RESEARCH DATA ALLIANCE

Summary
Data Practices Report

Peter Wittenburg
Max Planck Data & Compute Center
former MPI for Psycholinguistics

research data sharing without barriers

rd-alliance.org

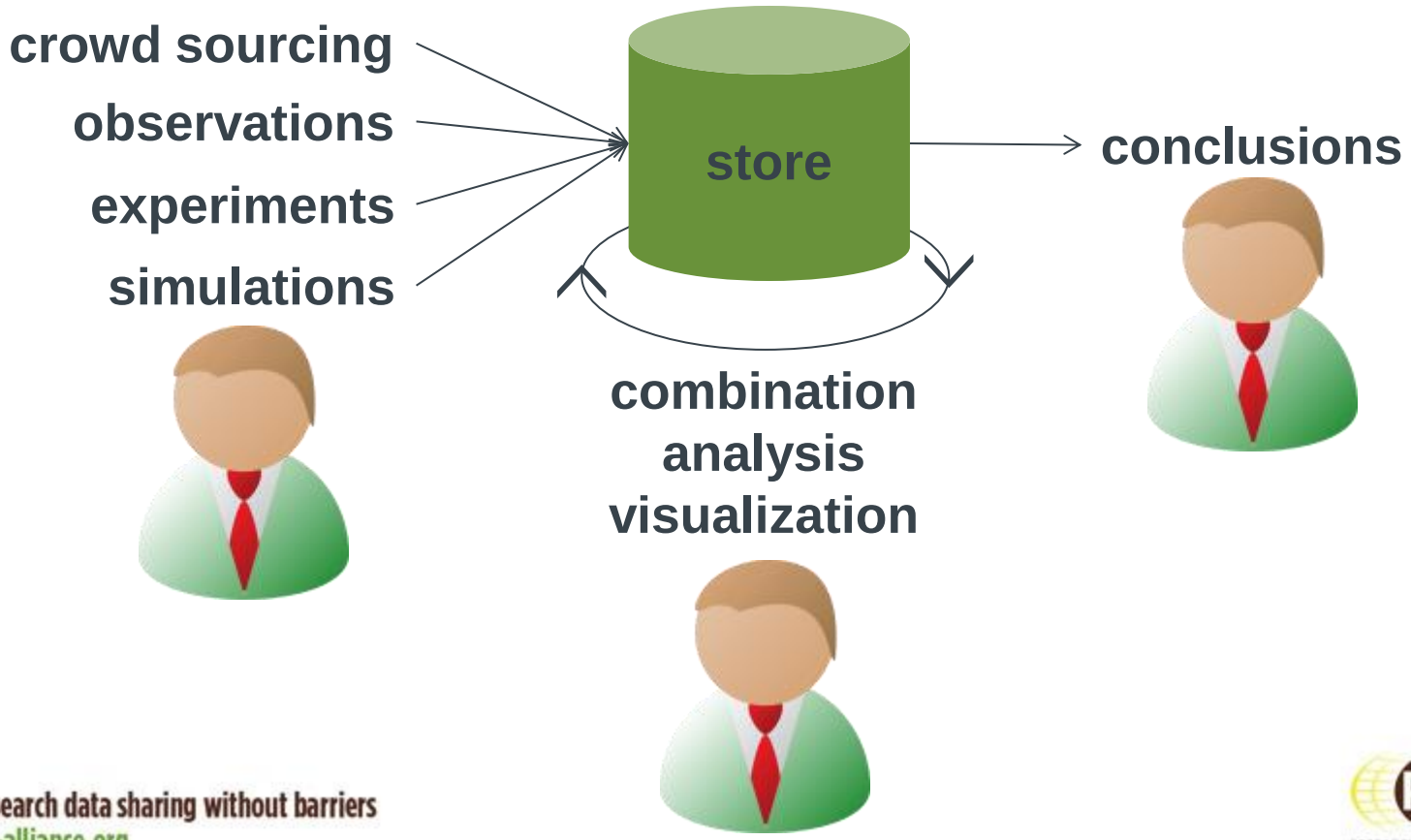
Topics for Data Science

- Relevance of Data
- Trends in Data Domain
- Requirements for Data Domain
- Data Practices

Why talking about data?

data is the oil driving research and economy

data is key to understanding big challenges



Why talking about data?

