# Implementing the FAIR Data Sharing Principles: Technology and Experiences

Khaled El Emam
David Sibbald
Rodrigo Barnes

17th *June 2021*

kelemam@replica-analytics.com
david.sibbald@aridhia.com
rodrigo.barnes@aridhia.com

# Agenda

**Context** 1 The demand for more robust data sharing mechanisms and technology

**Research Lifecycle** 2 Overview of a typical research lifecycle

**Data Synthesis** 3 An overview of data synthesis methods

**Demonstration** 4 A live demonstration of integration of data synthesis within a data sharing platform

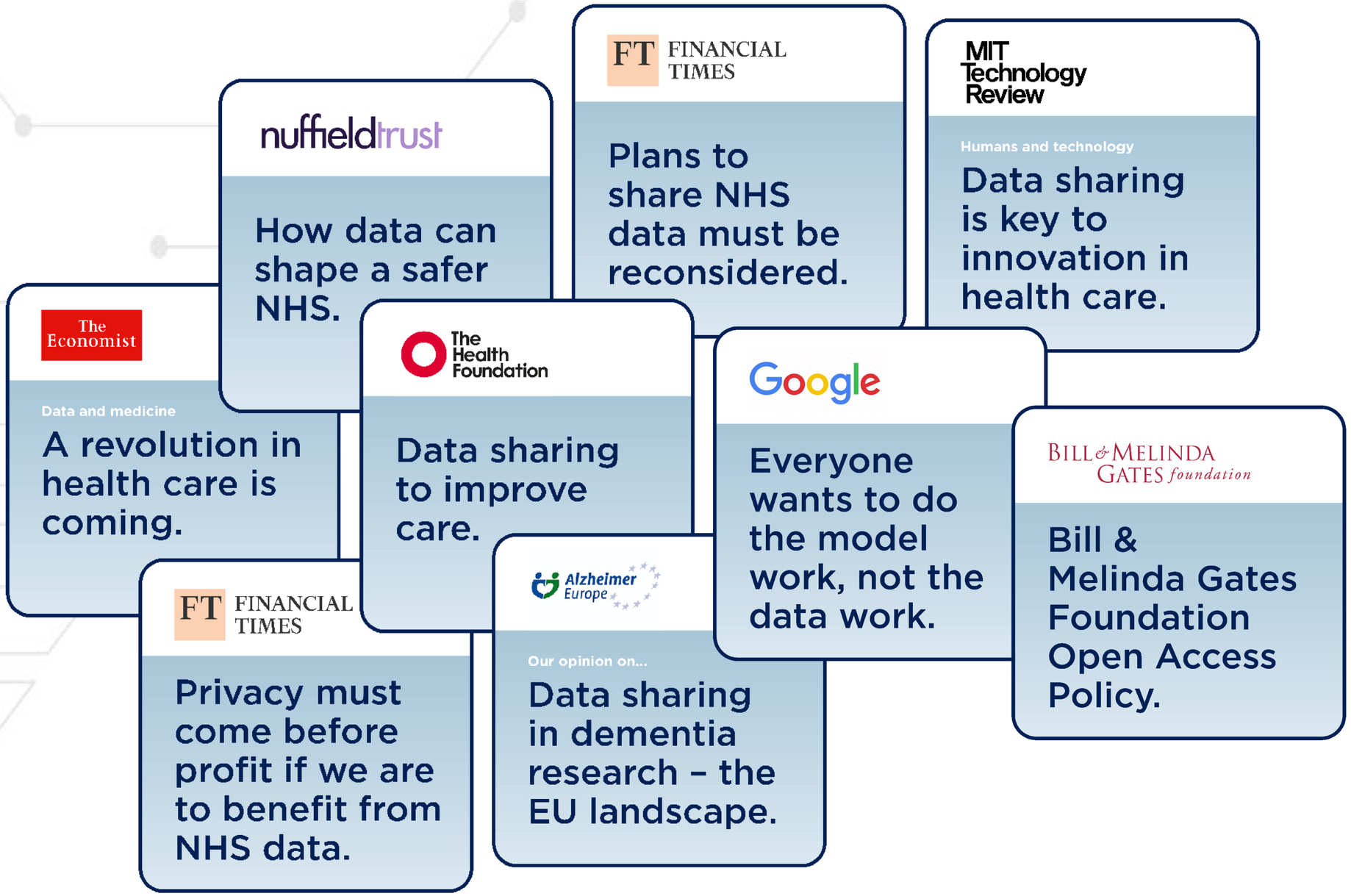




