# Datanomics

## Costs and Value of Research Data


Illustration by Jørgen Stamp digitalbevaring.dk CC BY 2.5 Denmark

Neil Beagrie

(Charles Beagrie Ltd)

Counting on Reproducibility

DPC/Jisc Briefing Day, Birmingham, May 2019

Charles Beagrie

# Some (KRDS) Partners

# Outline

- Costs
- Valuing Intangible Assets (Quantitive Approaches)
- Where Do We Go From Here?
  - »“What to Keep” study
  - »Infonomics – an industry perspective
- Concluding thoughts

# Costs

# Costs

Charles Beagrie

## Activity Cost Models

- Many examples – a few generic (intended for broadly based community), most organisation-specific (derived)
- Substantial effort to create

## Activity Cost Data

- Can be created in a consistent form using a ACM
- Cost Data still takes significant effort to collect and may be incomplete. “Total Costs of Curation” can be distributed across many budget centres/departments

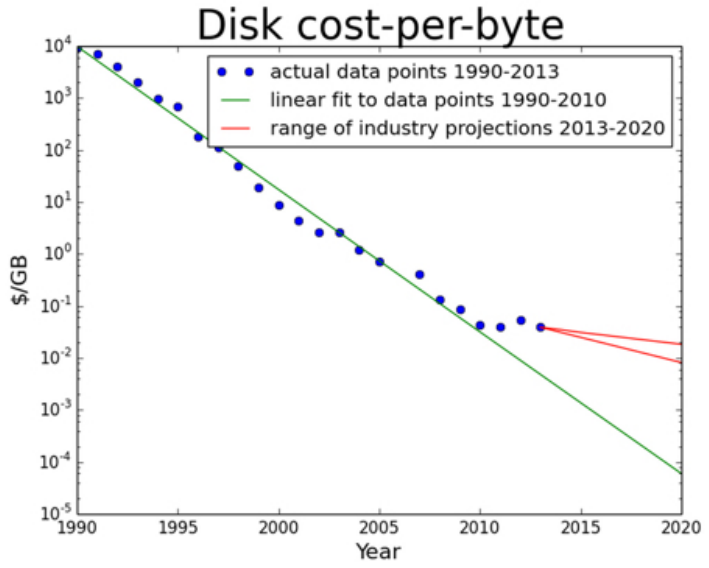
## Cost Trends

- Cost data can give trends and “Laws” or “rules of thumb” that are very powerful tools

# Effort and Use Knowledge Pyramid: for Costs

# Costs Rules of Thumb (1)

- Rules for a prolonged but not eternal period of time ("Laws")
- "Kryder's Law" – disk storage roughly halving in cost every year (comparable to Moores Law for processing power)
- A "re-set" in Kryder's Law from 2010 onwards documented by Rosenthal and Gupta




Kryder slowdown.
David Rosenthal.
Chart by Preeti Gupta at UCSC

# Costs: KRDS Laws/Rules of Thumb

1. Getting data in takes about Half of the lifetime costs, Preservation about a sixth, access about a third.
2. Preservation costs decline over time.
3. Fixed costs are significant for most data archives
4. Staff are the most significant Proportion of archive costs.

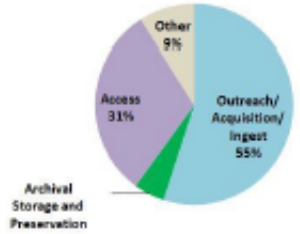
Note recent Dutch Digital Heritage Network research provides further independent validation of “KRDS Laws”

KRDS “Rules of Thumb”

***Getting data in takes about half of the lifetime costs, preservation about a sixth, access about a third***

KRDS found acquisition and ingest are the biggest costs over the preservation lifetime of research data. The costs of archival storage and preservation activities are consistently a very small proportion of the overall costs and significantly lower than the costs of acquisition/ingest or access activities for all the KRDS case studies.

Percentages varied between different archives but a consistent pattern emerged suggesting this rule of thumb from the Archaeology Data Service cost data as a rough guide to overall lifetime costs (Beagrie et al. 2010, pp. 31-52). It is potentially significant for those building business models and needing to fund archiving from depositor’s research grants. Ingest costs may be within the timespan of the research grant and can be a significant part of lifetime costs.




