The Digital Object Architecture and the Enhanced Robust Persistent Identification of Data

Rob Quick

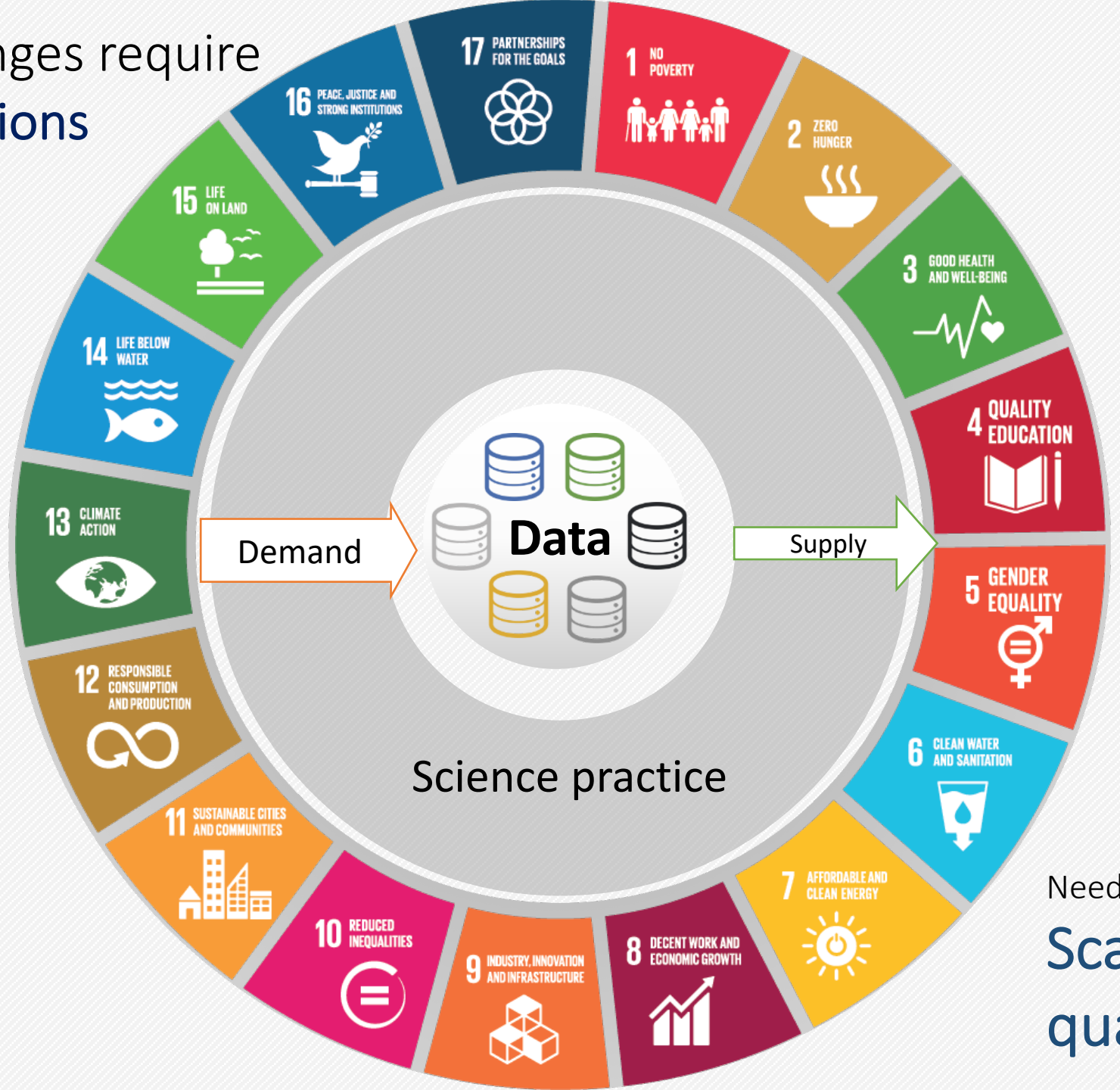
With Slides Contributed by Peter Wittenburg, Dimitris Koureas, George Strawn, Beth Plale & Larry Lannom

Rob Quick rquick@iu.edu @robquick5

Associate Director, **Science Gateways Research Center**
PI, **Enhanced Robust Persistent Identification of Data (ERPID) Testbed**
ECSS L3 Manager, **Science Gateways**
Co-Chair, **RDA Data Fabric Interest Group**

