based cohort study. **BMJ Open 2020;10:e043828**

UCLPartners IQVIA Genomics england NHS The Leeds Teaching Hospitals NHS Trust UNIVERSITY OF LEEDS QUEEN'S UNIVERSITY BELFAST

Health challenges in our region



Third lowest life expectancy in England for men and women

Life expectancy varies by as much as **10 years**

On average, women live for almost **21 years** with ill health

Second highest levels of obesity of England

Highest rates of smoking prevalence in England

Rated red or amber for all 28 indicators of healthcare and premature mortality

Yorkshire & Humber Care Record

Resources in our region

5.5m citizens

£12bn annual health and social care investment

22 Acute Trusts

6 are also Teaching Hospitals

15 Local Authorities

5 Mental Health Trusts

20 Clinical Commissioning Groups

770 GP Practices

3 Integrated Care Systems

12 Universities

1,450 care homes

INTERWEAVE
CONNECTING CARE

What is Interweave ?

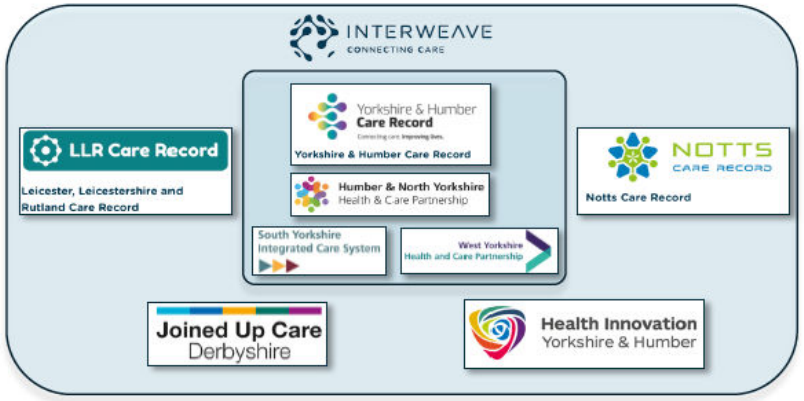
Interweave is a shared care record solution in use across six Integrated Care Systems which together cover a total population of 9.409m, 15.3% of the population of England.

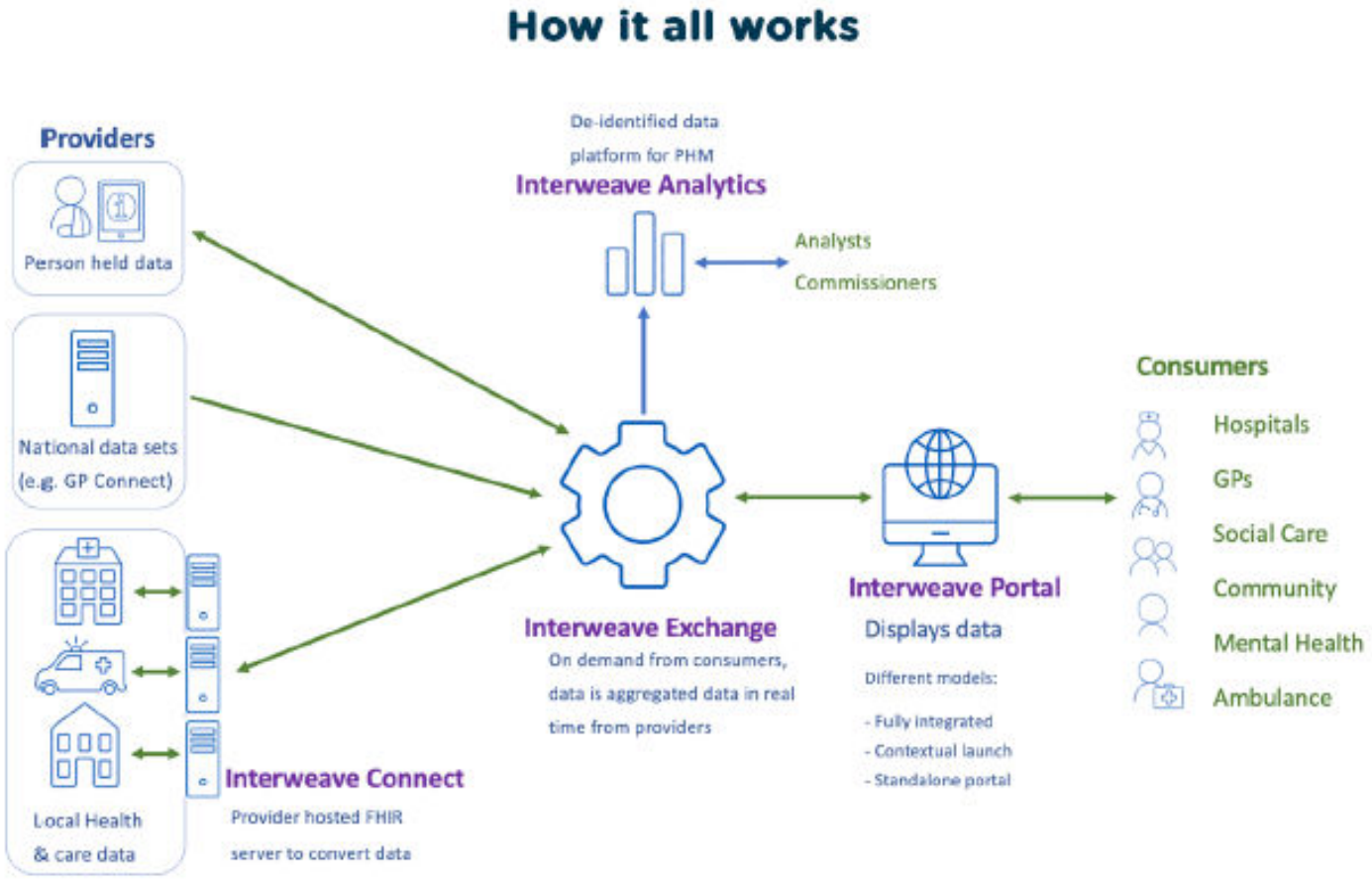
In terms of contiguous ICSs, Interweave supports only a slightly smaller population than the single largest Shared Care Record deployment, in London, but over a much wider and diverse geography.