**Approximate Activity Data Costs for the Archaeology Data Service**
(after Beagrie et al. 2008). CC-BY licensed

***Preservation costs decline over time***

KRDS found a trend of relatively high preservation costs in the early years reducing substantially over time for data collections. An example is the preservation costs projected for the Archaeology Data Service (ADS) based on their experience of the first 10 years of operating the data service. (Beagrie et al. 2008, pp.4-6).This long-term decline in costs reflects a number of factors: partly the effect of Kryder’s Law on technical storage costs but mainly the growth in collections over time and the effect of economies of scale. Again it is potentially significant for those building business models, particularly if considering one-time fixed payment deposit fees or endowment for a dataset.

***Fixed Costs are significant for most data archives***

KRDS (Beagrie et al 2010, pp. 31-52) found that data archive costs are dominated by fixed costs that do not vary with the size of the collections. For most social science data archives, fixed costs such as core staffing and technical set-up will be significant.

Fixed costs are eventually not fixed but you have to scale up quite a way before that applies. Activities characterised by significant fixed costs can reduce the per-unit cost of long-term preservation by leveraging economies of scale. These factors may have implications for cost-benefit of small collections (as relative costs can be higher) and for collection policies (economies of scale, lower costs and higher impact may come from collecting in adjacent areas such as population health data or the humanities, or via international data collaborations such as CESSDA).

***Staff are the most significant proportion of archive costs***

KRDS consistently found that staff are the major cost component overall, sometimes as high as 90% of the total costs (Beagrie et al 2010, pp. 31-52). This finding was also made in another recent costs study (NCDD 2017). Equipment costs are a relatively small proportion of total costs. There is a minimum base-level of staff and skills required for any service. It is important to note that staff are the most significant component of fixed costs (see above) and economies of scale will be largely driven by staff costs and data volumes.

# Valuing Intangible Assets

# Valuing Intangible Assets

- Valuable approach to digital preservation and intangibles by Laurie Hunter and since adapted for research data
- We measure value of data services not just data alone
- Measuring value of intangible assets is hard – much harder than for physical assets
- Economic methods are well established but difficult to get the cost and value data to use in them
- Counter-factuals – a baseline – are important
- Collaboration with John Houghton to move beyond qualitative value to financial measures of value