Our grand challenges require

Data-driven solutions



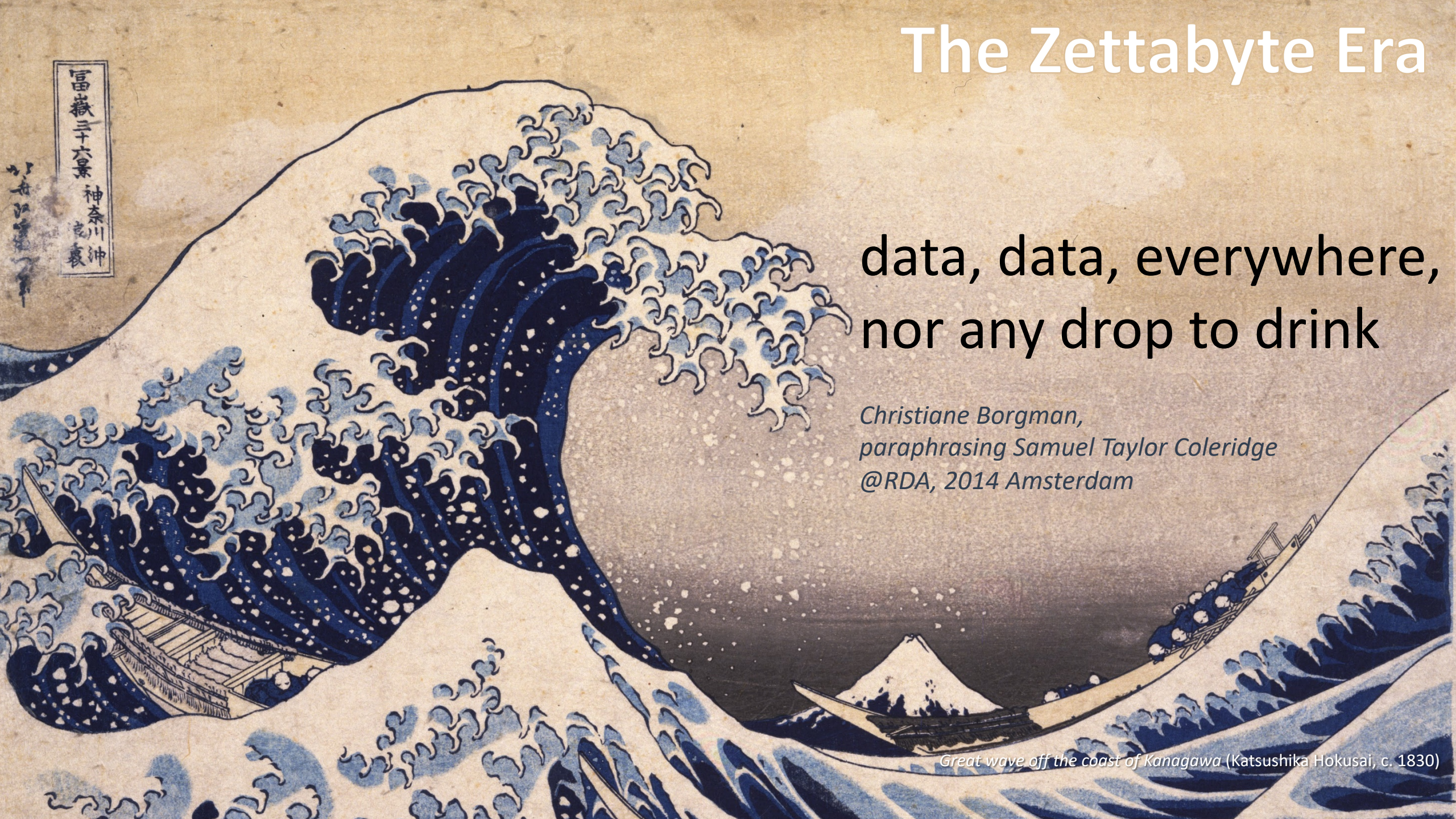
Need to deliver data at the

Scale, form and quality required

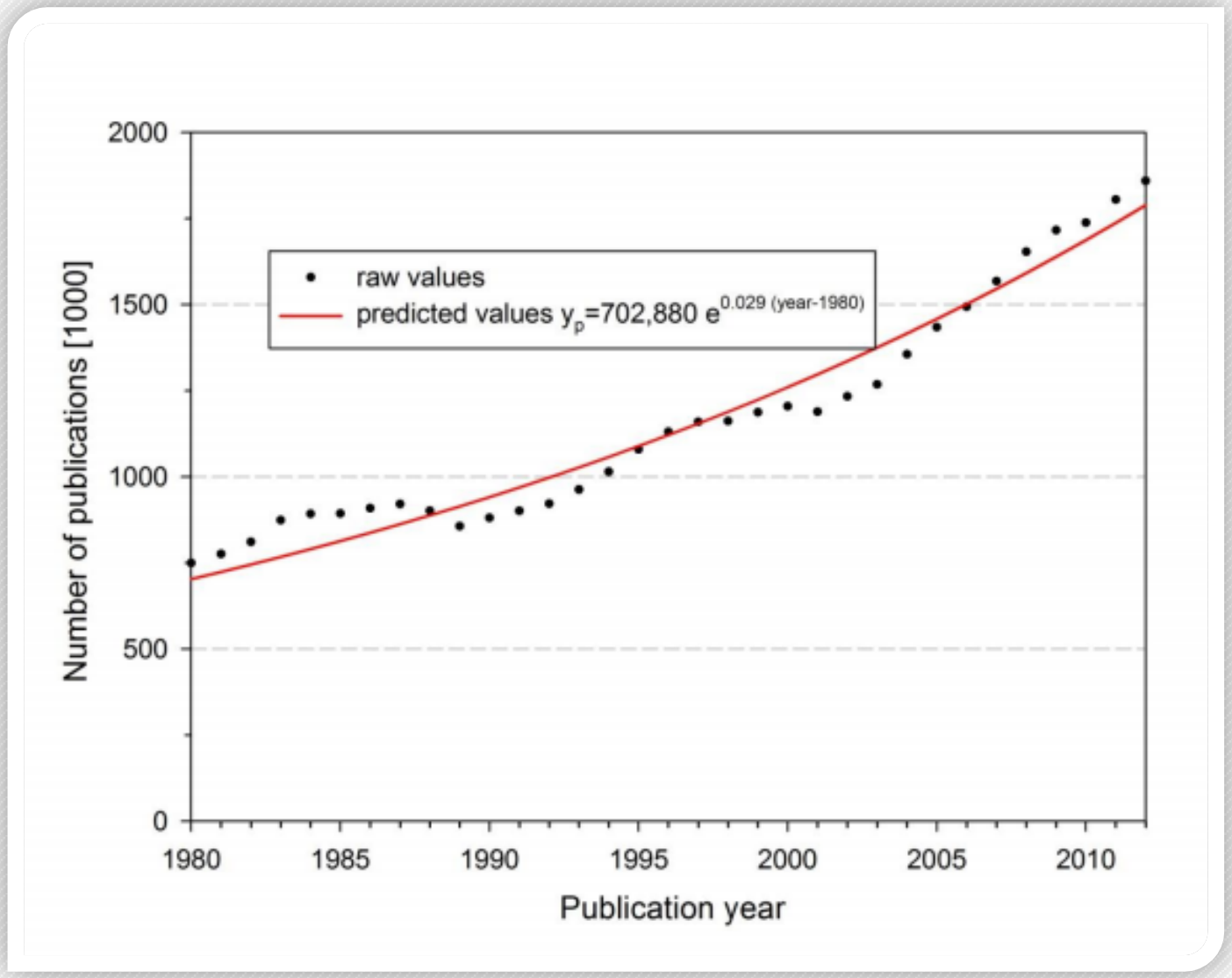
The Zettabyte Era

data, data, everywhere, nor any drop to drink

Christiane Borgman,
paraphrasing Samuel Taylor Coleridge
@RDA, 2014 Amsterdam



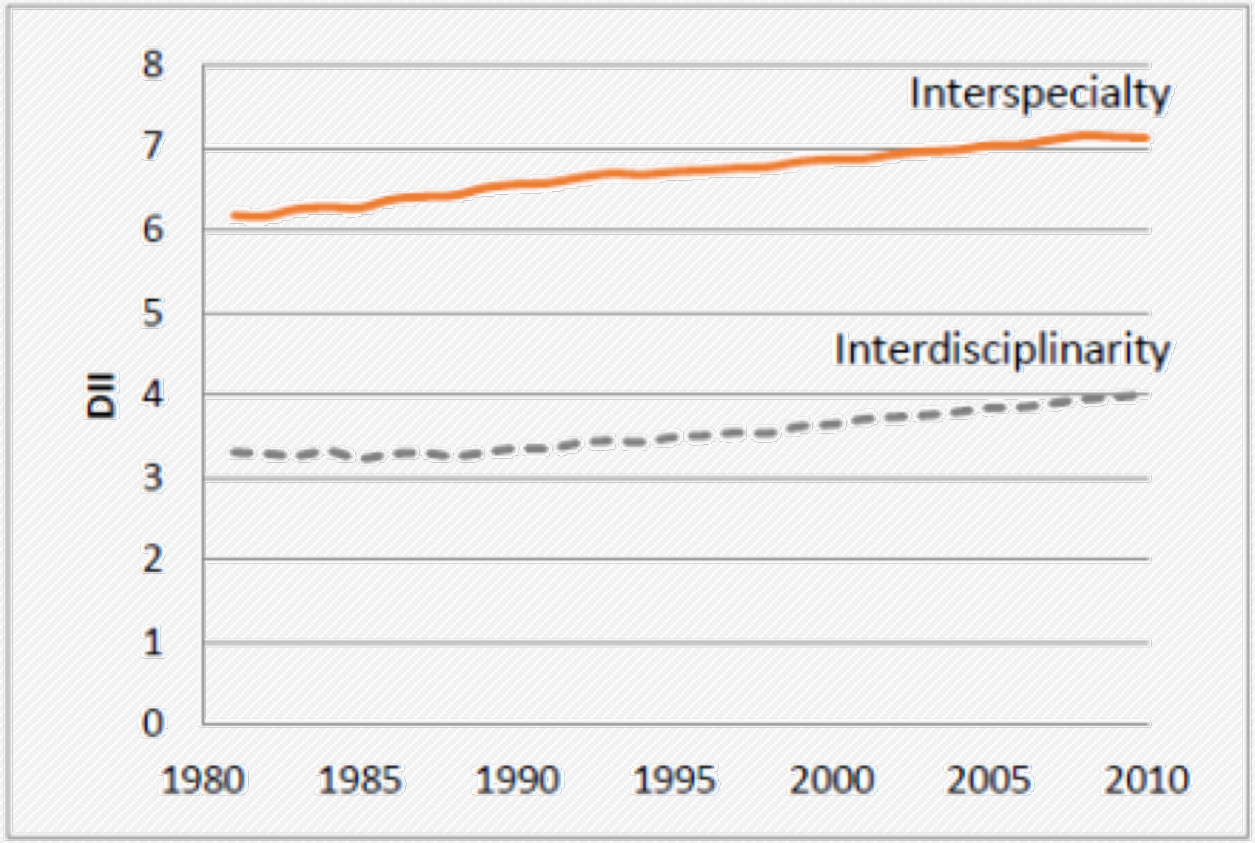
Great wave off the coast of Kanagawa (Katsushika Hokusai, c. 1830)



L. Bornmann & R. Mutz, 2014 **arXiv:1402.4578**

Ever-increasing rate of global scientific products

Does data ‘availability’ affect scientific outputs rate?



Impact Indicator of interdisciplinary research from 1981–2010

Chen, Shiji, et al. "Interdisciplinarity patterns of highly-cited papers: A cross-disciplinary analysis." *Proceedings of the American Society for Information Science and Technology*51.1 (2014): 1-4.