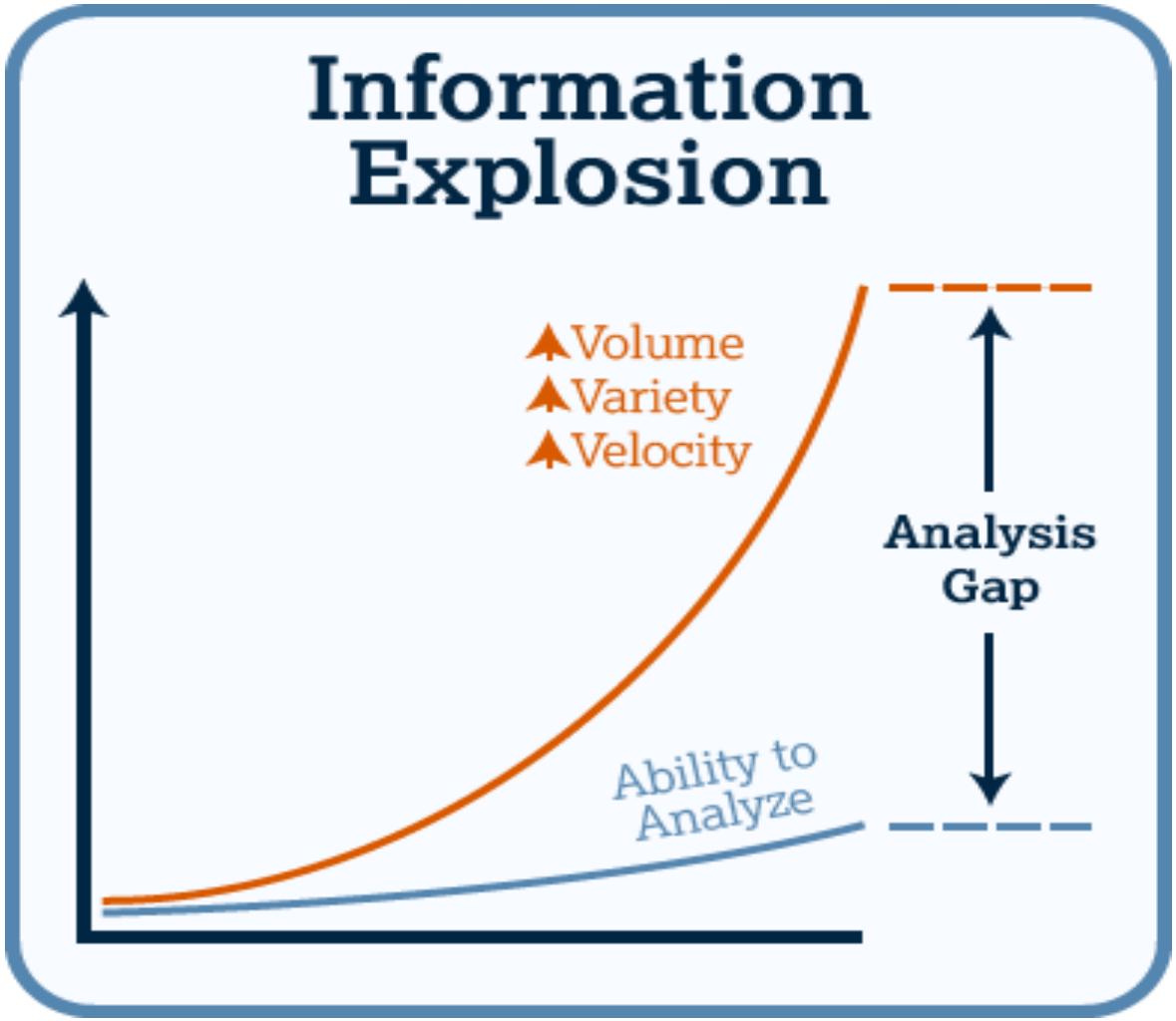
data is the oil driving research and economy

data is key to understanding big challenges

if this is true for most disciplines ...
... can we observe trends?
... can we specify requirements?
... can we correlate with practices?

combination
analysis
visualization

Trends I – 3 Vs



well known:

- **Volume**
- **Variety**
- **Velocity**

Trends II – Complexity

from simple structures ...