# Tangible and Intangible Assets

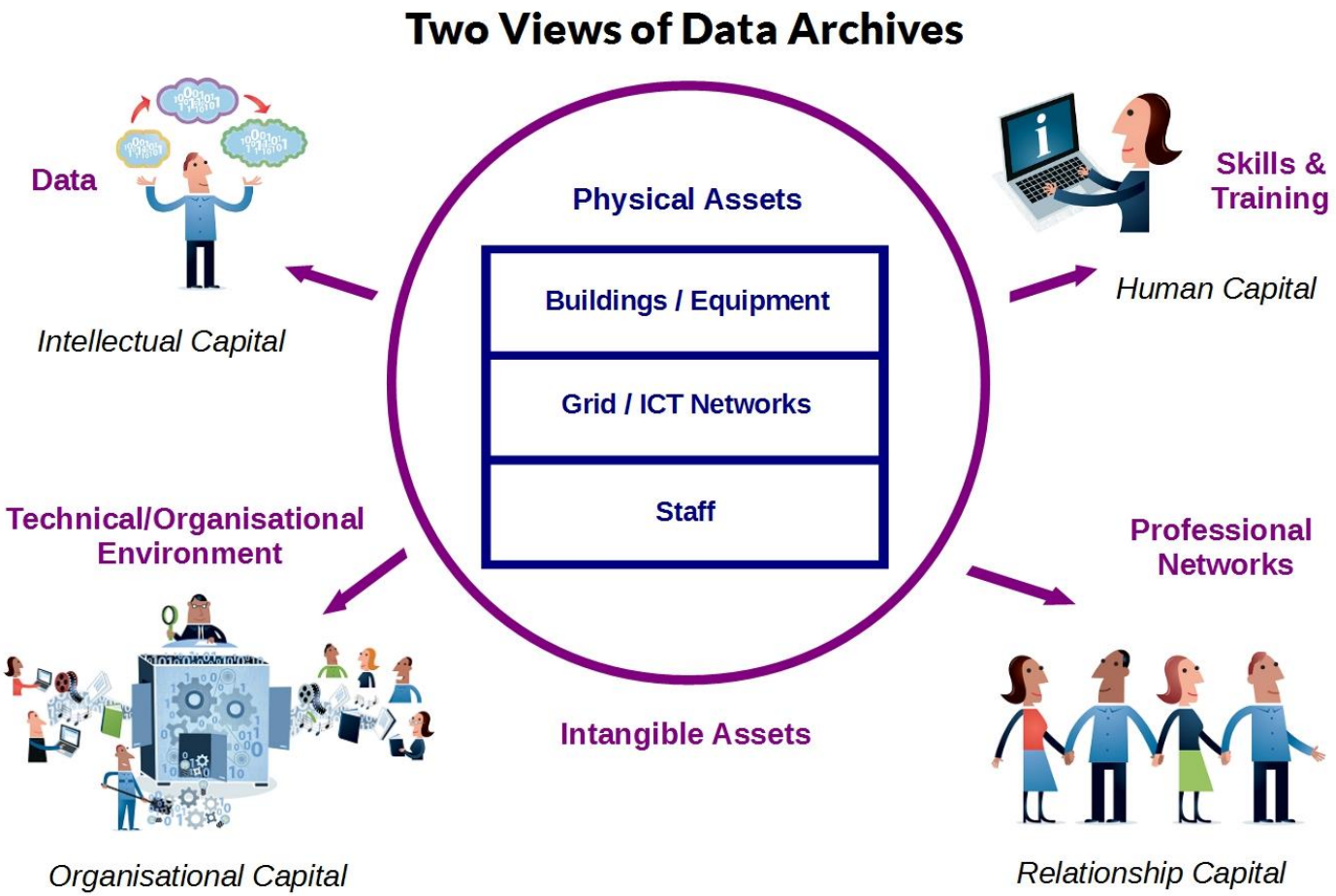



Illustration by Charles Beagrie Ltd ©2016 incorporating images by Jorgen Stamp digitalbewaring.dk. CC-BY licensed

# Value + Economic Impact Analysis

John Houghton (Victoria University) + Neil Beagrie (Charles Beagrie Ltd) 4 joint studies to date. Methods applied to:

Economic & Social Data Service (ESDS)