aridhia DRE

Replica Analytics

# What's so hard about data sharing?



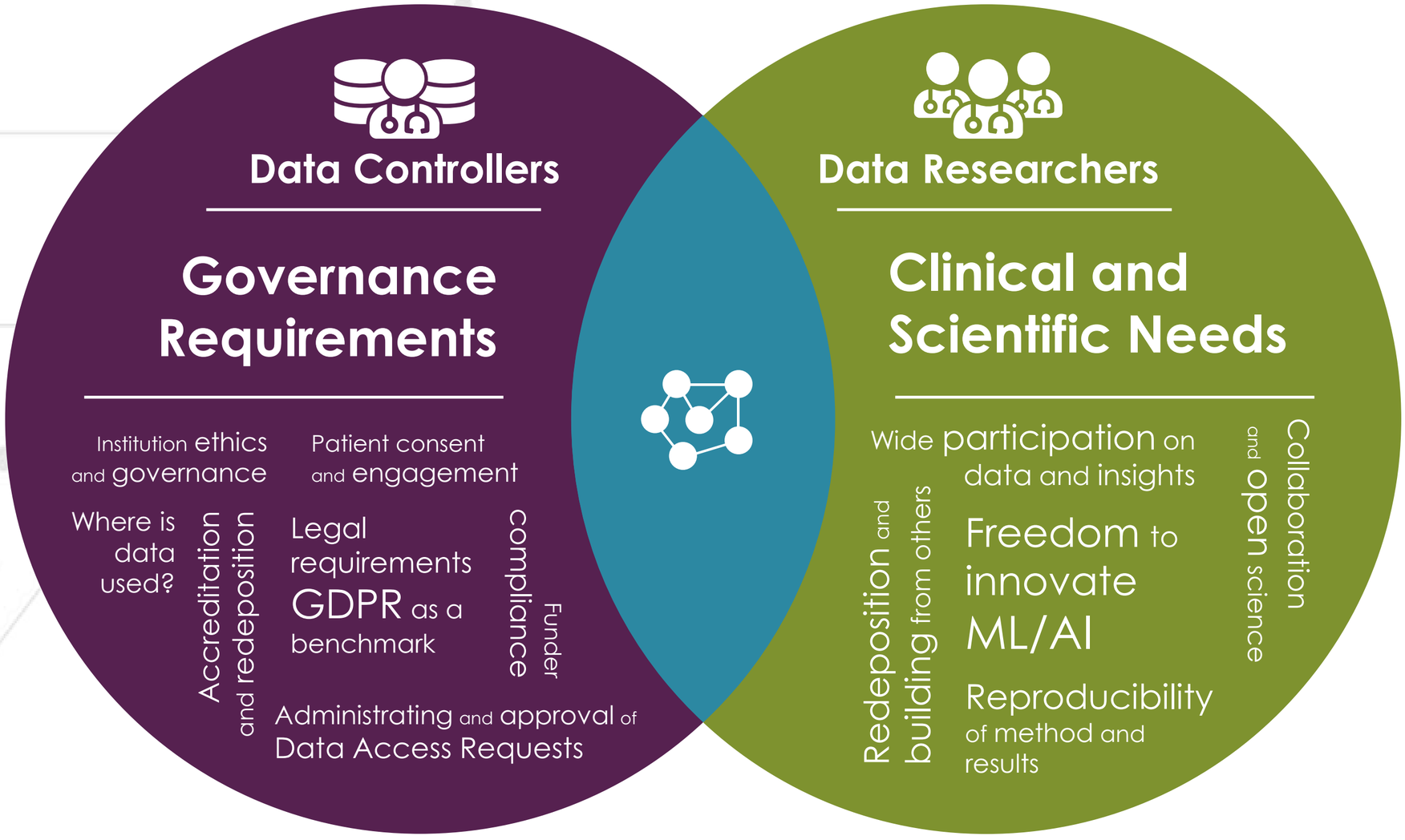





aridhia DRE

Replica Analytics

# The need for proportionality



# Benefits of Data Sharing


| | New business models | Faster and more innovative product development | New or enhanced customer services and experiences | More efficient or innovative manufacturing | Greater speed and visibility across supply chains | Cybersecurity or prevention of fraud |
|---|---|---|---|---|---|---|