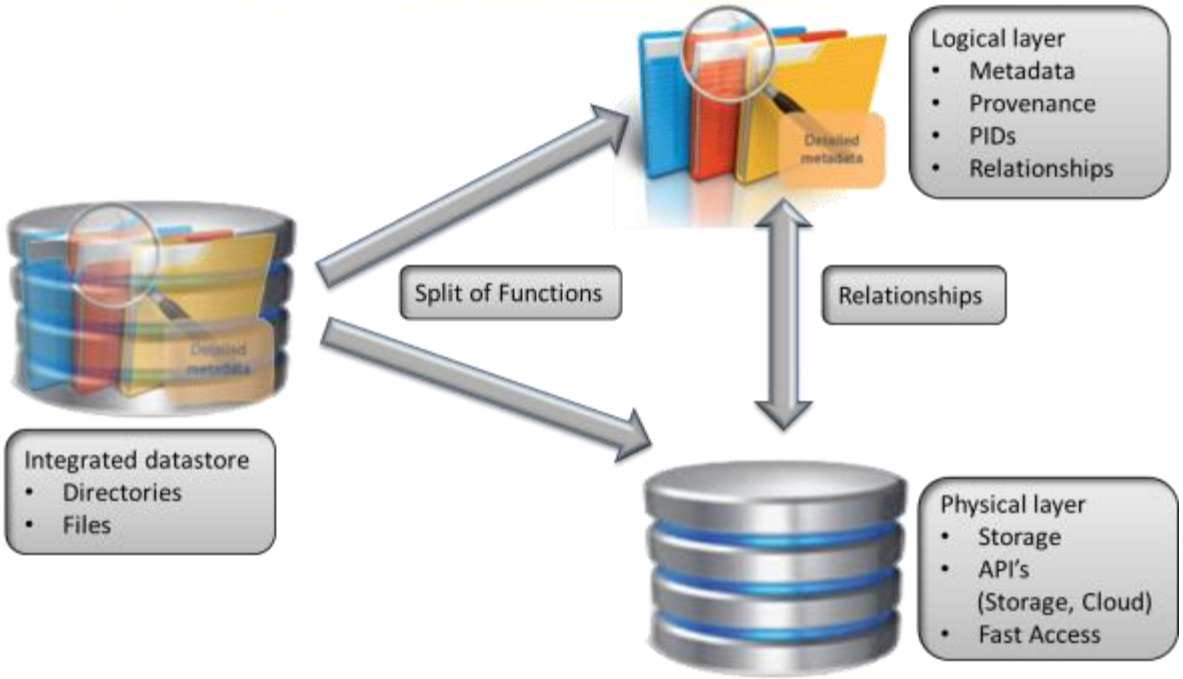
... towards complex relationships

not so often mentioned:

- lots of different relationships and dependencies
- relationships at data object level and at content level
- reproducibility gap