Uniquely it has been designed and delivered by an in-house team with support from an open source systems supplier partner. The code is open source, published on GitHub and NHS IP and has been designed to utilise modern technologies.

The Interweave Partnership

Interweave is owned and managed by its users – the 6 ICS's – and supported by Health Innovation Yorkshire & Humber. Each ICS is represented on the Interweave Management Board which is responsible for the development and operation of the solution, and which is accountable to the 6 ICBs.





[Home | Yorkshire & Humber Care Record]

Data Utility Framework - Development

HDRUK
Health Data Research UK

Importance of Healthcare Data

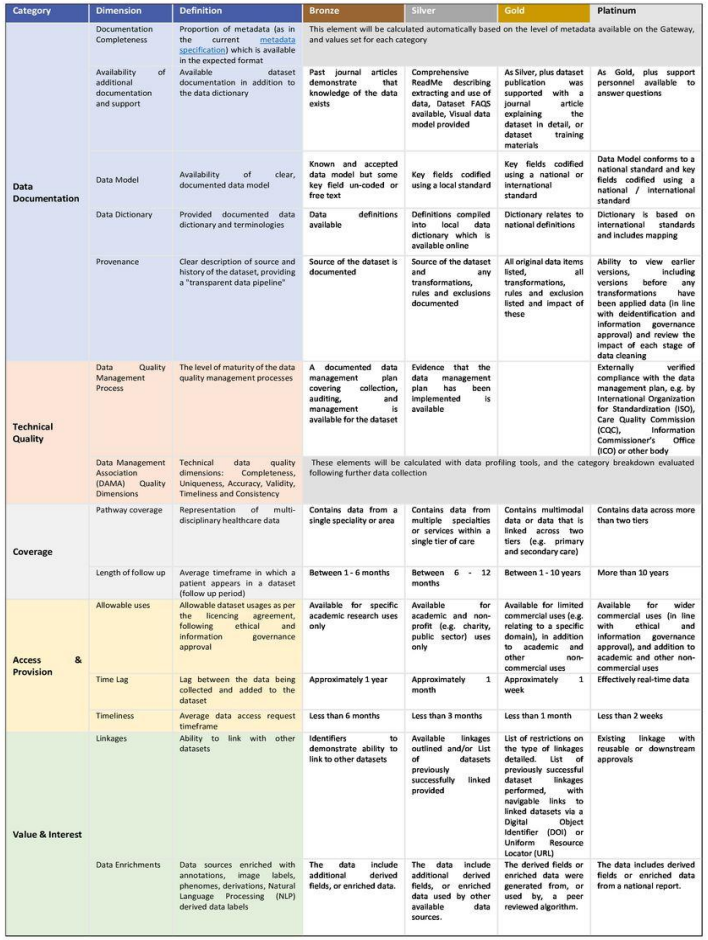
- The value of healthcare data was increasingly recognised so there was a need to improve dataset utility.
- No established evaluation mechanism existed, making it difficult to measure improvements.

Objective

- The goal was to develop a framework for evaluating and communicating healthcare dataset utility.

