

# Research support and data management challenges: Is there such a thing as too much data?

San Cannon  
Kansas City Fed



FEDERAL RESERVE BANK *of* KANSAS CITY

(standard disclaimer applies)

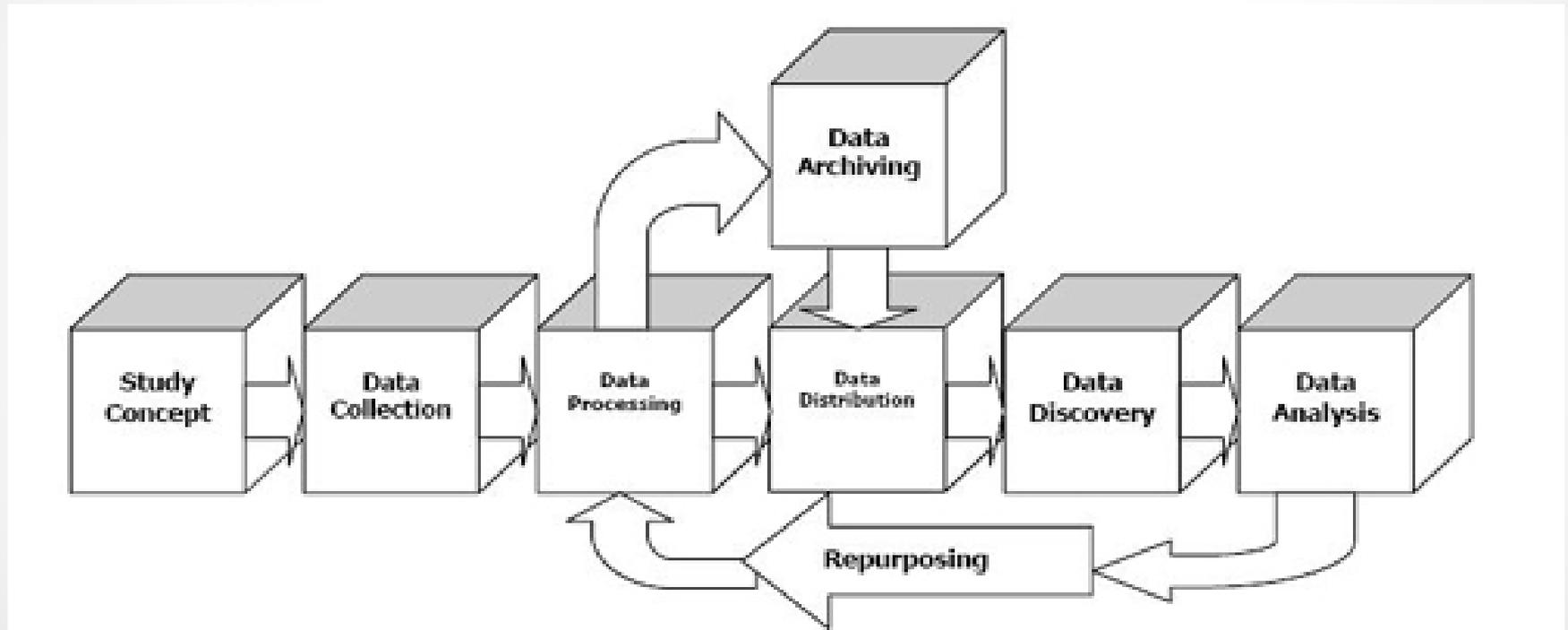


# More is always better, right?

- Past: carefully crafted studies collect information used to test hypothesis using tested statistical methods.
- Present: millions of transactional or administrative records are “mined” to see what patterns emerge.
- Is this research?
- What does it mean for research support?  
Computing? Data management?



# Data Life Cycle



Source: Old DDI diagram still available from MIT libraries site.



# Beginning

- First stages: formulating the question THEN gathering the data. Right?
- There are questions that get asked based on theory or preconceived hypothesis.
- Then there are questions that get asked because they can be answered.



# Playing with data?

- My story: for my thesis I had a dataset with 4.5 million records.
- Firm level data: information on number of employees, location, payroll.
- I played with the data to see what could be done then wrote an empirical paper with two supporting theoretical papers.
- I was “accused” of data mining.



# Accusation to acceptance

- In 2008, Google showed that they could predict flu outbreaks faster than CDC.
- Because the CDC estimates were bad or because they had the data?
- Now more than 900 articles on Google scholar related to GFT.
- There are at least 5 sessions at ASSA 2015 concerning data including one on machine learning in economics.