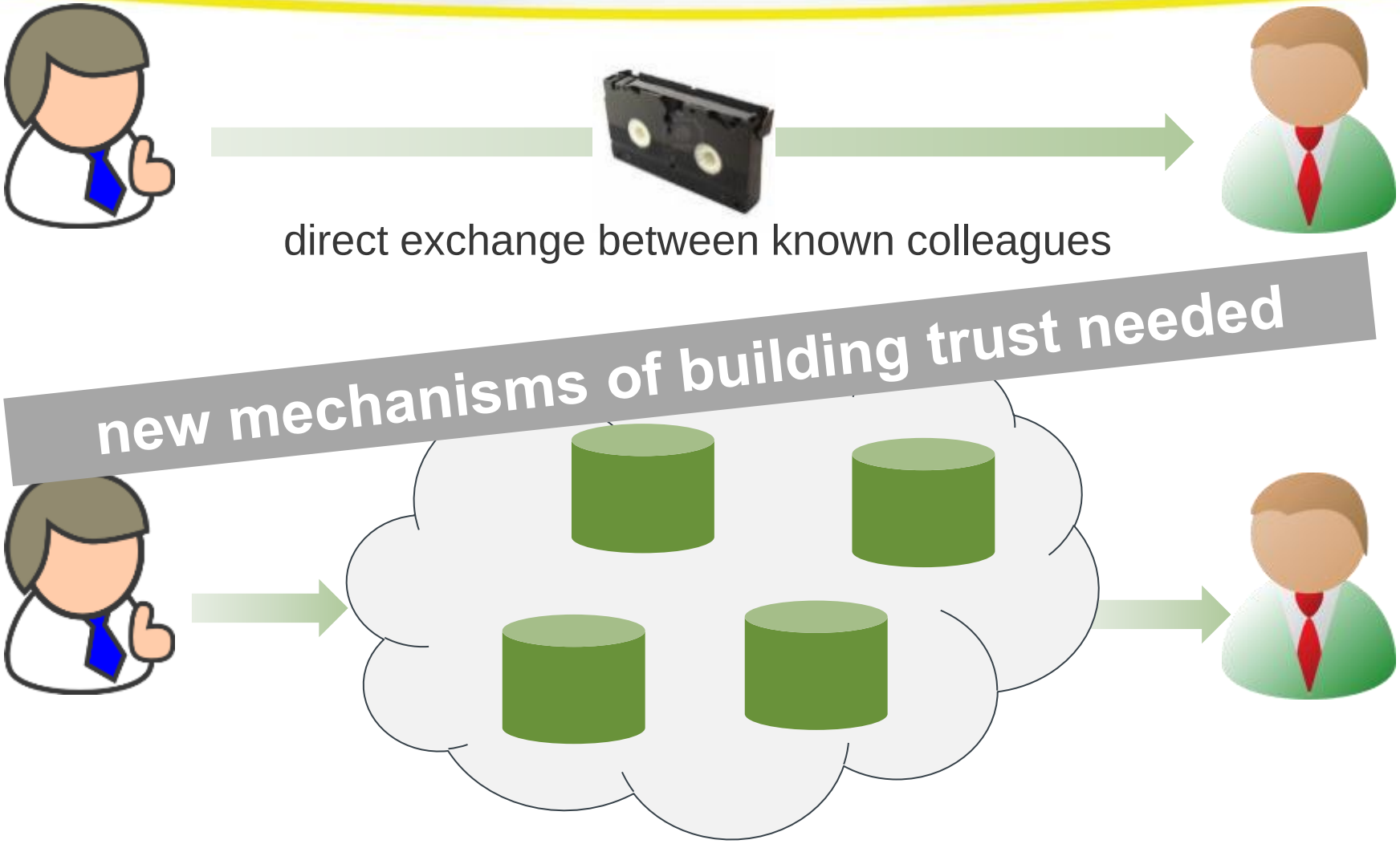
RDA RESEARCH DATA ALLIANCE

Trends III – Change in Approach

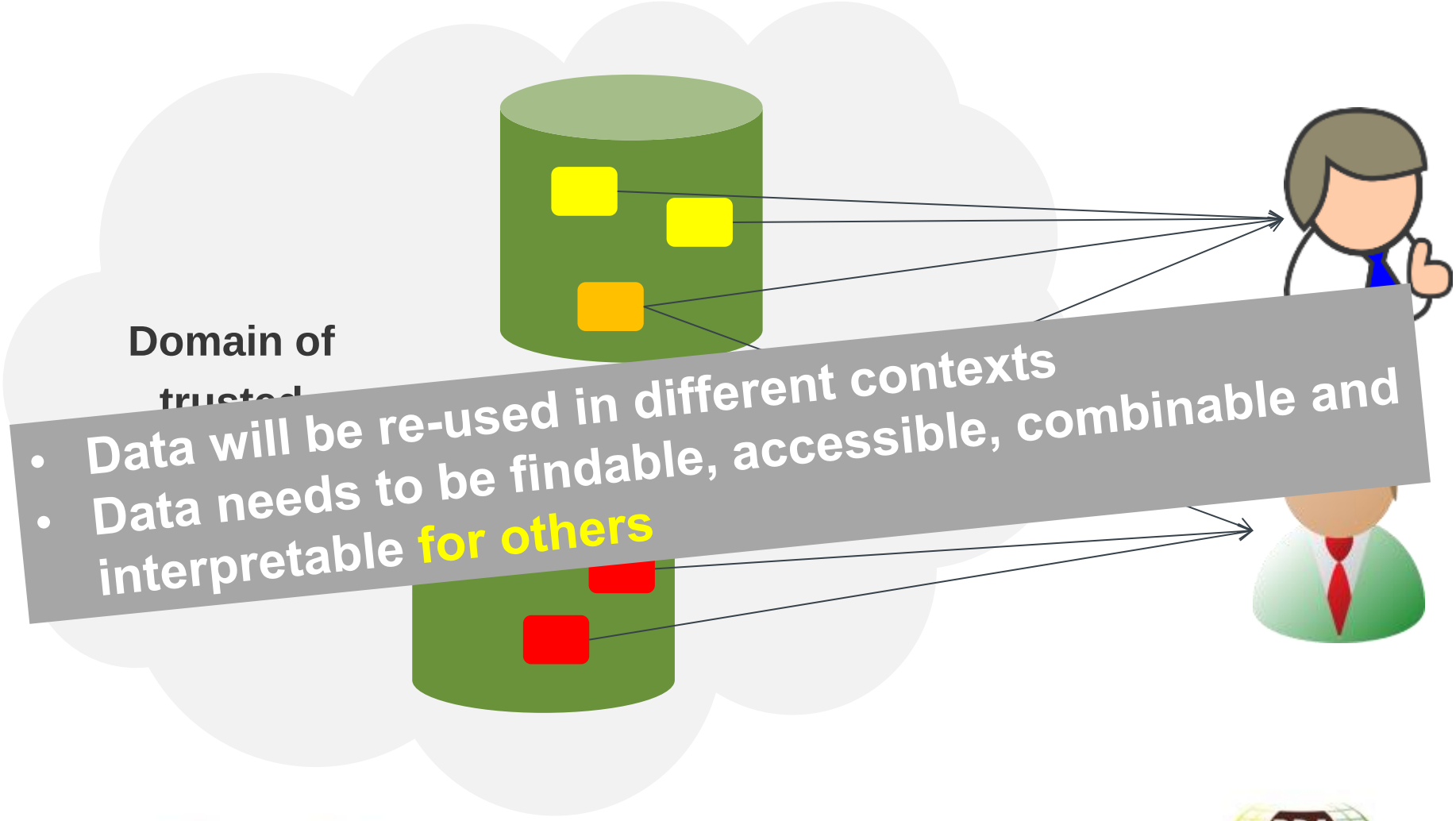


- split into **Physical Layer** to optimize performance (files, clouds, etc.)
- split into **Logical Layer** to optimize finding, tracing relationships, managing access, etc.

Trends IV - Anonymity



Trends V – Re-Usage



Trends VI – large federations

EUDAT example to

- **domain of registered data**
- **exchange data**
- **manage and preserve data**
- **bring data close to HPC and strong clusters**
- **offer improved data services**

- **all across disciplines and countries in EU**

Trends VII - large federations in Humanities

★ remote language archives
★ large data centers with copies

DOBES Example

- ~70 international teams collaborating + 1 archive
- changed culture in community

Requirements for Data Science

- let's use the G8 formulations – data should be
  - searchable -> create useful metadata
  - accessible -> deposit in trusted repository and use PIDs
  - interpretable -> create metadata, register schema and semantics
  - re-usable -> provide contextual metadata
  - persistent -> provide persistent repositories

- let's use the Knowledge Exchange formulations
  - promote data sharing norms and standards -> make all explicit
  - support services enabling effective sharing -> build infrastructure
  - provide a deposit infrastructure -> build infrastructure
  - develop flexible access mechanisms -> build infrastructure
  - allow linking between publications and data -> use PIDs and citation MD

Requirements for Funders

- Naoyuki Tsunematsu (Senior Advisor, JST)