Impact of Interdisciplinary research publications

Challenges **global**

- It needs global standards
- Global workflows
- Cooperation of global players

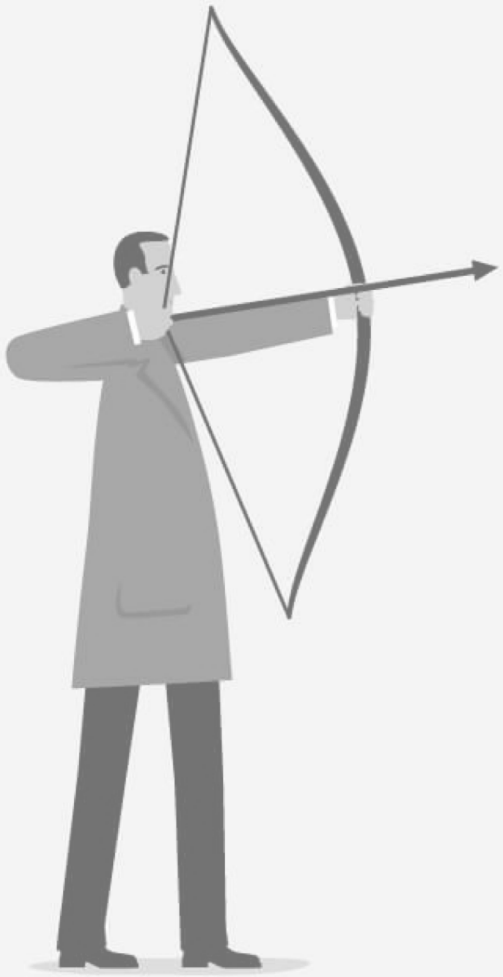
BUT

Science carried out **“locally”**

- By local scientists
- Being part of local infrastructures
- Having local funders

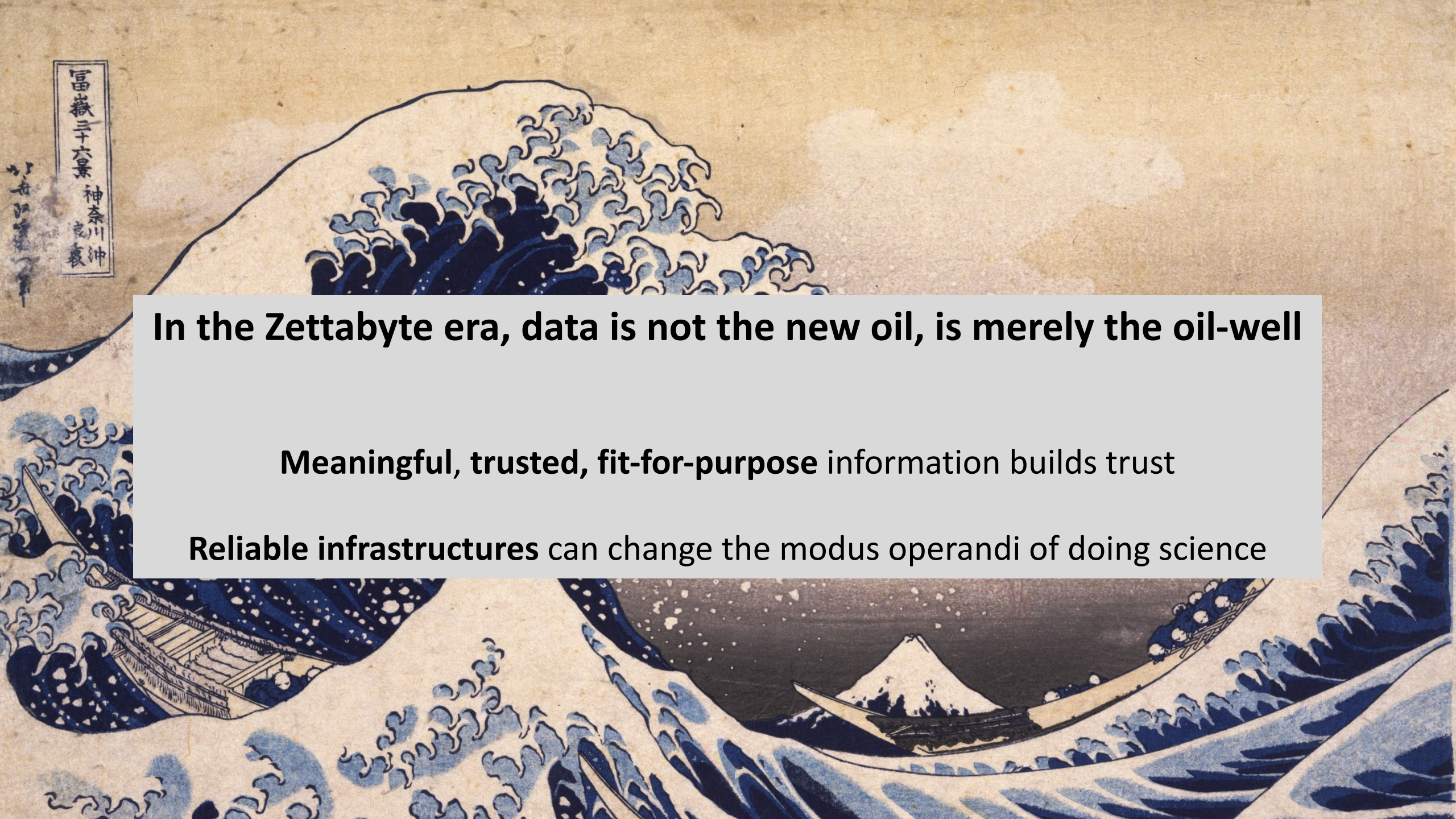
TRUST

Practices
People
Data



Trust lost when datasets disconnect from:

context in which they were created,
or
communities who created them.



In the Zettabyte era, data is not the new oil, is merely the oil-well

Meaningful, **trusted, fit-for-purpose** information builds trust

Reliable infrastructures can change the modus operandi of doing science

Where are we today?