Archaeology Data Service

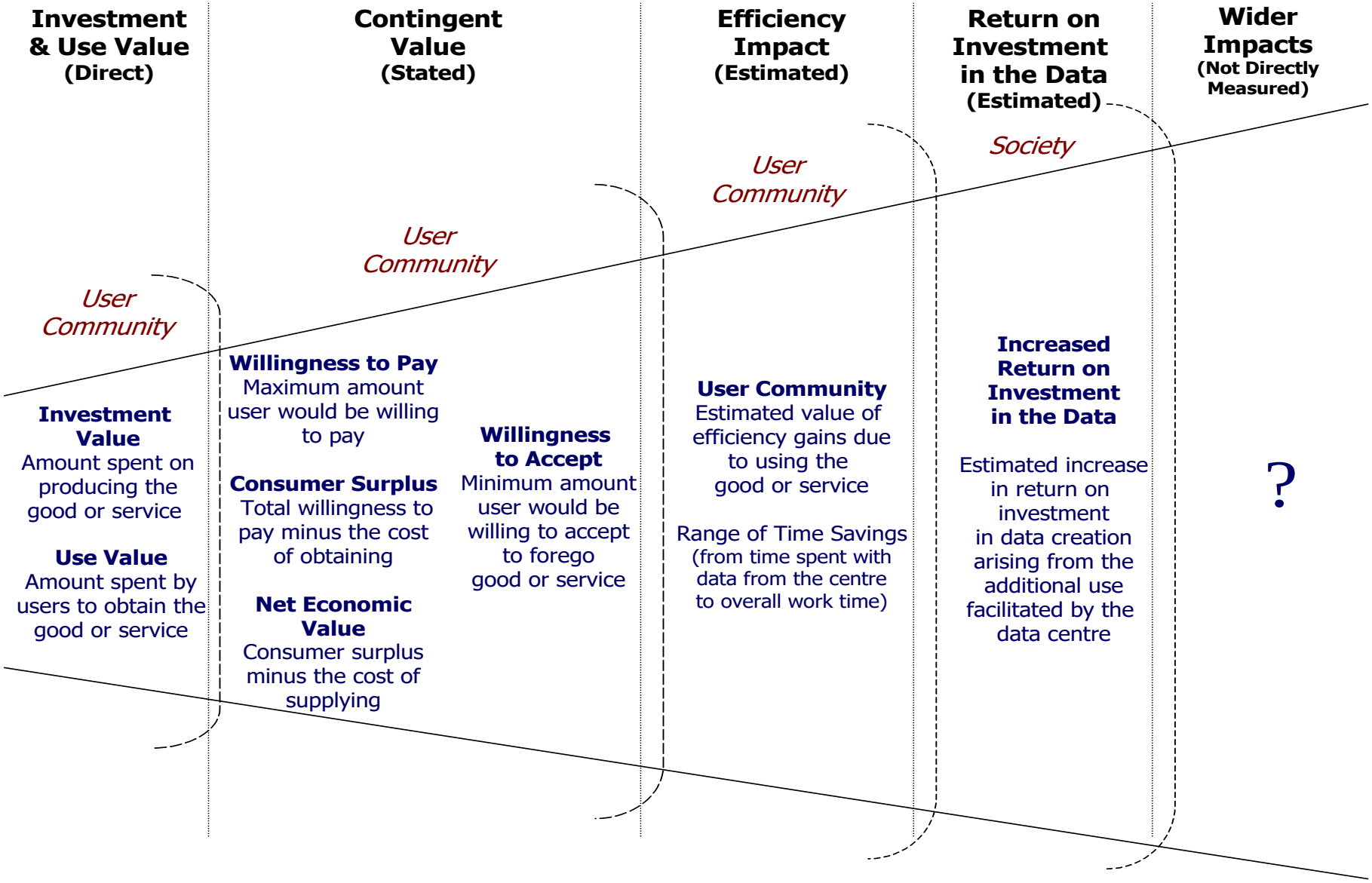
British Atmospheric Data Centre

European Bioinformatics Institute

# Economic Metrics Used

- **Investment value**: annual operational funding plus the costs that depositors face in preparing data for deposit and in making those deposits
- **Use value**: weighted average user access costs multiplied by the number of accesses
- **Contingent value**: the amount users are "willing to pay“ for or “willing to accept” in return for giving up access
- **Efficiency gain**: user estimates of time saved by using the Data Service resources
- **Return on Investment in the data service**: standard ROI
- **Return on investment in the data creation**: the estimated increase in return on investment to the funder(s) in the data creation due to the additional use facilitated by the data service