Conclusion

- A user-centered framework was created for objective evaluation.
- It provides value to both healthcare data users and custodians and helped optimise data curation investments



Category	Dimension	Definition	Bronze	Silver	Gold	Platinum
Data Documentation	Documentation Completeness	Proportion of metadata (as in the current metadata specification) which is available in the expected format	This element will be calculated automatically based on the level of metadata available on the Gateway, and values set for each category			
	Availability of additional documentation and support	Available dataset documentation in addition to the data dictionary	Past journal articles demonstrate that knowledge of the data exists	Comprehensive ReadMe describing extracting and use of data, Dataset FAQS available, Visual data model provided	As Silver, plus dataset publication was supported with a journal article explaining the dataset in detail, or dataset training materials	As Gold, plus support personnel available to answer questions
	Data Model	Availability of clear, documented data model	Known and accepted data model but some key field un-coded or free text	Key fields codified using a local standard	Key fields codified using a national or international standard	Data Model conforms to a national standard and key fields codified using a national / international standard
	Data Dictionary	Provided documented data dictionary and terminologies	Data definitions available	Definitions compiled into local data dictionary which is available online	Dictionary relates to national definitions	Dictionary is based on international standards and includes mapping
	Provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions documented	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including versions before any transformations have been applied data (in line with deidentification and information governance approval) and review the impact of each stage of data cleaning
Technical Quality	Data Quality Management Process	The level of maturity of the data quality management processes	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available		Externally verified compliance with the data management plan, e.g. by International Organization for Standardization (ISO), Care Quality Commission (CQC), Information Commissioner's Office (ICO) or other body
	Data Management Association (DAMA) Quality Dimensions	Technical data quality dimensions: Completeness, Uniqueness, Accuracy, Validity, Timeliness and Consistency	These elements will be calculated with data profiling tools, and the category breakdown evaluated following further data collection			
Coverage	Pathway coverage	Representation of multi-disciplinary healthcare data	Contains data from a single speciality or area	Contains data from multiple specialties or services within a single tier of care	Contains multimodal data or data that is linked across two tiers (e.g. primary and secondary care)	Contains data across more than two tiers
	Length of follow up	Average timeframe in which a patient appears in a dataset (follow up period)	Between 1 - 6 months	Between 6 - 12 months	Between 1 - 10 years	More than 10 years
Access & Provision	Allowable uses	Allowable dataset usages as per the licencing agreement, following ethical and information governance approval	Available for specific academic research uses only	Available for academic and non-profit (e.g. charity, public sector) uses only	Available for limited commercial uses (e.g. relating to a specific domain), in addition to academic and other non-commercial uses	Available for wider commercial uses (in line with ethical and information governance approval), and addition to academic and other non-commercial uses
	Time Lag	Lag between the data being collected and added to the dataset	Approximately 1 year	Approximately 1 month	Approximately 1 week	Effectively real-time data
	Timeliness	Average data access request timeframe	Less than 6 months	Less than 3 months	Less than 1 month	Less than 2 weeks
Value & Interest	Linkages	Ability to link with other datasets	Identifiers to demonstrate ability to link to other datasets	Available linkages outlined and/or List of datasets previously successfully linked provided	List of restrictions on the type of linkages detailed. List of previously successful dataset linkages performed, with navigable links to linked datasets via a Digital Object Identifier (DOI) or Uniform Resource Locator (URL)	Existing linkage with reusable or downstream approvals
	Data Enrichments	Data sources enriched with annotations, image labels, phenomes, derivations, Natural Language Processing (NLP) derived data labels	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from a national report.

Development of a data utility framework to support effective health data curation | BMJ Health & Care Informatics

Ben Gordon, Monica Jones, Neil Sebire et al. BMJ Health Care Informatics 2021;28:e100303

QUANTUM
THE HEALTH DATA QUALITY LABEL

HDRUK
Health Data Research UK

Developing a data quality and utility label for HealthData@EU

European Health Data Space (EHDS) for secondary use

What is QUANTUM goal?

- QUANTUM is an EU-funded project (2024-2026) that aims to **create a common label system** for Europe that guarantees the quality and utility of datasets for scientific and health innovation purposes.
- This label system will enable researchers, policymakers, and healthcare professionals to **identify high-quality data for research and decision making.**

Learn more here

QUANTUM & the secondary use of health data (HealthData@EU)

QUANTUM will address Article 56 of the European Health Data Space (EHDS) regulation proposal, which mandates labelling health datasets to show their **quality and usefulness for secondary use of health data.**

What is HealthData@EU?
It is the pilot infrastructure for the secondary use of health data in the European Health Data Space (EHDS), which will provide a system for reusing health data for **research, innovation, policy-making, and regulatory activities.**