- Claim: data can be made considerably more useful with the right research and use case development

- Application pull exists: the Internet of Things is just around the corner, deep learning is dependent on evermore data, and calls for a more Open Science are growing

Open Science

- The printing press enabled the Royal Society's call in the 17th century science for scientists to publish their results
- Networked computers could today enable the publishing of *all* science products: articles, data, software, workflows, etc
- The US National Academy of Sciences consensus report of July, 2018, titled *Open Science by Design,* included the recommendation that *all research products be made available according to the FAIR principles*

Big, Open, and FAIR Data

- In 2011, the US Federal interagency committee coordinating IT research established a senior steering group for *big data* research. This acknowledged that we could now store more data than we could effectively process

- In 2013, the US President's Science Advisor signed an executive order requiring all research products (articles, data, software, workflows, etc.) produced under Federal support to be *open* for public access

- In January, 2014, a workshop was held at the Leiden University Lorentz Center under the leadership of Professor Barend Mons, to consider what characteristics open data should have to be useful. The result: *FAIR* data

Enabling Hardware

- In 1970, one thousand transistors could be constructed on a chip, in 1990 it was one million, and in 2010 it was one billion
- Fiber optic/laser communication bandwidth has increased even more rapidly from mega-bits per second in the 1980s, to giga-bps in the 1990s, to tera-bps in the 2000s, to (experimental) peta-bps in the 2010s
- Disk prices have dropped from $500,000 per gigabyte in 1981 to $0.03 per gigabyte today (a four terabyte drive for $100)
- These great increases in performance and the equally important decreases in cost have enabled data-intensive science, machine learning, and other use case advances

Machine Learning

- Supervised
- Unsupervised
- In both cases “understanding” the data is a requirement

Interoperable Computing

- Creating interoperable computing elements by introducing *new levels of abstraction* has been a powerful computer science technique
- High level languages and their interpreters solve the interoperability problem for heterogeneous computers
- The Internet solves the interoperability problem for heterogeneous networks
- The Digital Object Architecture could solve the interoperability problem for heterogeneous data

Era's of IT?

- The first era (from 1950 to 1995) was one of many computers and many datasets
- The second era (from 1995 to 2025?) has been one of a single computer and many datasets. Recall SUN Computer's marketing slogan, "the network is the computer"
- The anticipated third era (2025?-) will be one of a single computer and a single dataset. That is, the desired state of the *interoperability of* all *heterogeneous data*

What is the Digital Object Architecture?

- Infrastructure: technology whose primary purpose is to enable or improve the use of other technologies
  - eg, highways for cars, air traffic control and airports for planes, electric grids for power, the Internet for computers

