Globally Distributed Secure Data Exchange Fabrics

Cees de Laat

Systems and Networking Laboratory
University of Amsterdam

Contributions from:

Leon Gommans, Paola Grosso, Wouter Los, Yuri Demchenko, Lydia Meijer, Tom van Engers, Sander Klous, Rodney Wilson, Marc Lyonais, Inder Monga, Reggie Cushing, Ameneh Deljoo, Sara Shakeri, Lu Zhang, Joseph Hill, Lukasz Makowski, Ralph Koning, Gleb Polevoy, Tim van Zalingen, and many others!

All for one and one for all

- All for one
  - Many infrastructures centered around compute and workflows
- One for all
  - Now we need to get a fluid data layer that frees data to be shared and used by (unforeseen) applications

- Efforts as FAIR and ScienceDMZ / DTN fabrics pave the way to solve the data problem that is also encountered by industry.

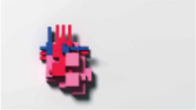
Harvard Business Review

ECONOMY

Managing Our Hub Economy

by Marco Iansiti and Karim R. Lakhani

FROM THE SEPTEMBER–OCTOBER 2017 ISSUE



THOMAS M. SCHEER/EYEEM/GETTY IMAGES

I. The Problem

The global economy is coalescing around a few digital superpowers. We see unmistakable evidence that a winner-take-all world is emerging in which a small number of “hub firms”—including Alibaba, Alphabet/Google, Amazon, Apple, Baidu, Facebook, Microsoft, and Tencent—occupy central positions. While creating real value for users, these companies are also capturing a disproportionate and expanding share of the value, and that’s shaping our collective economic future. The very same technologies that promised to democratize business are now threatening to make it more monopolistic.