			Level 1	Level 2	Level 3	Level 4	Level 5		
Serial	Dimension	Definition	**Initial:** a new or undocumented process, unpredictable, poorly controlled and reactive	**Emerging (or Documented, Structured):** the process is documented sufficiently so repeating the same steps may be attempted	**Defined:** the process is defined and confirmed as a standard process	**Managed:** the process is quantitatively and proactively managed in accordance with agreed-upon metrics	**Optimised:** process management includes deliberate process optimisation	Reference	Comments
1	Data collection process	The processes relating to the capture of data from multiple sources for secondary purposes	No data collection or capture process for secondary purposes	Ad hoc processes exist for data collection	Defined processes exist for some data collection	Defined processes exist for all data collection and are used for secondary purposes	Processes are automated for all data collection, focus on continuous process improvement.	ISO 8000-62 data maturity model	
2	Data management - governance	The level of maturity of the data management processes	No documented data management governance	A documented data management plan covering collection, auditing, and management is available for the dataset	Evidence that the data management plan has been implemented is available	Demonstrated compliance with the data management plan	Externally verified compliance with the data management plan	Adapted HDRUK	Covers more than just data management planning
3	Data management - infrastructure	The level of implementation and development of the data holder's data management infrastructure	No data management infrastructure	An emerging data management infrastructure, some validation and verification	Data management infrastructure defined and confirmed as a standard process	Data management infrastructure, with partially automated, verified and validated [real time] data management	A robust and comprehensive data management infrastructure, with fully automated, verified and validated real time data management	European Health Data Space / FAIR	Cover security of infrastructure (throughout the dimensions)
4	Data provenance	Clear description of source and history of the dataset, providing a "transparent data pipeline"	No documented provenance	Source of the dataset is documented	Source of the dataset and any transformations, rules and exclusions	All original data items listed, all transformations, rules and exclusion listed and impact of these	Ability to view earlier versions, including "raw" or "source" dataset, and review the impact of each stage/step	Adapted HDRUK	
5	Data access	How well defined and implemented are data access processes, from a legal, ethical and technical perspective	No data access processes or procedures	Have the processes and procedures but don't respond in timely and constistent manner	Have the processes and procedures, and respond in timely and constistent manner	Data access system that covers both technical and policy areas, in accordance with agreed metrics	A comprehensive data access system that covers technical, ethical and policy areas (allowable uses, API documentation, access and approvals) compliant with EU Policy	European Health Data Space / FAIR	Cross reference to EHDS Trusted Data Holder
6	Data analytics environment	Analytical services, tooling and access to [secure] data environments	No data environment available	Requested analysis can be undertaken by internal teams and provided back in anonymised format to data requestors	The dataset can be used in a secure data environment (SDE)	The dataset can be used in an SDE and other data and tools can be brought in as required	The dataset can be used in federated organised environment	Adapted HDRUK	
7	Data enhancement - augmentation	The application of various techniques to make data more useable for specific purposes	No data augmentation	Some techniques to make data more useable for specific purposes	Defined techniques to make data more useable for specific purposes	Managed techniques to make data more useable for specific purposes	Comprehensive application of various techniques and mapping to data model i.e. OMOP to make more useable for specific purposes	OMOP / OpenEHR / HL7 FHIR / ISO/IEC 22989:2022	This is different from Data enrichment. They are both subtypes of Data Enhancement.
8	Data enhancement - enrichment	Data sources enriched for example with annotations, image labels, phenomes, derivations, NLP derived data labels	The data has no additional derived fields, or enriched data.	The data include additional derived fields, or enriched data.	The data include additional derived fields, or enriched data used by other available data sources.	The derived fields or enriched data were generated from, or used by, a peer reviewed algorithm.	The data includes derived fields or enriched data from an [inter] national report.	Adapted HDRUK	This is different from Data augmentation. They are both subtypes of Data Enhancement.
9	Data Model	Availability of clear, documented data model that provides structure and standardisation	There is no data model	Known and accepted data model but some key field uncoded or free text	Key fields codified using a local standard and updated over time	Key fields codified using a national or international standard and updated	Data model conforms to an [inter] national standard and key fields codified using a national / international standard	Adapted HDRUK / OMOP / OpenEHR / HL7 FHIR	Interoperability and Standardisation
10	Data Dictionary	Provided documented data dictionary and terminologies	No Data Dictionary	Data definitions available	Comments complete and local data dictionary which is available online	Dictionary relates to national definitions	Dictionary is based on international standards and includes mapping	Adapted HDRUK / EHDS	Link to Metadata Standards and Definitions

Home - QUANTUM: The health data quality label

UNIVERSITY OF LEEDS

Data Strategy – What was needed?