# Economic Methods Applied

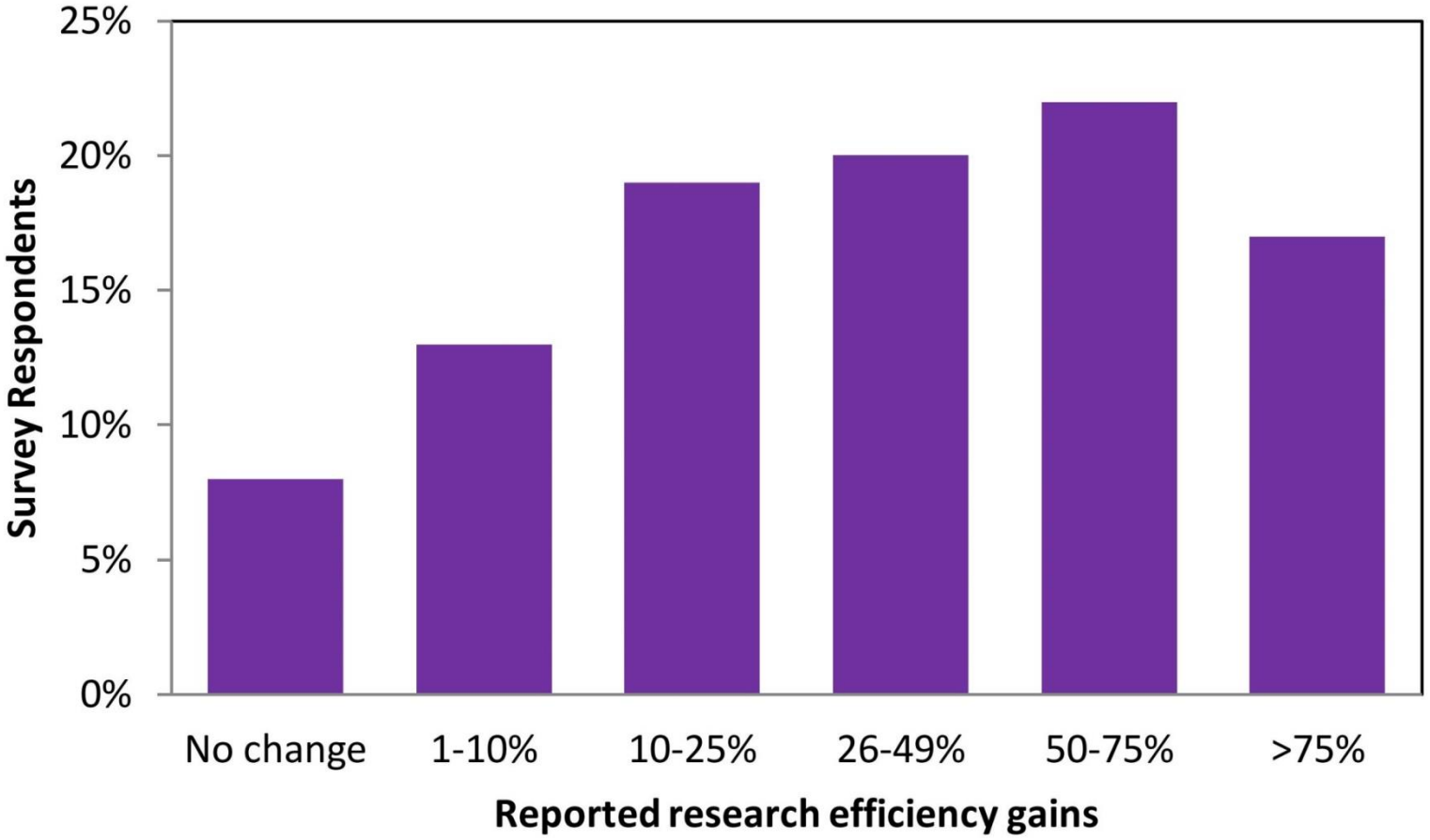


# ESDS Value/Impact Analysis

Benefit/cost ratio of net economic value to ESDS operational costs

**£5.40 to £1**

# ESDS Study: Researcher Efficiency Gains




Impact of using ESDS data and services on research efficiency

(after Beagrie et al 2012, p77, Figure 15)

Economic and Social Research Council © 2012 CC-BY licensed

# Counter-factuals – "Costs of Inaction"

**"Ideally, economic impact assessments should estimate the counterfactual – i.e. what would occur in the absence of the facility…However, counterfactuals are rarely addressed in the [c.100] studies reviewed due to lack of data. We found two exceptions that address this issue partially. One is the evaluation of the economic impacts of ESDS (2012) which partially explores the counterfactual through a users' survey…Another exception is a review of economic impacts of large-scale science facilities in the UK (SQW, 2008) … however, this estimation is not done rigorously and relies mostly on the estimation of the local benefits."**

***Big Science and Innovation - Report to BIS - Technopolis 2013***

# Costs of Inaction

| Costs of Inaction: reported metrics for archiving via individual researchers | | | |
|---|---|---|---|
| **Absolute loss** | Rate of loss of research data sets | 17% per annum | (Vines et al 2014) |
| **Partial information loss** | Rate of loss of working contact emails | 7% per annum | (Vines et al 2014) |
| | Rate of loss for web-links to data on personal websites | c.5.5% per annum | (Pepe et al 2014) |
| **Access** | Data requests fulfilled | 25.7%<br>44%<br>59% | (Wicherts et al 2006)<br>(Krawczyk and Reuben 2012)<br>(Vines et al 2013) |
| **Delay** | Elapsed time to fulfill data requests | Up to 6 months<br>Within 1-3 weeks<br>(mean 7.7 days) | (Wicherts et al 2006)<br>(Vines et al 2013) |

Illustration by Charles Beagrie Ltd ©2017. CC-BY licensed

Although these reported metrics are from studies of different disciplines and study dates, they contrast sharply with the excellent preservation record, very high fulfilment rates, and rapid online access rates of public data archives in the social sciences. The public data archives also are appreciating as opposed to

depreciating assets with improving rather than decreasing trends in value over time.