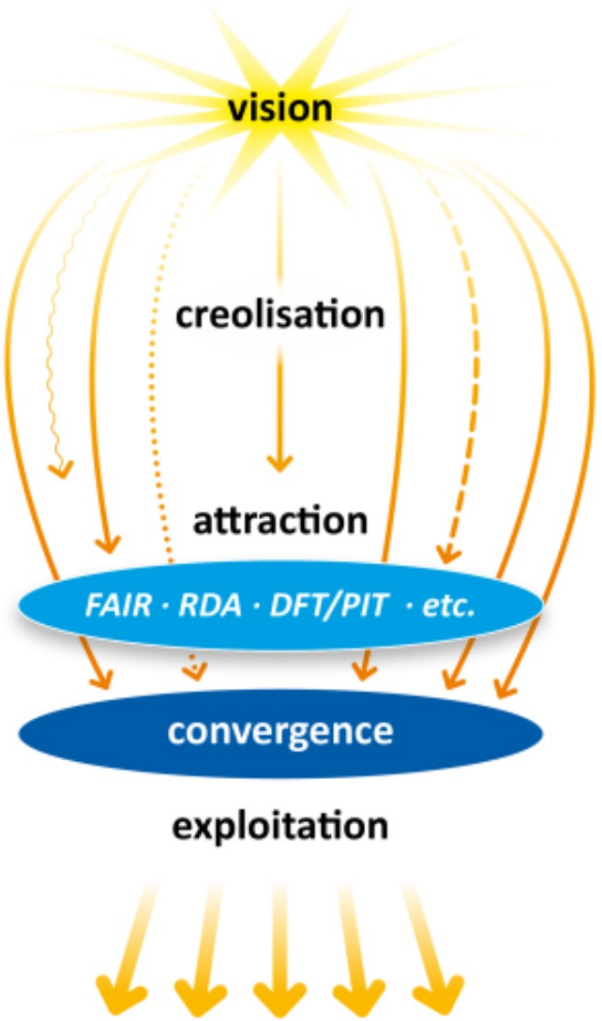
Historical Patterns

of

Infrastructure

Development

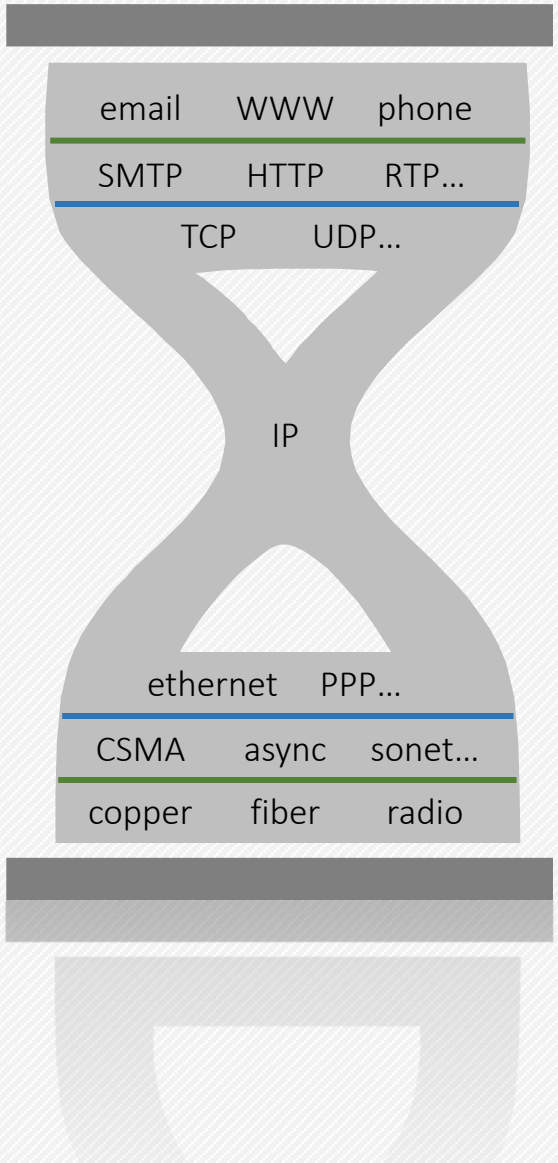
- Vision
- Creolization
- Attraction
- Convergence
- Exploitation



Historical Patterns - Internet

- Vision — One terminal connecting to any computer
- Creolization — Vender-specific nets, ARPAnet, OSI
- Attraction — TCP/IP, packet switching
- Convergence — NSFnet
- Exploitation — Email, remote logon, file transfer (WWW)

Hourglass Model: Internet



Value Added Services

Internet Protocol Suite

Network Technology

Historical Patterns - WWW

- Vision — Easy access to information on the Internet
- Creolization — Gopher, WWW
- Attraction — HTTP, HTML
- Convergence — Mosaic, Netscape, Internet Explorer, etc
- Exploitation — Google, Amazon, Facebook, etc

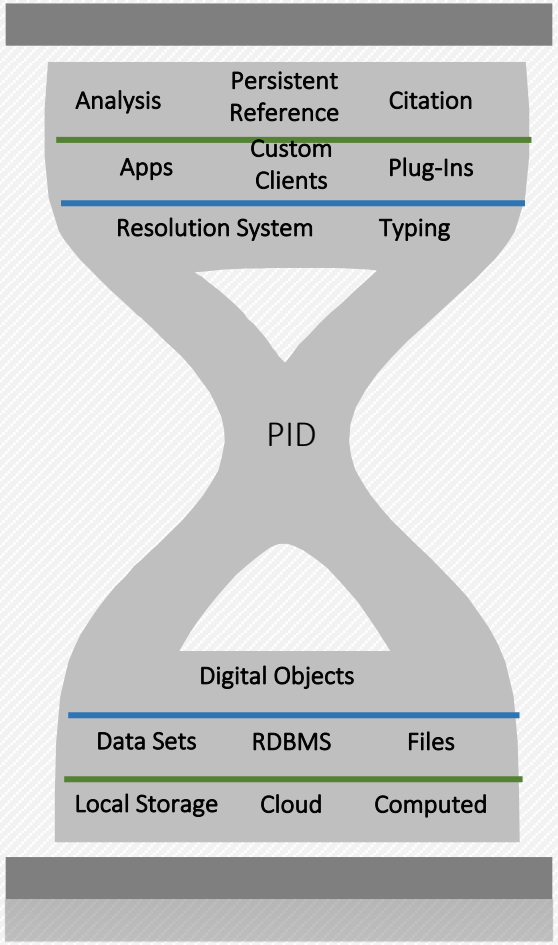
Historical Patterns - Data

- Vision — Interoperable, Heterogeneous, Big Data
- Creolization — No-SQL, Map-Reduce, Cloud computing, etc
- Attraction — Permanent Identifiers (PIDs), FAIR data
- Convergence — ?
- Exploitation — ?

Aspiration

- Vision — Interoperable, Heterogeneous, Big Data
- Creolization — No-SQL, Map-Reduce, Clouds, etc
- Attraction — Permanent Identifiers (PIDs), FAIR data
- Convergence — Global Open FAIR, *Digital Objects* (DOs)
- Exploitation — Open Science, “Recreating Capitalism,” automation of “data wrangling”, etc

Hourglass Model: Information Management on Networks



Value Added Services

Persistent Identifiers

Data Sources

The DO Attracter

- Originally, programmers built their own complex data structures out of the arrays and records provided by programming languages
- Next “abstract data types” emerged, hiding data structure details and accessing the data structure only by defined operations (this led to “object-oriented programming”)
- Now *Digital Objects* have been defined that are self-identifying, so that programs can query an object to find out what it is and which operations can be applied.