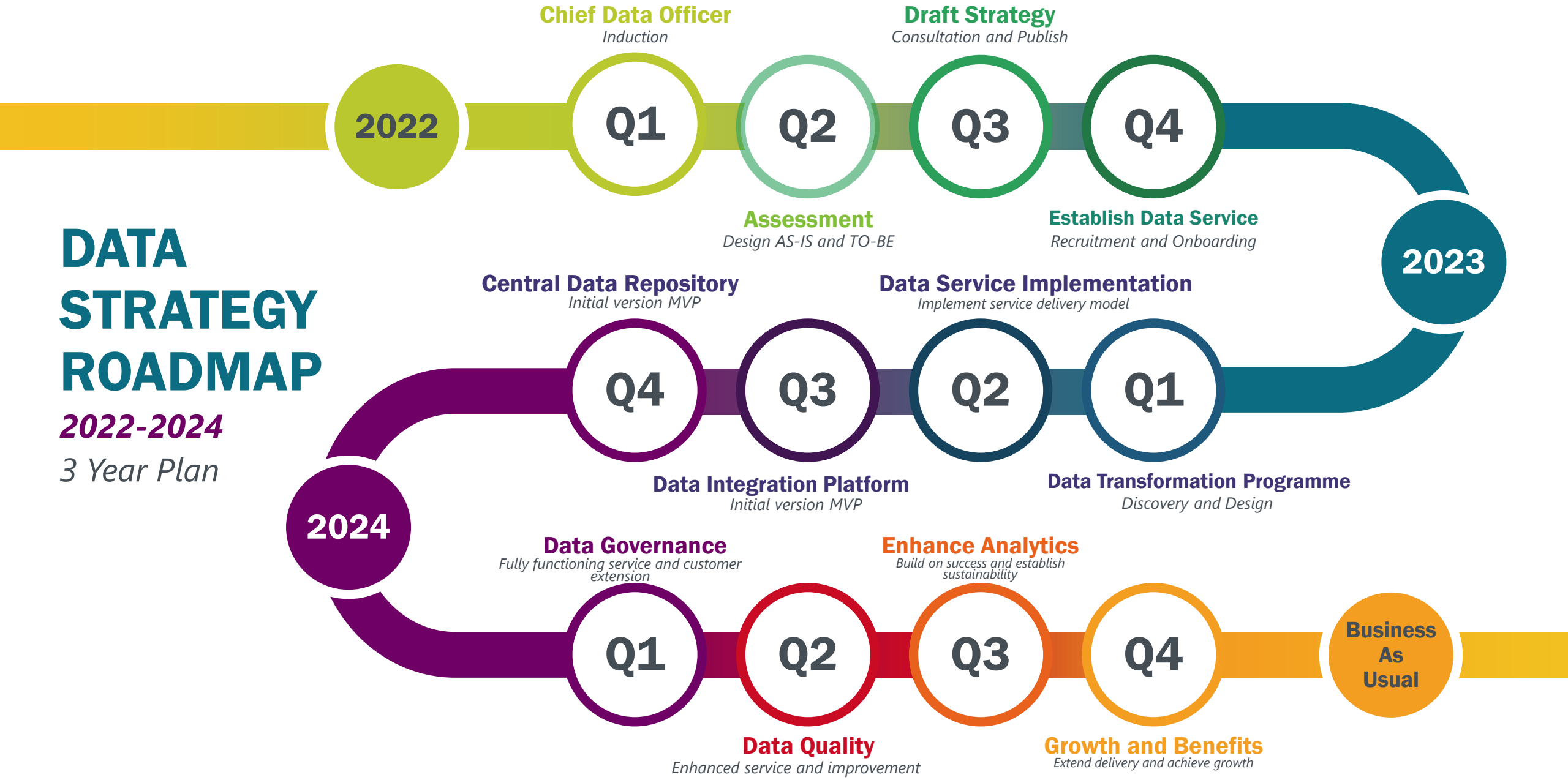
- Data is essential to day-to-day operation as well as longer term planning and performance management
- A Data Strategy sets out a vision and a clear action plan to make better use of data
- It provides an overarching narrative and plan to address the current cultural, behavioural and structural barriers in the system with the ultimate goal of a system that is underpinned by high quality, readily available data
- It powers vital analysis and research to discover new and innovative approaches

Find the Data Strategy here
For further information,
please contact data.services@leeds.ac.uk

DATA STRATEGY ROADMAP

2022-2024

3 Year Plan



Consolidating Disparate Data Sources

One of the most critical steps in preparing for integration is consolidating data from various sources. Organisations often have data scattered across different departments, databases, and even legacy systems. This fragmentation can lead to inconsistencies, duplication, and a lack of coherent data narratives, which impede effectiveness.