HARVARD BUSINESS PUBLISHING

Data value creation monopolies

➔

Create an equal playing field

➔

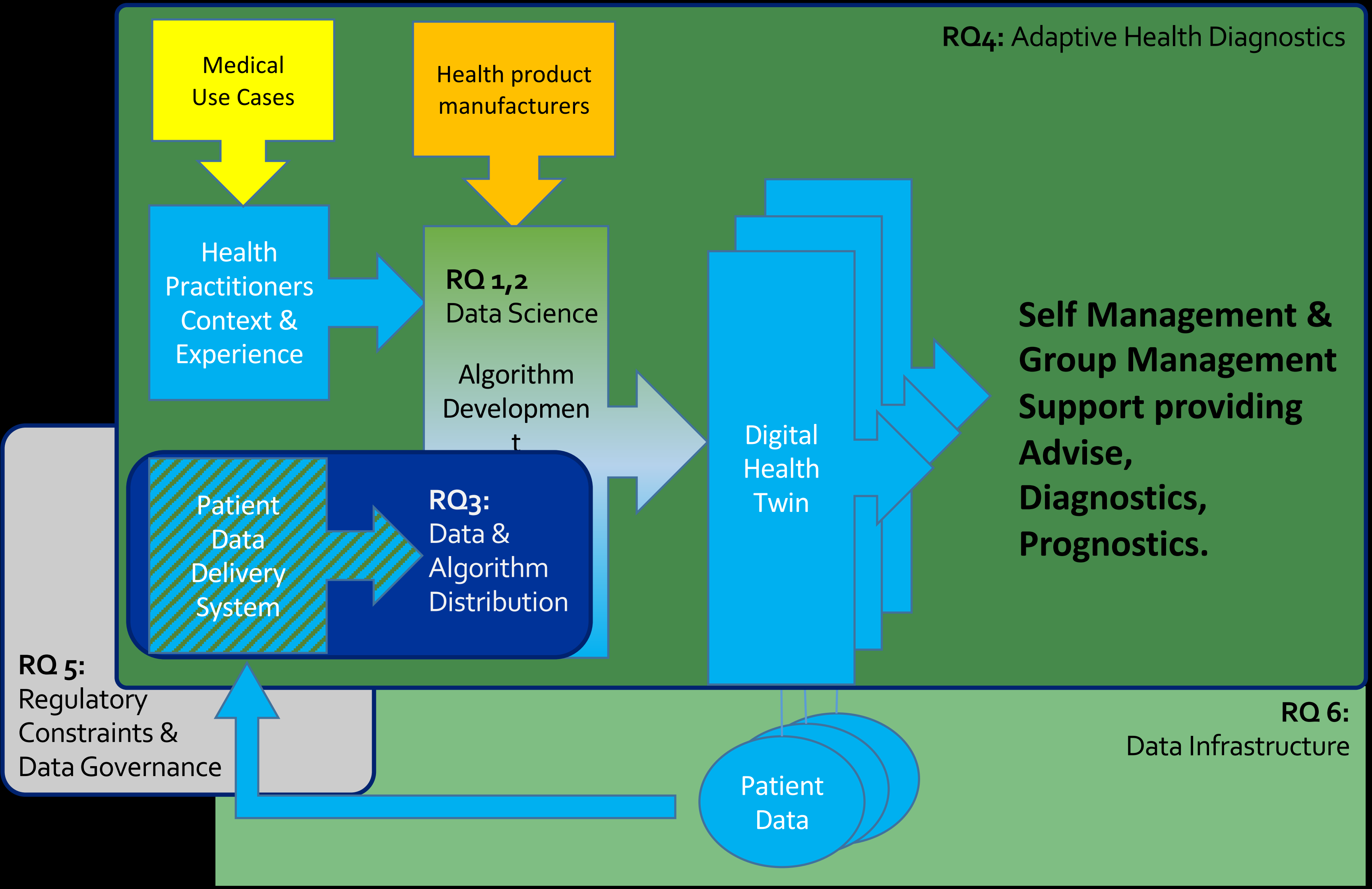
Sound Market principles

https://hbr.org/2017/09/managing-our-hub-economy

Main problem statement

- There is lots of data out there that is not shared (99%)
- FAIR is typically not fair ;-), but limited by policy and/or law
  - the A in FAIR is about access, trust is hard to implement across domains
- Organizations that normally compete have to bring data together to achieve a common goal/benefit!
- The shared data may be used for that goal but not for any other!
- Expected use is fine but unexpected use/mission creep…
- Data processed by alien algorithms in foreign data centers… Hmmm…
  - How to organize data processing alliances?
  - How to enforce policy using modern Cyber Infrastructure?
  - How to translate law policy from strategic via tactical to operational level?
  - What are the different fundamental data infrastructure models to consider?

Health Use Case
Enabling Personal Interventions



The overall aim of this project is to explore the use and effectiveness of data driven development of scientific algorithms, supporting personalized self- and joint management during medical interventions / treatments.

The key objective is to use data science promoting health practically with data from various sources to formulate lifestyle advice, prevention, diagnostics, and treatment tailored to the individual and to provide personalized effective real-time feedback via a concept referred in this proposal as a digital health twin.

https://delaat.net/epi

Big Data Sharing use cases placed in airline context

Global Scale

National Scale

City / regional Scale

Campus / Enterprise Scale

Aircraft Component Health Monitoring (Big) Data
NWO **CIMPLO project**
4.5 FTE

Cargo Logistics Data
(C1) DL4LD
(C2) Secure scalable policy-enforced distributed data Processing
(using blockchain)