  Question 1: what is the value proposition for publically funded research?

  - Publically funded research is about stimulating competitiveness
  - new strand for competitiveness:

    Knowledge Discovery based on smart data collection

- role of funders seems to change:
  - build infrastructures to make data visible and accessible
  - invest in human capacity

Requirements for Culture

- Naoyuki Tsunematsu (Senior Advisor, JST)

  Question 2: is culture of sharing a relevant issue?

  - need for data exchange (and thus the need for proper data management) are yet difficult to convey in Japanese Science
  - parallel trends observed for Japanese Science
    - not so often included in collaborations anymore
    - not so often represented in the top papers
    - decrease in international ranking
  - serious worries about counterproductive lack of openness

- this concern seems to be relevant for all of us

  G8 Open Data Report: UK 90, US/CA 80, FR 65, IT 35, **JP 30**, **DE 25**, RU 5

Data Practices I – Commons

- many obstacles for sharing (mostly well–known)
  - cultural, sociological -> technological
  - lacking widely used mechanisms
  - lack of proper mechanisms and incentives
  - research is about competition & collaboration
- even more obstacles for re-using
  - often doubted whether useful in particular across disciplines
  - Big Data claims are controversial
  - lack of trust, information, quality, etc.
  - too time consuming frequently

Data Practices I – Survey

- ~120 Interviews/Interactions
- 2 Workshops with Leading Scientists (EU, US)

- too much manual work or via ad hoc scripts
- still creating huge amounts of legacy formats