1. **Unified Data Repository**: Establishing a centralised data repository, such as a data lake or a data warehouse, is essential. This allows seamless access to all organisational data, breaking down silos and enabling a holistic view of information.

2. **Data Integration Tools**: Utilising advanced data integration tools can streamline the process of merging disparate data sources. These tools can automate the extraction, transformation, and loading processes, ensuring that data from different systems is harmonised and ready for analysis.

3. **APIs and Data Connectors**: Implementing APIs and data connectors facilitates real-time data integration from various sources. These technologies ensure that the centralised repository is always up-to-date, providing accurate and timely data for applications.

Enhancing Data Quality

Algorithms are only as good as the data they are trained on. Poor data quality can lead to inaccurate predictions, biased outcomes, and ultimately, failed initiatives. Therefore, enhancing data quality is paramount.

- **Data Cleansing**: Regular data cleansing routines must be established to remove duplicates, correct errors, and fill in missing values. Automated data cleansing tools can significantly reduce manual effort and improve accuracy.

- **Data Governance Framework**: Implementing a robust data governance framework helps in setting standards for data quality, ensuring consistency, accuracy, and reliability. This framework should include data quality metrics, policies for data usage, and accountability measures.

- **Master Data Management (MDM)**: MDM systems help in maintaining a single, consistent view of key business entities across the organisation. By consolidating master data, organisations can ensure that AI models are trained on the most accurate and up-to-date information.

Implementing Cutting-Edge Data Management Practices

Advanced data management practices are crucial for creating a fertile ecosystem where technologies can thrive. These practices encompass the latest methodologies and tools designed to enhance data accessibility, security, and usability.

Data Cataloguing: Implementing a data catalogue can help in organising and managing data assets. A well-structured data catalogue enables data scientists and practitioners to easily discover and understand the available data, fostering efficient data utilisation.

Data Privacy and Security: With the increasing focus on data privacy, it is essential to implement stringent data security measures. This includes encryption, access controls and compliance with regulations such as GDPR and Data Protection Act 2018. Secure data management practices ensure that sensitive information is protected, thereby maintaining trust and compliance.

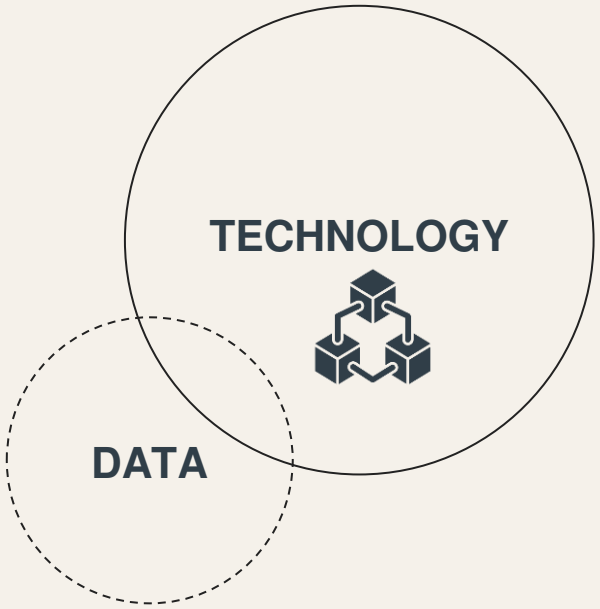
Scalable Data Architecture: Adopting scalable data architectures, such as cloud-based platforms, ensures that the data infrastructure can grow with the organisation's needs. Cloud solutions offer flexibility, scalability, and advanced analytics capabilities, which are critical for supporting initiatives.

Automated Data Workflows: Automation in data workflows can drastically improve efficiency. Implementing tools for automated data ingestion, processing and analysis reduces the time and effort required to prepare data for applications, allowing data scientists to focus on model development and innovation.

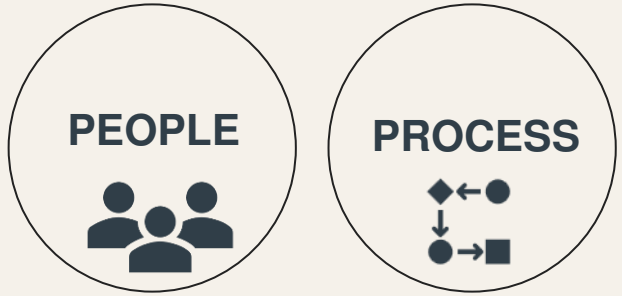
Change is More Than Just Technical Solutions

The organisation now needs to look beyond tech alone in order to adopt a data driven culture to improve the student experience – we will do this through the Data Literacy Programme