| Consumer goods and retail | | X | | | X | |
| Financial services | X | | | | | X |
| Travel and hospitality | | X | | | X | |
| IT and telecoms | | | X | | | X |
| Manufacturing | | X | | X | X | |
| Pharma and healthcare | | X | | | X | |
| Transport and logistics | | X | | | X | |
| Energy and utilities | X | | | | X | |
| Media and marketing | | X | X | | | X |
| Professional services | | | | | X | X |
| Government | | X | | | X | |

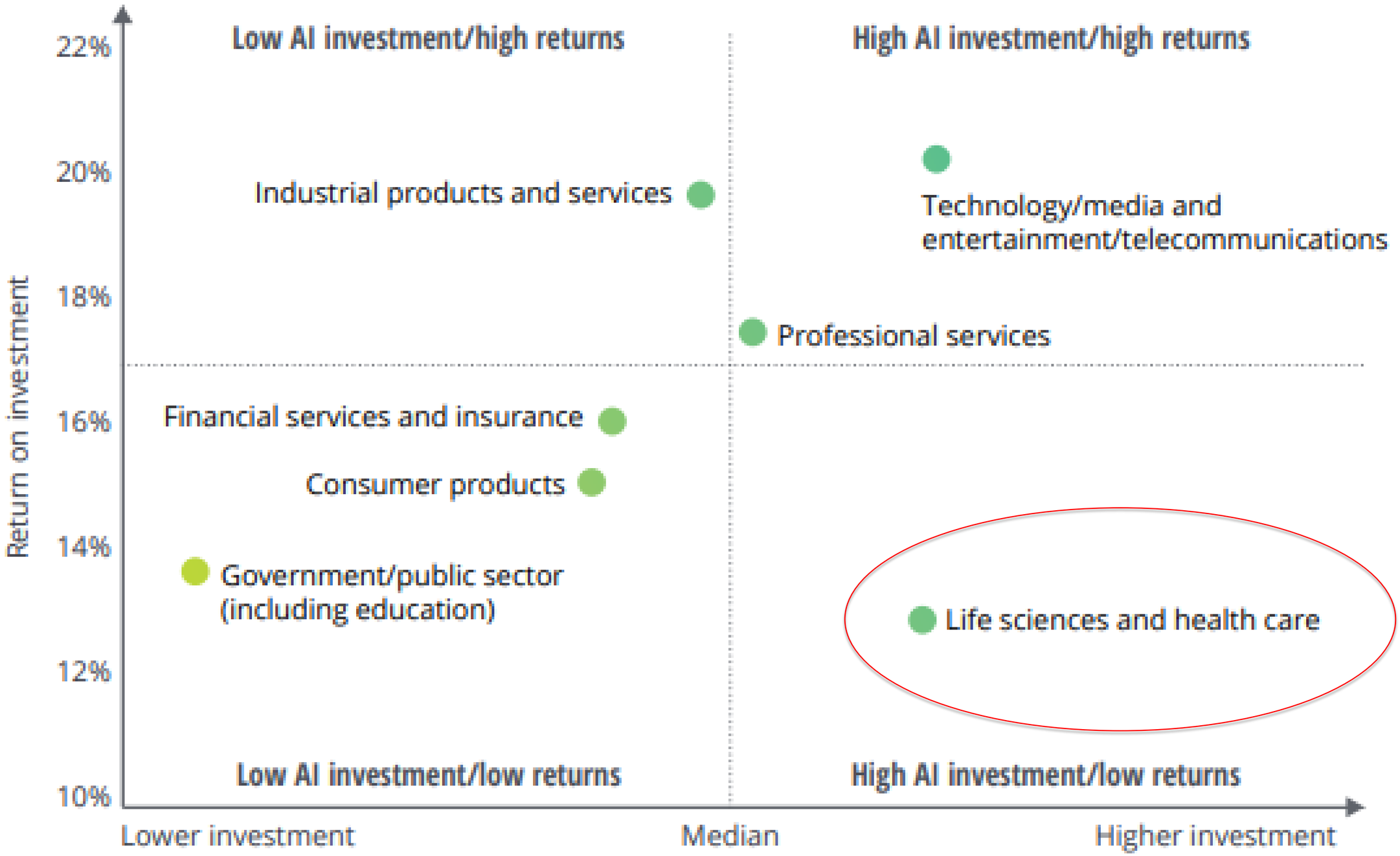
Source: MIT Technology Review Insights survey, 2020