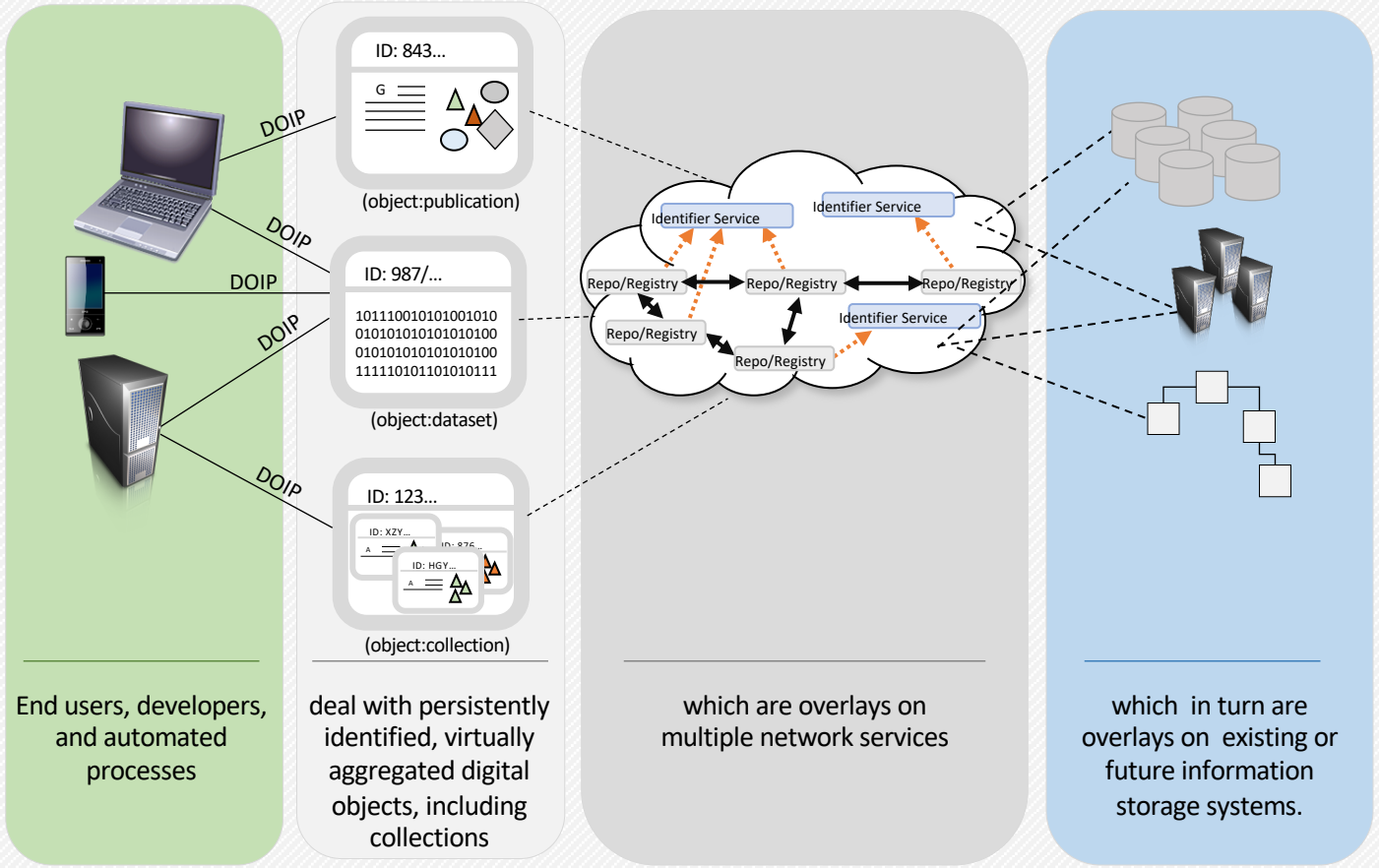
DO Architecture

- PIDs point to digital objects (DOs), just as URLs point to web pages
- A *DO Interface Protocol* (DOIP) plays the same role for DOs as HTTP plays for Web pages (transferring DOs and facilitating other operations)
- DOs can contain other DOs; in particular, the “outer most” DO in any assemblage is a special DO called a *repository*
- DOs contain not only data, software, etc, but also (and especially) *metadata*

The FAIR Attracter

- Findable. Analogous to google searching, but generalized to include indexing of machine-readable *metadata* (and not just keywords)
- Accessible. "Open access" to persistent machine-readable *metadata* whether or not the data is behind a pay wall
- Interoperable. Intelligent use of machine-readable *metadata* to enable heterogeneous data to be utilized by different software without manual "data wrangling"
- Reusable. Liberal application of automated copy-write law enabling direct use *and* derived works wherever possible

High-Level View of Digital Object Infrastructure



Motivation

- If the data infrastructure coalesces as the Internet and Web did, the data infrastructure will be revolutionary
- The Internet of Things will require a better data infrastructure to reach its potential value
- Even where data remains in silos, a global FAIR data infrastructure could greatly automate data wrangling (which currently takes an estimated 80% of data scientists' time)
- It could also facilitate the emergence of a "data market"
- It would facilitate the emergence of *Open Science*, which advocates believe will be as revolutionary as the 17th century science revolution

How do we realize the DOA?

The E-RPID Vision

- Starts with data network based on Digital Object Architecture (DOA), a distributed architecture of services spread worldwide that together identify and resolve digital objects
- DOA first espoused by Internet founder Robert Khan in the mid’80’s.
- DOA is a network of Handle servers at its core

Required Components

- Identifier – Globally persistent unique identifier
- Identifier Resolution – State information related to object (including location but not the actual data object)
- Data Typing – Definition of state information for programmatic interfaces
- Mapping/Brokering Services – Prevent refactoring of existing repositories
- Repositories – Provide access to digital objects
- Operational Protocol – Allow interaction with objects

The RPID Testbed

- Identifier – Globally persistent unique identifier
- Identifier Resolution – State information related to object
- Data Typing – Definition of state information for programmatic interfaces
- Mapping/Brokering Services – Prevent refactoring of existing repositories
- Repositories – Provide access to digital objects (Use Cases)
- Operational Protocol – Allow interaction with objects

NSF Grant No. 1659310

The Digital Object Architecture serves as base infrastructure only. DOA is silent on issues of modeling data objects themselves: their *content*, their *relationship to their own metadata*, and *relationship between data objects*