# Where Could We Go From Here?

# What to Keep

## Recent Jisc research data study

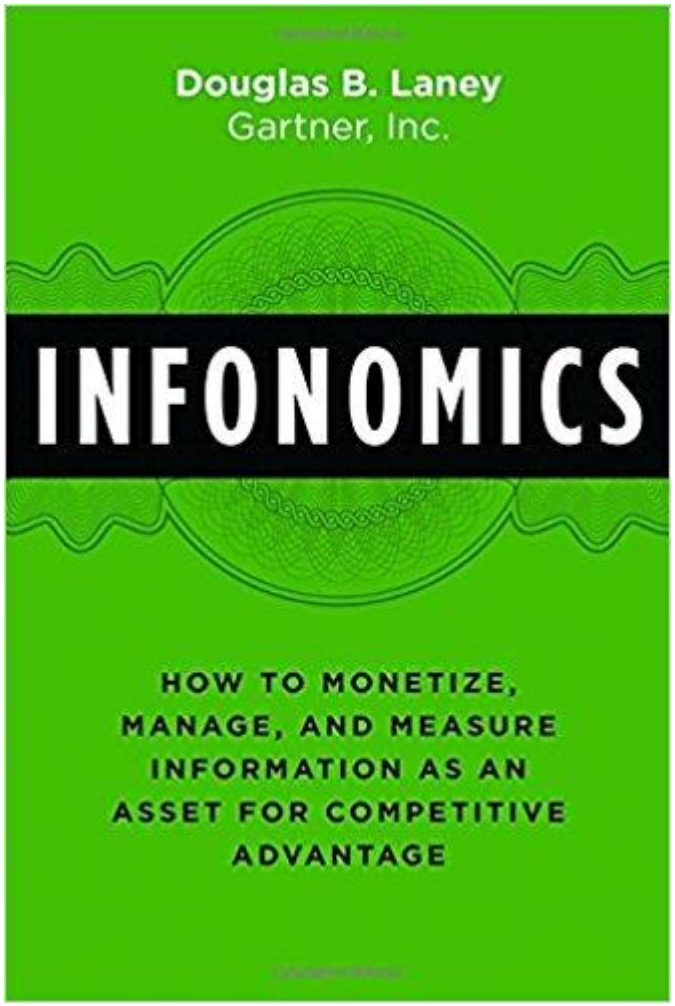
# Recommendations

- **Recommendation 4**: Investigate the relative costs and benefits of differential curation levels, storage, or appraisal for what to keep for the two major use cases (Research Integrity, and Reuse) identified in the study.

# Levels of Curation

**US National Science Board 2005 Long-lived Data Collections**

two-tier system with differential curation levels used by the UKDS or the DANS data archive's systems - DataverseNL for short-term data management (up to 10 years) and EASY for long-term archiving, in the Netherlands. Both these examples in the UK and the Netherlands have different time horizons (how long the data is kept), costs in terms of metadata and preservation care (how it is kept) for their two systems, with the option to move from short-term to long-term systems and curation levels after future appraisal (or alternatively be maintained in their existing short-term system/ or deleted).