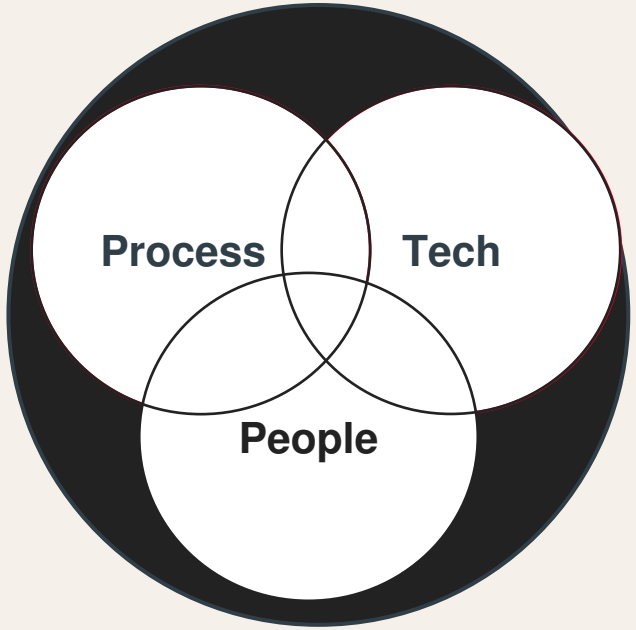
1 Institutions often lead with technical solutions and ‘data-by-product’



2 While under-investing in the people and process elements that use and adopt the tech solutions



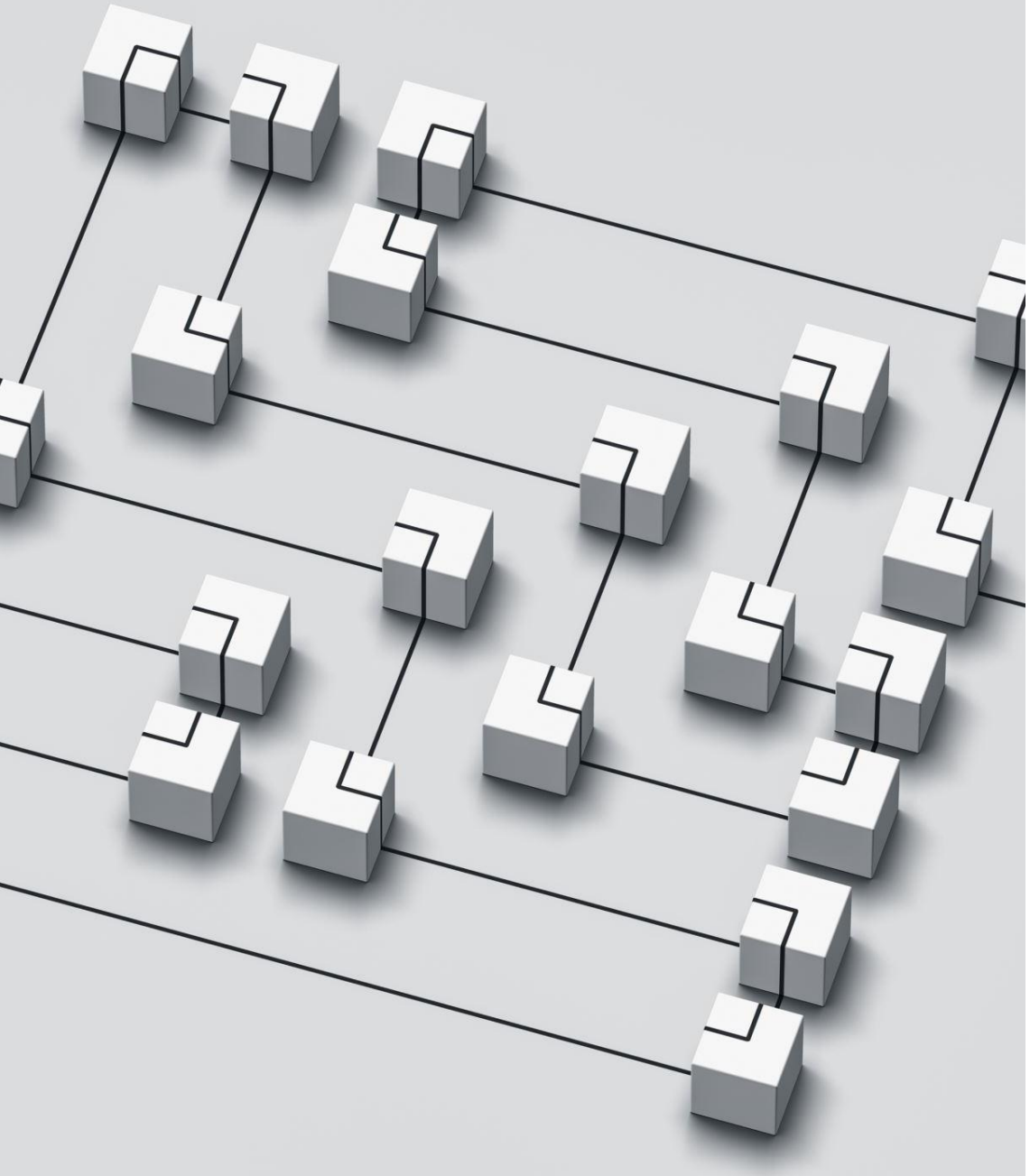
3 And avoid thinking about how these work together to deliver outcomes.