For object modeling we turn to FAIR principles and PID Kernel Information

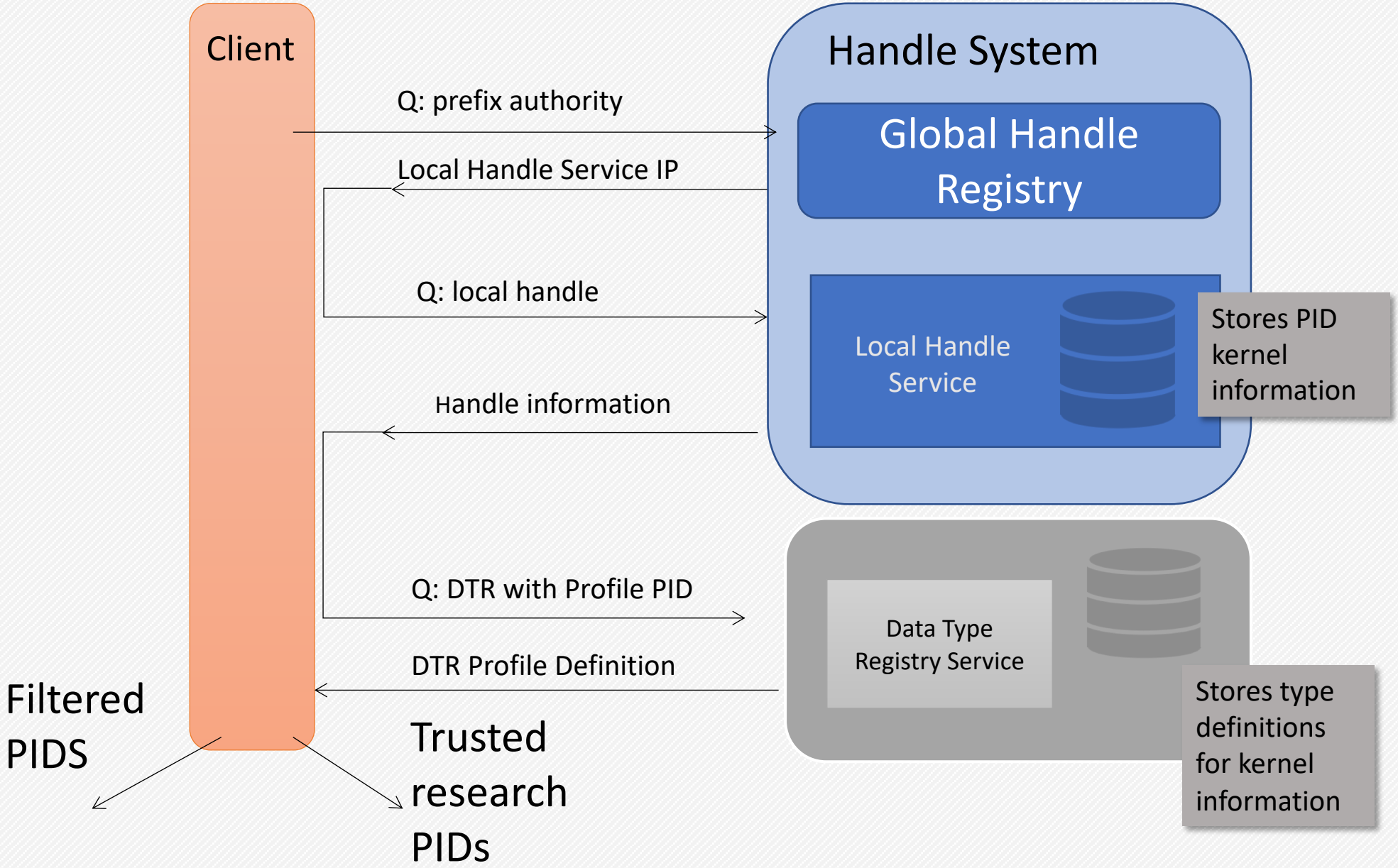
Imagine an Internet-scale data client that is handed a list of a 100,000,000 PIDs.

How does client quickly sift through list to find research data objects?

Further suppose client is able to winnow list down to just research data objects, how does it then quickly discard fakes?

plale@indiana.edu

PID Kernel Information Use case: Client filters list of millions of PIDs to identify research data and makes simple determination of trust



RPID Outcomes

- A client working with RPID services and PID Kernel Information looks at each PID is a list, accepts those it ”likes”.
  - KI profile is stored in DTR
  - PID KI holds time amount of data provenance from which basic trust can be derived
- Exploration driven by identifying and evaluating minimal information that can go into Kernel Information that can help make Data Objects FAIR and less dependent on the repository system to enforce FAIRness
- Kernel information has potential to spawn new ecosystem of data services for smart data objects

ERPID Testbed

- Identifier – Globally persistent unique identifier
- Identifier Resolution – State information related to object
- Data Typing – Definition of state information for programmatic interfaces
- Mapping/Brokering Services – Prevent refactoring of existing repositories
- Repositories – Provide access to digital objects (Use Cases)
- Operational Protocol – Allow interaction with objects

E-RPID Work Items

- Evaluate and Create Mapping/Brokering Services
- Provide a testbed and feedback for DOIP
- Provide training and education materials
- Move services to an Open Stack environment (JetStream)
- Continue the User Advisory Group
- New use cases – SGRC and Galaxy
- Recruit more use cases

Conclusions

- IT is poised to do for data what the Internet did for networks
- ERPID and other DOA services could be the start
- Interoperable data will become a core infrastructure, like the Internet has already become
- We will wonder how we ever got along without it!

Science is a ‘light’s better’ endeavour in that research effort is **not directed at areas where the work is technically infeasible**. Research is directed where real, interpretable results may be obtained.

We do, in fact, conduct research where the light’s better.

But, when the light changes, so does science.

With better illumination, we look in new areas.

We find new things…