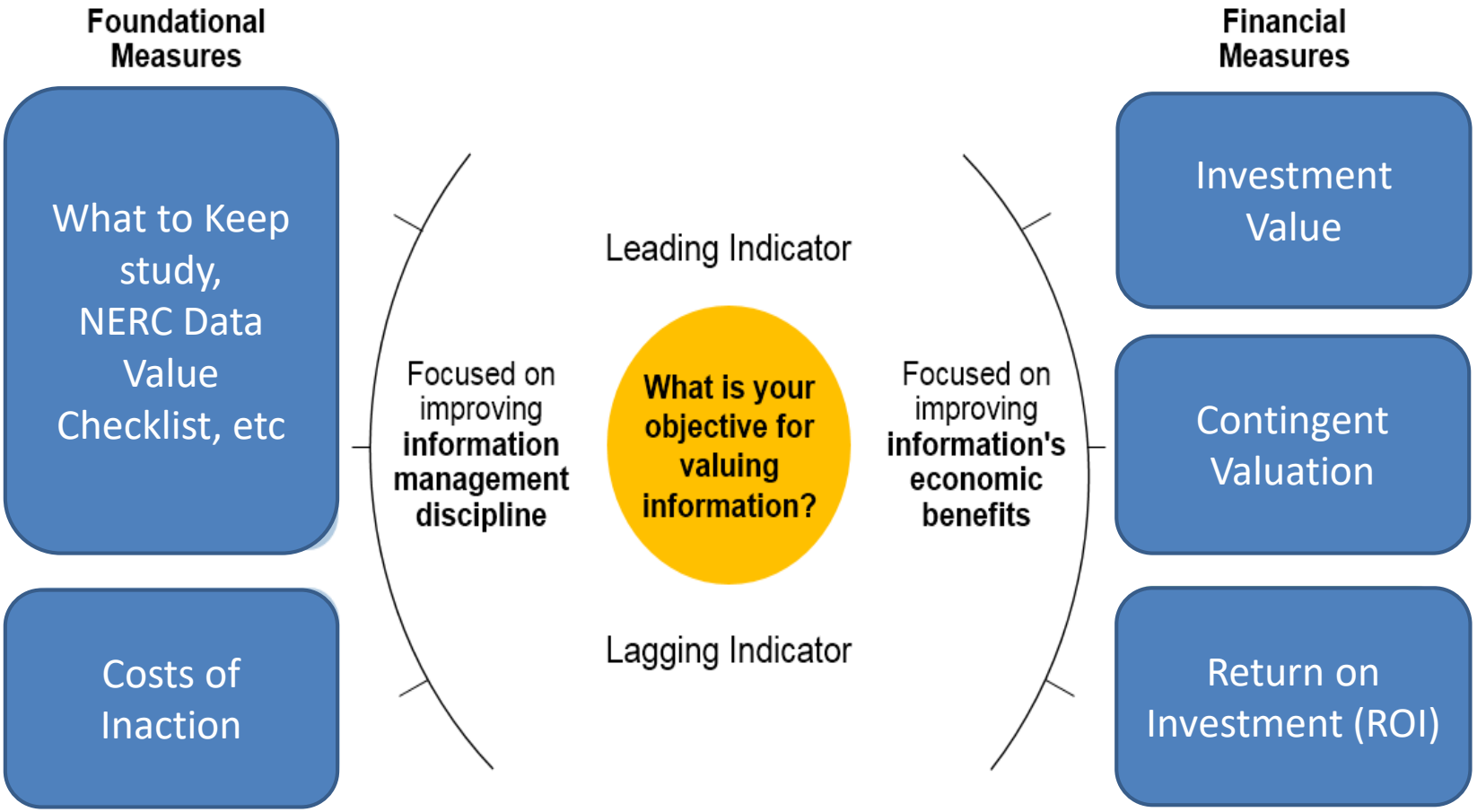

# An industry-centric view of the value of information

# Accounting for Information

## Some Infonomics Quotes

- “Five or six decades since the beginning of the Information Age, the namesake of this age, and the major asset driving today’s economy, is still not considered an accounting asset”
- “Corporations typically exhibit greater discipline in tracking and accounting for their office furniture than their data”
- Bottom line - Data stewards are not alone in seeing this as an anomaly. There are others pressing for changes to insurance and accounting practices.

# Gartner's Information Valuation Models




From *Why and How to Measure the Value of Your Information Assets* by Douglas Laney



10 CONFIDENTIAL AND PROPRIETARY I © 2016 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner and ITxpo are registered trademarks of Gartner, Inc. or it's affiliates.

# Conclusions

- We can use collections of cost data to look for trends – rules of thumb are probably the most widely useful cost information
- “Datanomics” and “Infonomics” have synergies - we may be able to leverage efforts within our community and industry
- Need to investigate the relative costs and benefits of differential curation levels, storage, or appraisal for the two major use cases (Research Integrity, and Reuse) identified in the What to Keep study.
- We have HSM in IT – in time can we look towards automating some decisions as Hierarchical Curation Management?

# Further Information

- Costs, Benefits, and ROI for Research Data
  - CESSDA SaW Cost-Benefit Advocacy Toolkit, http://dx.doi.org/10.18448/16.0013
- Economic Impact Studies of Research Data Services
  - The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf
- Douglas B. Laney 2017 Infonomics: How to Monetize, Manage, and Measure Information as an Asset for Competitive Advantage
  - ISBN-13: 978-1138090385