aridhia DRE

Replica Analytics

# High Investment in AIML .. But




Deloitte 2018

# High Investment in AIML .. But

Top challenges for AI initiatives: Ranked 1–3, where 1 is greatest challenge

| | Ranked 1 | Ranked 2 | Ranked 3 | Ranked top three |
|---|---|---|---|---|
| Implementation challenges | 13% | 14% | 12% | 39% |
| Integrating AI into the company's roles and functions | 14% | 13% | 12% | 39% |
| Data issues (e.g., data privacy, accessing and integrating data) | 16% | 13% | 10% | 39% |
| Cost of AI technologies/ solution development | 13% | 12% | 11% | 36% |
| Lack of skills | 11% | 10% | 10% | 31% |
| Challenges in measuring and proving business value | 10% | 11% | 9% | 30% |

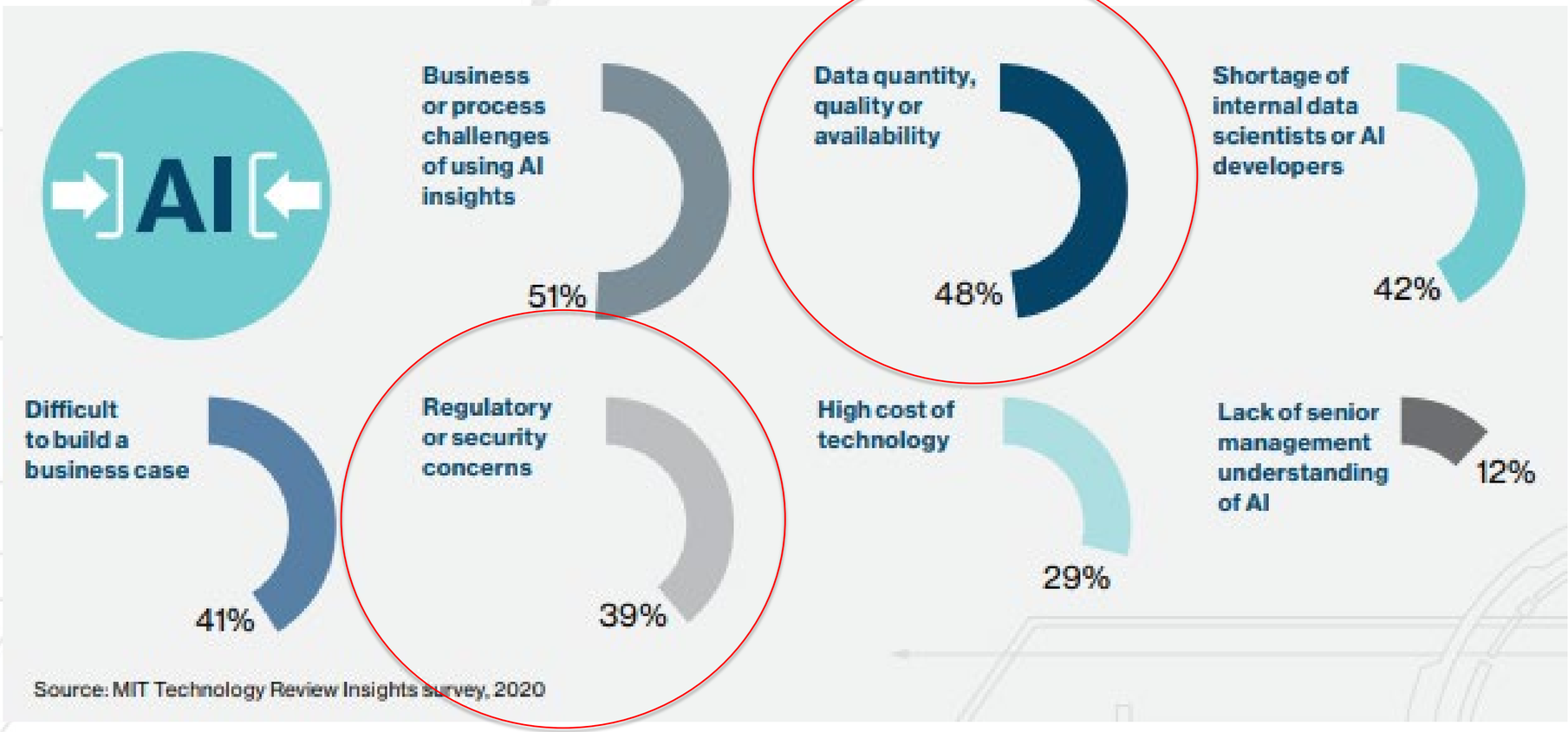
Deloitte 2018



aridhia DRE

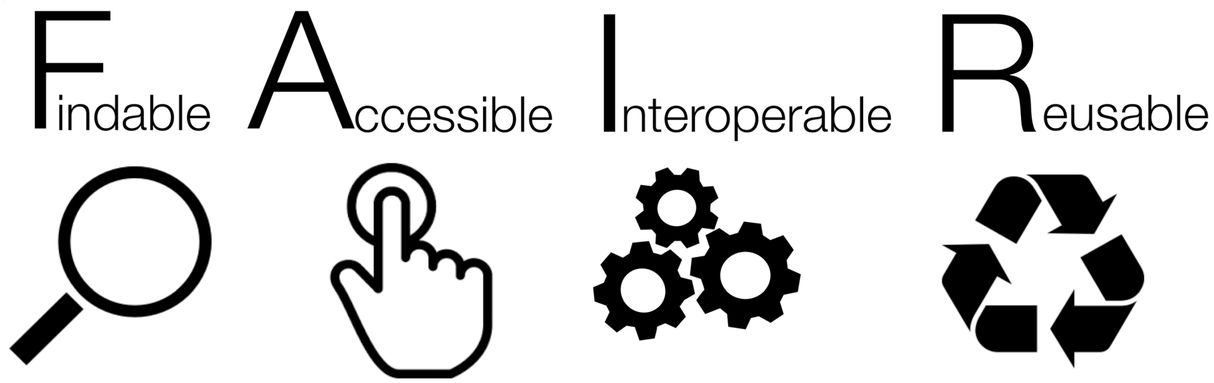
Replica Analytics

# Biggest Constraints



# FAIR Principles

https://snd.gu.se/sites/default/files/inline-images/fair.jpg




### FINDABLE

Making data discoverable, identifiable and searchable via the assignment of metadata and unique identifiers.

### ACCESSIBLE

Available and retrievable data with access via authentication and authorisation procedures.

### INTEROPERABLE

Parseable and semantically understandable data allowing the broadest possible data exchange.

### REUSABLE

Accurately described data with associated provenance and well documented, easily shared usage rights.