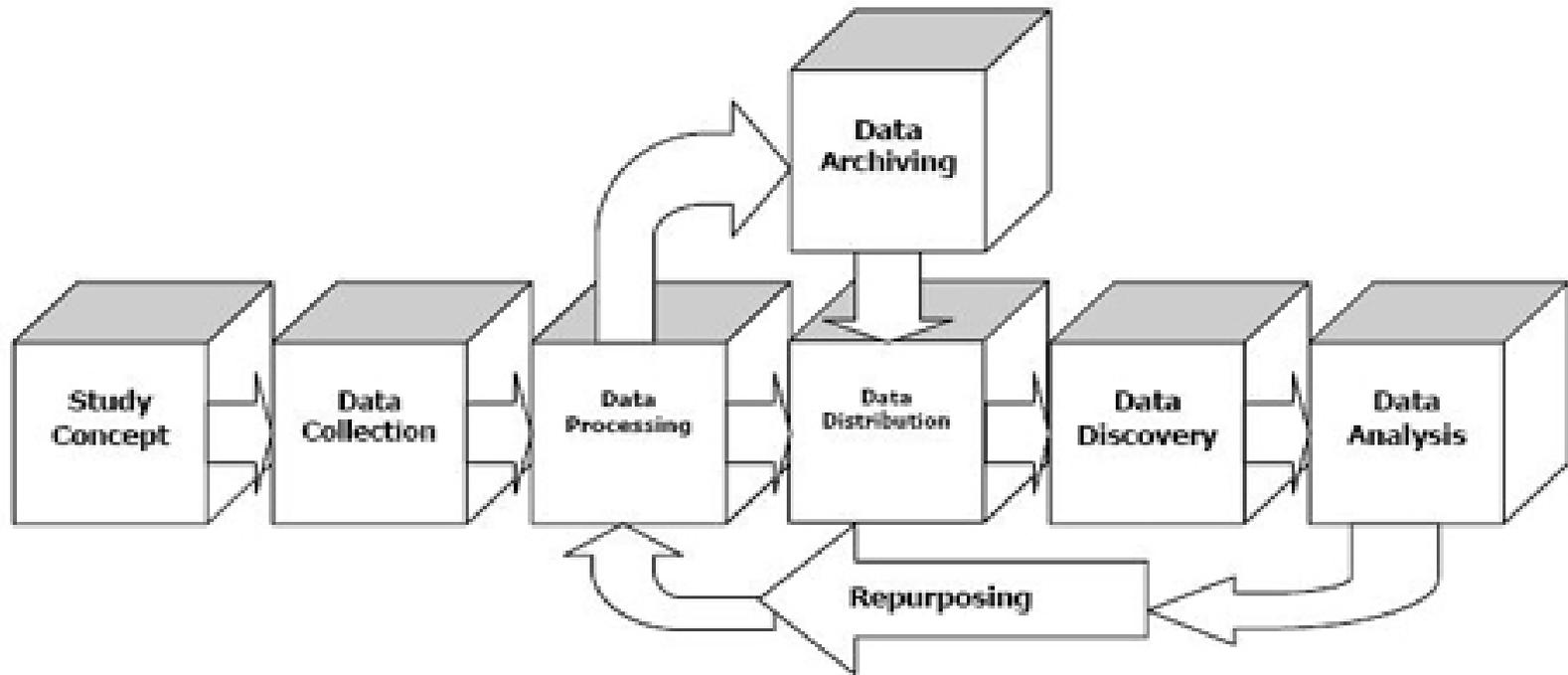


# Chicken and egg

- As more data are available through more data sources, it is harder to disentangle the first two lifecycle stages.
- Is that bad?
- Does “playing with the data” equate to scholarly research?
- Commercial data usage doesn't seem to care.



# Data Life Cycle



Source: Old DDI diagram still available from MIT libraries site.



# Data collection

- Is it now really data acquisition or data choice?
- Now data are commercially or openly available from a variety of sources
- Comparing carefully designed data collection (small) with data captured with a design concept (administrative records, sensor data, internet “exhaust”)
- What are the implications?