  (no PID, MD, lack of explicitness)
- there are positive project examples etc. but ...
  - DM and DP not efficient and too expensive

    (Biologist for 75% of his time data manager)
  - federating data incl. logical information much too expensive
  - hardly usage of automated workflows and lack of reproducibility

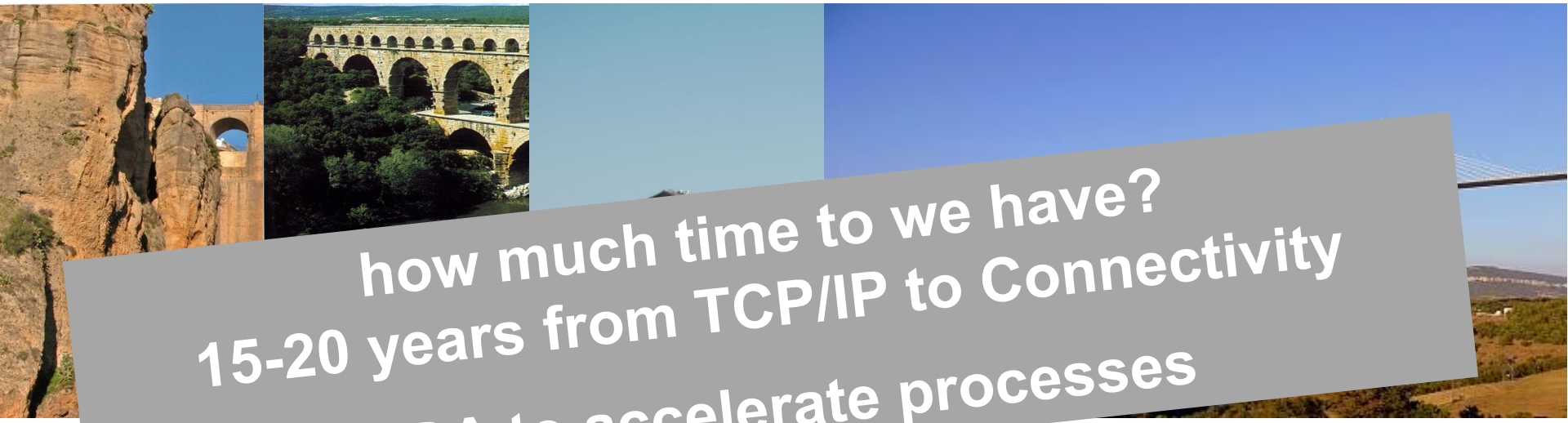
Data Practices I – Survey

- ~120 Interviews/Interactions
- 2 Workshops with Leadi

- many researchers **can't participate**
- pressure towards DI research is high, but only some departments are **fit for the data challenges**
- **Senior Researchers: can't continue like this!**
  - need to move towards **proper data organization** and **automated workflows** is evident
  - but changes now are risky: **lack of trained experts, guidelines and support**

- incl. logical information much too expensive
  - hardly usage of automated workflows and lack of reproducibility

Promiss of RDA



how much time to we have?
15-20 years from TCP/IP to Connectivity
RDA to accelerate processes

*about building the **social and technical bridges** that enable global open sharing and re-use of data. **Researchers, scientists, data practitioners** from around the world are invited to work together to achieve the vision*

*Funders: NSF, EC, AU, Japan, Brazil, **DE?, UK?, ZA?, FI?, etc.***

Thanks for your attention.

http://www.rd-alliance.org
http://europe.rd-alliance.org

Next RDA Plenary P6: 23-25. September Paris

Data Practices III – DOBES Example

- ~ 70 teams, ~ 100 languages
- DOBES changed culture
  - researchers agreed to deposit and share **pre-final versions**
  - researchers agreed on basic **access mechanisms** and layers
  - researchers started with **cross-language analysis** work using collections
  - agreement on **data structuring and metadata principles**
  - researchers used broadly **modern technology**
  - archivists established **archiving principles and technology**
  - technology all built on **Open Standards**
  - result is a large online data archive with many open resources
- all fine?
  - NO - could not solve **archive sustainability** issue yet in a smooth way

Summary

- science is changing and data accessibility will be crucial
- trends of working with data requires culture change
- currently working with data is inefficient, expensive and mostly non-reproducible
  - data intensive science is limited to power institutes
- essential components must be persistent
- connecting nodes cost 15-20 years
- RDA founded to accelerate towards efficient data science
- still RDA is a very young initiative

  it is a chance