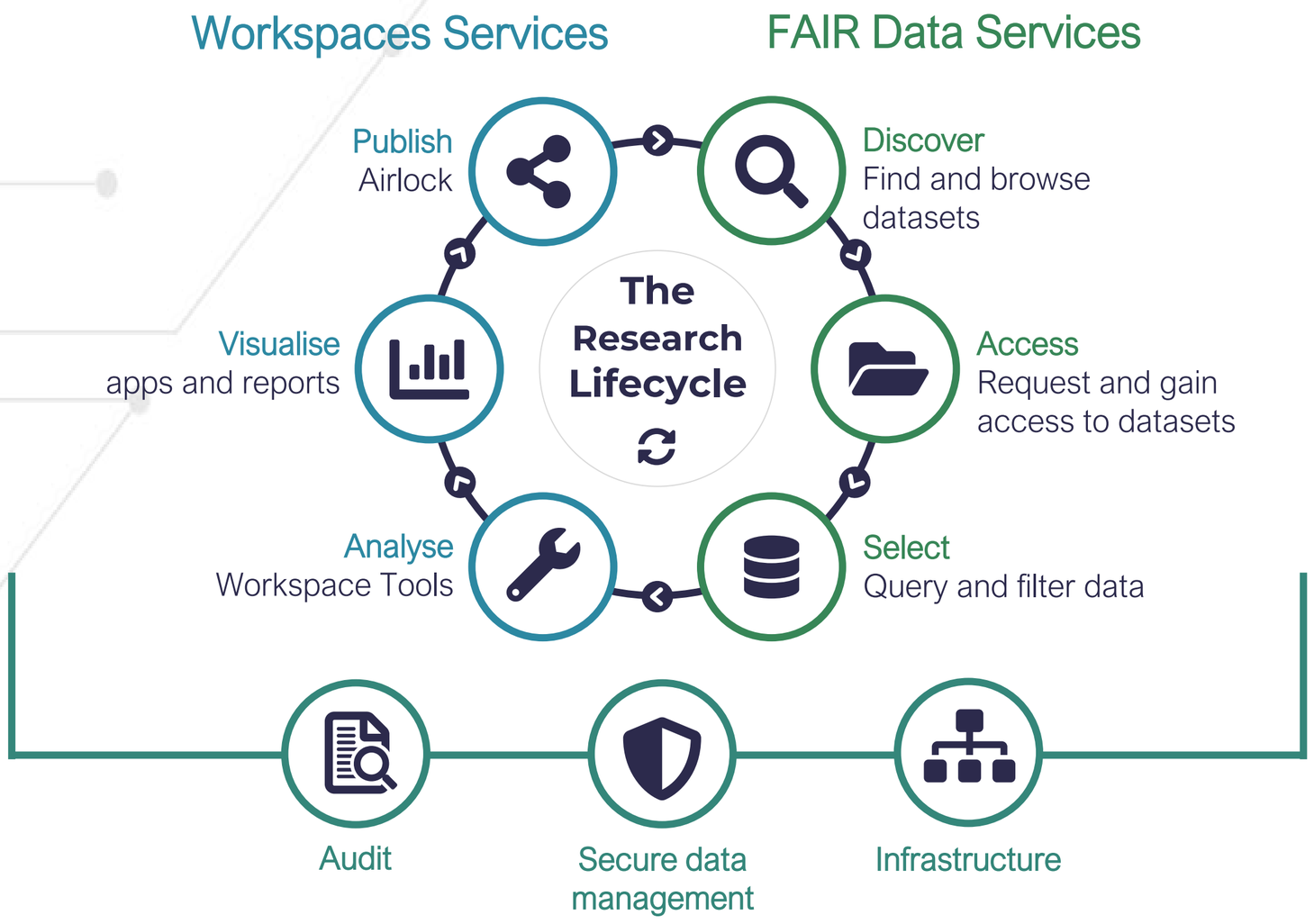




aridhia DRE

Replica Analytics

# Research Lifecycle



# Synthetic data on demand

## FAIR Data Services

## Workspaces

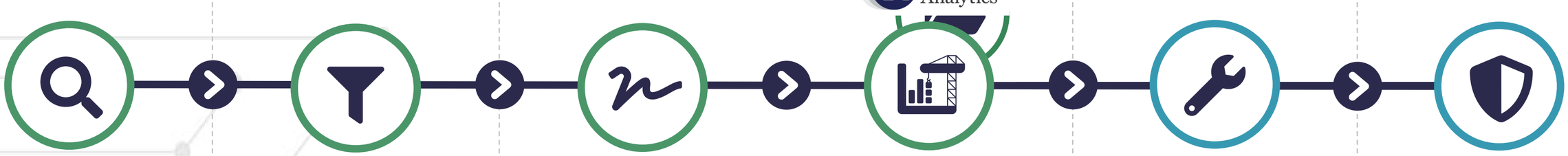
***Gateway 1***
*Accredited to select*

***Gateway 2***
*Request for data*

***Gateway 3***
*Inbound Airlock*

***Gateway 4***
*Outbound Airlock*

Replica Analytics



**Data discovery**

*Open*

**Selection & filtering**

*Accredited User*

**Authorisation**

*Data Steward*

**Synthetic Data**

*Automated*
*Policy driven*

**Analysis**

*Accredited User*

**Airlock**

*P.I. or Authorised Approver*

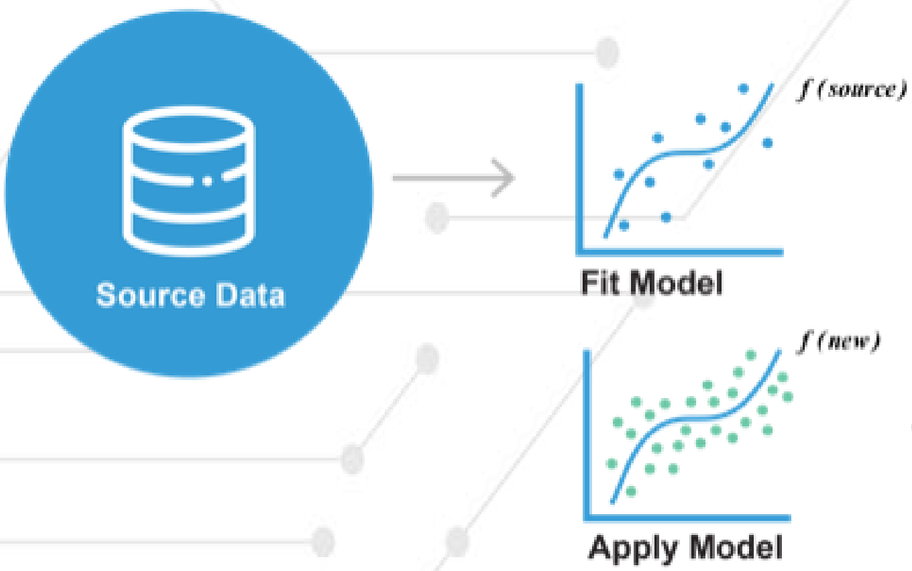
***Audit and Governance Reporting***



aridhia DRE

Replica Analytics

# The Synthesis Process




| COU1A | AGECAT | AGELE70 | WHITE | MALE | BMI |
|---|---|---|---|---|---|
| United States | 2 | 1 | 1 | 1 | 33.75155 |
| United States | 2 | 1 | 1 | 0 | 39.24707 |
| United States | 1 | 1 | 1 | 0 | 26.5625 |
| United States | 4 | 1 | 1 | 1 | 40.58273 |
| United States | 5 | 0 | 0 | 1 | 24.42046 |
| United States | 5 | 0 | 1 | 0 | 19.07124 |
| United States | 3 | 1 | 1 | 1 | 26.04938 |
| United States | 4 | 1 | 1 | 1 | 25.46939 |





aridhia DRE

Replica Analytics

# Synthetic Data Uses

- Data Sharing and Data Access
  - AI and data science projects
  - Software testing
  - Proof of concept and technology evaluations
  - Open data/open science
  - Hackathons and data competitions/challenges
- Data Amplification and Data Augmentation