# Data compilation

- What about combining data sets?
- Traditional (small) data often couldn't be efficiently combined because of differences in study design.
- More often data are now being combined with other data with potential for unintended consequences
- The "mosaic effect": big picture from smaller pieces may not have been seen previously



# Data management

- Traditional (small) data are relatively easy: limited number of variables or observations, uses traditional data scrubbing or quality measures, variety of storage options
- New (big) data are different: many variables or observations, little or no data scrubbing or quality checks employed, storage options more limited



# Data analysis

- Traditional approach: econometrics, sociometrics, biometrics, statistics
- New approach: predictive analytics and data mining
- Difference is goal: explanation versus prediction.
- Correlation vs. causation – where is one more relevant than the other? When is it okay to only answer the “what” without the “why”?

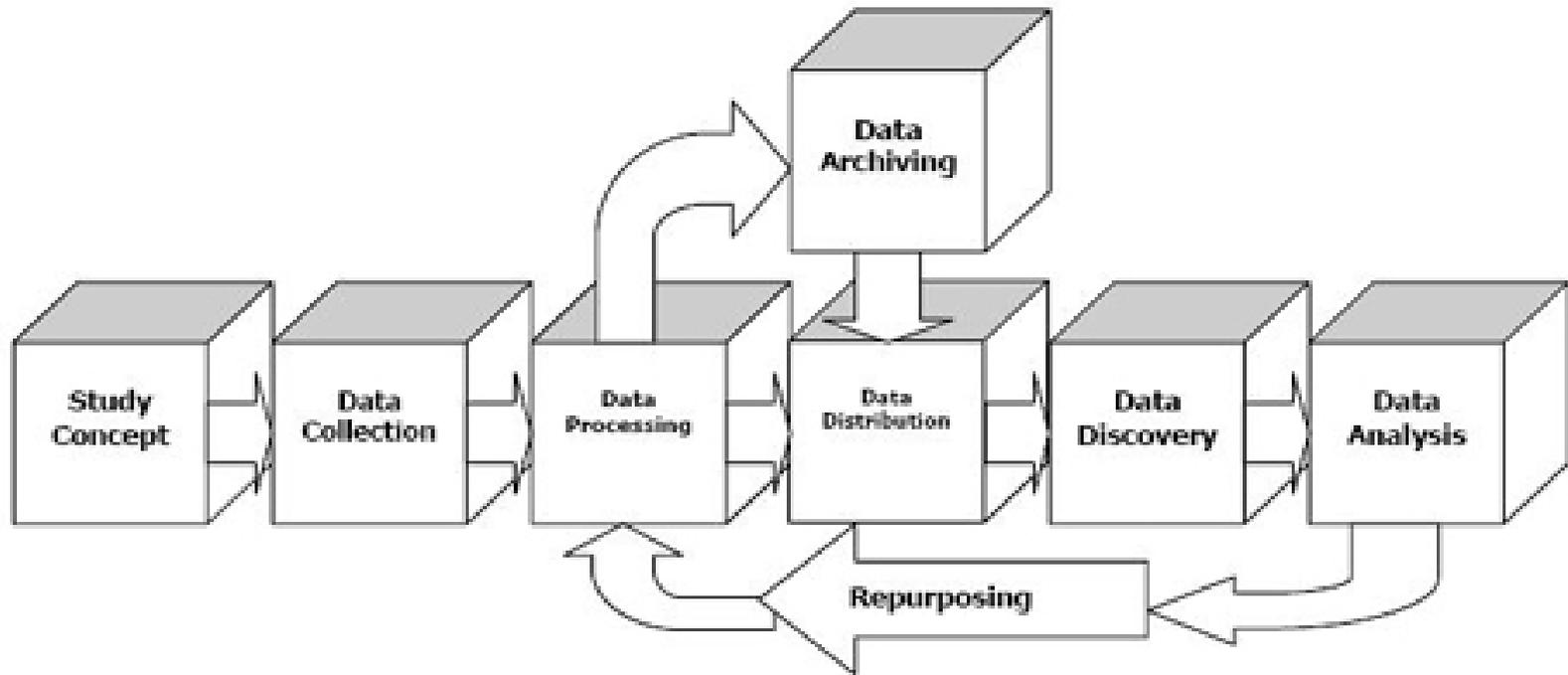


# Technology and infrastructure

- New types of data (quality and quantity) require new storage and retrieval tools
- New (to us) analytical techniques may require new software or programming approaches
- Lots of new things are likely to have to intersect with lots of old things – interoperability challenges abound.



# Data Life Cycle



Source: Old DDI diagram still available from MIT libraries site.