A data-driven culture requires balanced thinking across people, process, technology and data to deliver scalability, value realisation and a competitive advantage.

Lessons Learned

- View your organisation’s wider strategy through a data lens. Make the data strategy relevant to those it affects and keep it clear, succinct and achievable.
- Adopt the approach that business change management should be established first.
- Consult with the wider business and internal networks. Engage individuals to establish what that one thing is that they cannot do that would provide genuine benefit. Deliver on this.
- Recruit and implement a central Data Service department before the implementation of the transformation strategy, to take ownership of the data strategy and support adoption of end users.
- View data as an asset and a core function to enable transformation from burden to asset. Recognise the potential in primary data collections e.g. statutory to release secondary uses e.g. curriculum, improving student services.
- Organise these data sets to enable secondary uses with an emphasis on the ‘as is’ to the ‘to be’.



Top Tips

- Create and promote a set of Data Principles.
- Convene a Data Design Authority (DDA) to act as a mechanism for advising and benchmarking projects and initiatives against the data principles to ensure governance, assurance and standardisation.
- Emphasise the benefits and added value through developing a benefits dependency network. Showcase these through internal workshops and network events.
- If the project/activity does not deliver on the benefit, question the validity to the objective and subsequently the strategic driver. All activities should have a purpose.
- Document and be transparent around commitments and future plans.
- ***Finally, simply talk to people and listen to their concerns and ideas.***

Assessing Impact

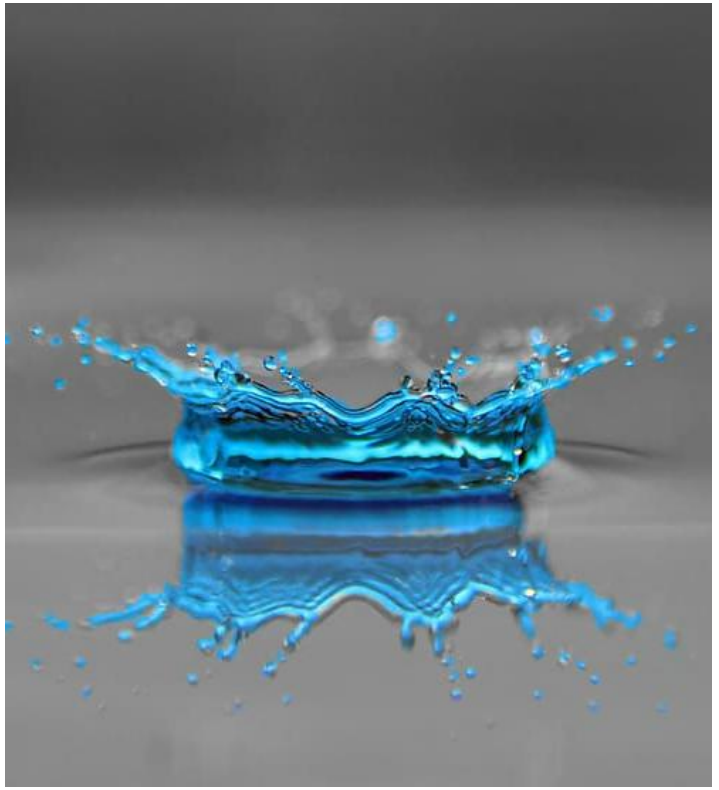
A fully implemented central Data Service removed the burden from end users and turned defensive behaviours into positive collaborations, instilling a ‘data as an asset’ culture;

A trusted data repository, metadata and cataloguing capabilities provides the infrastructure to build on;

Reusable and reliable data adds value and improves data quality;

Increased data-driven decision making;

Benefits realisation work has proven to be a positive way of sense checking suitability of projects across the business. Stakeholders have really identified and engaged with this initiative.



[Implementing a data strategy to create a trusted data culture – Data analytics]

Questions ?

m.c.m.jones@leeds.ac.uk

www.linkedin.com/in/monicacmjones