  - Amplifying small datasets
  - Correct bias

# Privacy Risks

| Dataset | Fully Synthetic Data | Original Data |
|---|---|---|
| **Washington Hospital Data** | 0.0197 | 0.098 |
| **Canadian COVID Data** | 0.0086 | 0.034 |

A commonly used risk threshold = 0.09





aridhia DRE

Replica Analytics

# Live Demonstration

# QUESTIONS

# References


- Z. Azizi, C. Zheng, L. Mosquera, L. Pilote, K. El Emam: “Replicating Secondary Studies Using Synthetic Clinical Trial Data”, *BMJ Open*, 11:e043497, 2021.
- K. El Emam, L. Mosquera, E. Jonker, H. Sood: “Evaluating the Utility of Synthetic COVID-19 Case Data”, *JAMIA Open*, 14(1):ooab012, January 2021.
- K. El Emam, L. Mosquera, and C. Zheng, “Optimizing the Synthesis of Clinical Trial Data Using Sequential Trees,” *JAMIA*, 28(1): 3-13, 2021.
- K. El Emam, L. Mosquera, and J. Bass, “Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation,” *JMIR*, vol. 22, no. 11, Nov. 2020. [Online]. Available: https://www.jmir.org/2020/11/e23139.
- K. El Emam, L. Mosquera, and R. Hoptroff, Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O’Reilly, 2020.
- K. El Emam, “Seven Ways to Evaluate the Utility of Synthetic Data,” *IEEE Security and Privacy*, July/August, 2020.

# Thank you

- For more information about Replica Analytics:

  www.replica-analytics.com

- For more information about Aridhia:

  www.aridhia.com

- Please contact us for more information about the integrated DRE with synthetic data generation capabilities