# Archiving

- How to store stuff that is soooooo big?
- Curation becomes more challenging with more than one of the Vs (volume, velocity, variety)
- Consider – the Library of Congress is archiving EVERY tweet.
- MARC record for 140 characters of text?
- Other traditional metadata fields may also not fit new data types well.



# Governance

- Or can you store it? Or use it in a particular manner?
- Many commercially available data products have restrictions on how data may be preserved or stored.
- “Free” data from the web often comes with restrictions that no one notices.
- Currently there is no universal legal opinion on the validity of “browse through” contracts but....



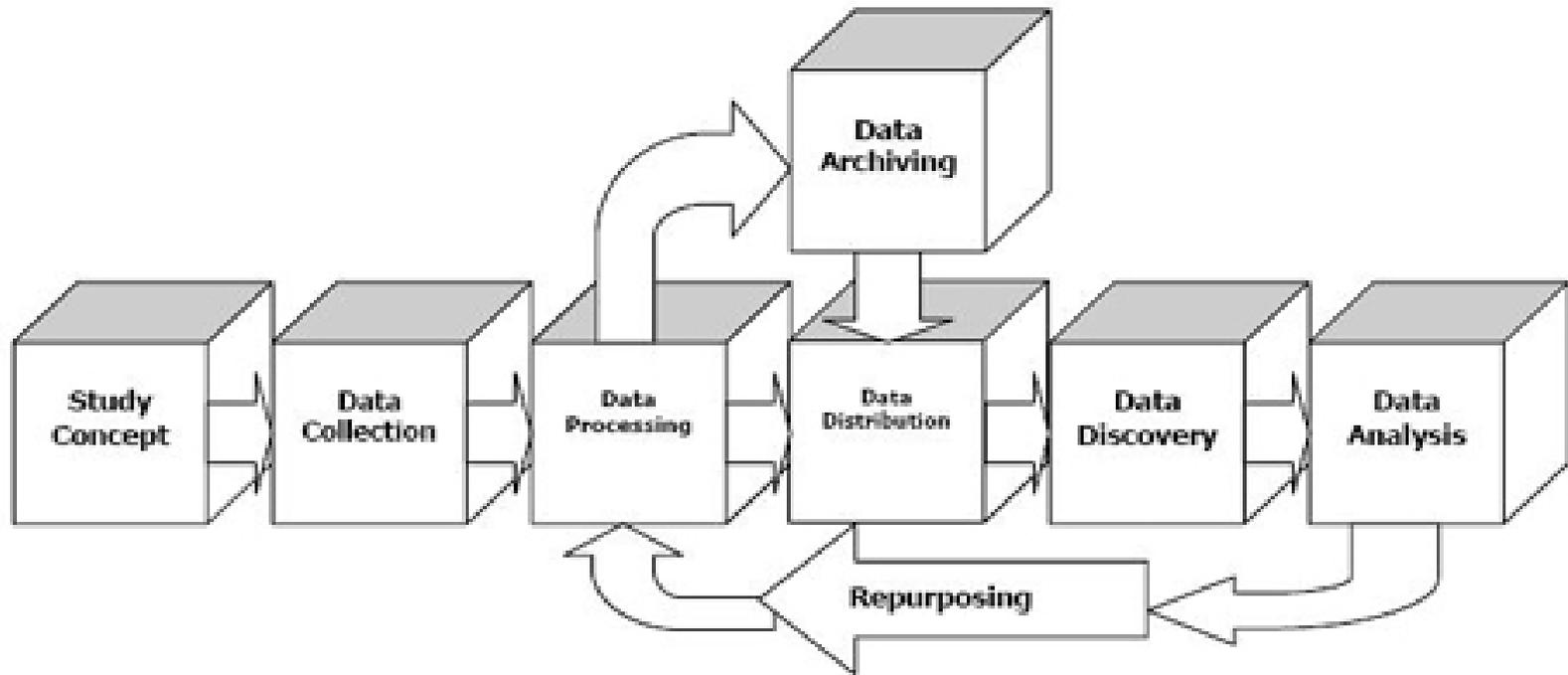
# Fine print

- “You further acknowledge and agree that, unless NYSE Euronext, its applicable affiliate, and/or the applicable Third Party Provider give you prior written permission, you **will not sell, license, rent, modify, print, copy, reproduce, download, upload, transmit, distribute, disseminate, publicly display, publicly perform, publish, edit, adapt, compile or create derivative works** from any Content or materials (including, without limitation, through framing or systematic retrieval to create collections, compilations, databases or directories) or otherwise transfer any of the Content to any third person (including, without limitation, others in your company or organization).”<sup>1</sup>

<sup>1</sup> NYSE website <http://www.nyx.com/terms-use> accessed 4/12/14



# Data Life Cycle



Source: Old DDI diagram still available from MIT libraries site.



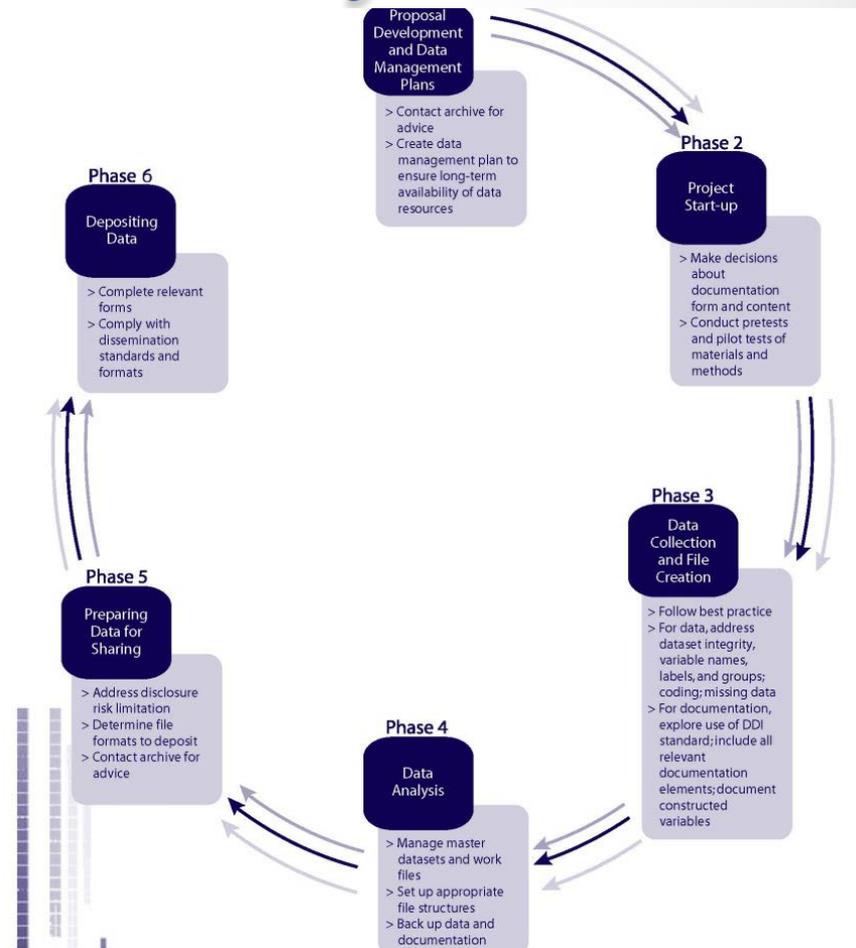
# Distribution and sharing

- At the heart of the data management plan (DMP)
- Sharing data has some of the same sticking points as archiving: size and restrictions.
- National Science Foundation guidance says lots of things “will be determined by the community of interest through the process of peer review and program management.”



# New data life cycle

- Phase 1: Proposal development and metadata plans
- Phase 2: Project start-up
- Phase 3 – Data collection and file creation
- Phase 4 – Data analysis
- Phase 5 – Preparing data for sharing
- Phase 6 – Depositing data



# Suggestions?

- I've raised lots of questions without offering any answers.
- Some I don't know yet; others no one knows; yet others there is no agreement upon
- What I do know is that there is the potential for new types of research support and data management activities
- Maybe not new to everyone but many are not consistently available



# “Creative” services

- Development of data management plans
- Technology consulting
- Data librarianship
- Data archiving
- Data citation assistance
- Ontological development and linking services
- Data dissemination/sharing preparation



# Moral of the story

- Lots of data is the new normal.
- It is “big data” if it’s bigger than you are used to.
- It is “too much” if you don’t know what to do with it.
- With consideration, careful planning, and creativity, big data doesn’t have to be too much.



# Last slide

Thanks for listening!

[sandra.cannon@kc.frb.org](mailto:sandra.cannon@kc.frb.org)

(816) 881-2596