NLIP iShare project

iSHARE
powered by NLIP

Cybersecurity Big Data
NWO COMMIT/
SARNET project
3.5 FTE

https://delaat.net/dl4ld

System and Network Engineering

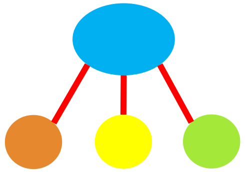
AIR FRANCE KLM

RESEARCH WORKING ALONGSIDE IT INDUSTRY

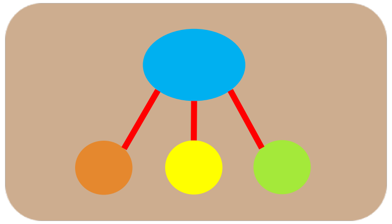
NETWORK RESEARCH INFRASTRUCTURES

COMMERCIAL DATACENTER INFRASTRUCTURE AS NEUTRAL GROUND

Data Sharing Infrastructure Model Research using Future Internet capabilites



Research collaboration

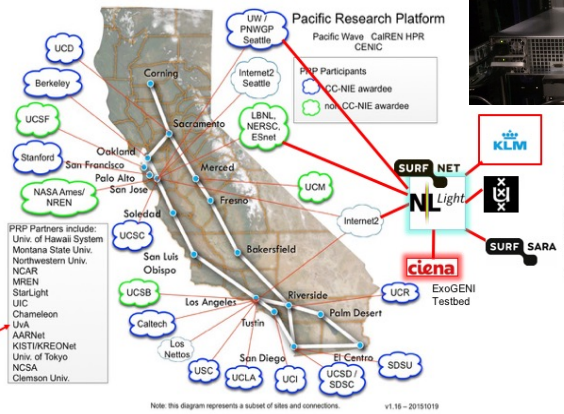


Goal: How to create a Digital Marketplace Ecosystem

PACIFIC RESEARCH PLATFORM

prp.ucsd.edu

As foundation of the National Research Platform

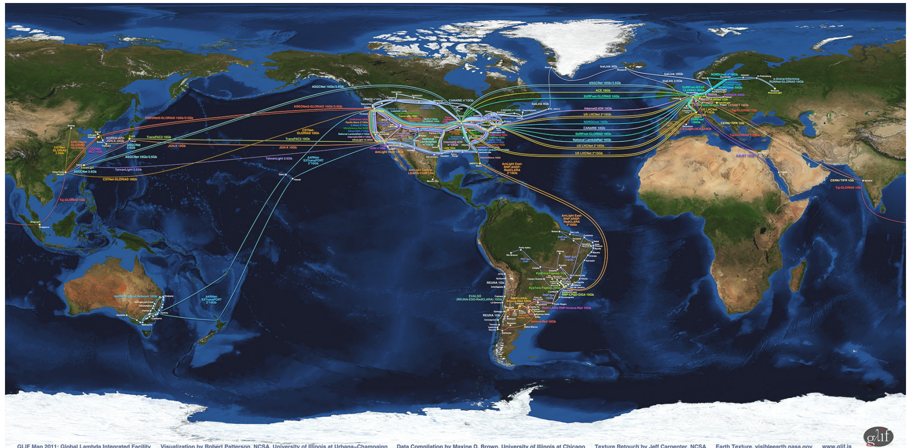


Data Transfer Node at KLM fieldlab with 100 gb/s link to enable SDMP research thanks to UvA, SURFnet and Ciena

AM3 and AM4 Datacenters Science Park Amsterdam

SV10 Datacenter Silicon Valley

Gobal Lambda Integrated Facility ➔



NCE KLM

SAE Use Case envisaged **research** collaboration

Funding Agency	NSF — Big Data Hub / Spoke or Industry initiative funding			NWO Netherlands Organisation for Scientific Research — Topsector Funding
International Networking	Internet2, ESnet Energy Sciences Network			GÉANT
Regional / National Networking	CENIC Connecting California	SoX Southern Crossroads	LEARN Lonestar Education and Research Network	SURF NET
Local University	Stanford	Georgia Tech	UT Dallas	Universiteit van Amsterdam
Aircraft MRO, OEM & Operators	Boeing	Delta Air Lines	Bell Helicopter	Air France KLM
Industry Standards Body	SAE International — SAE AeroSpace Group HM-1 working group Use Case on aircraft sensor Big Data			

Approach

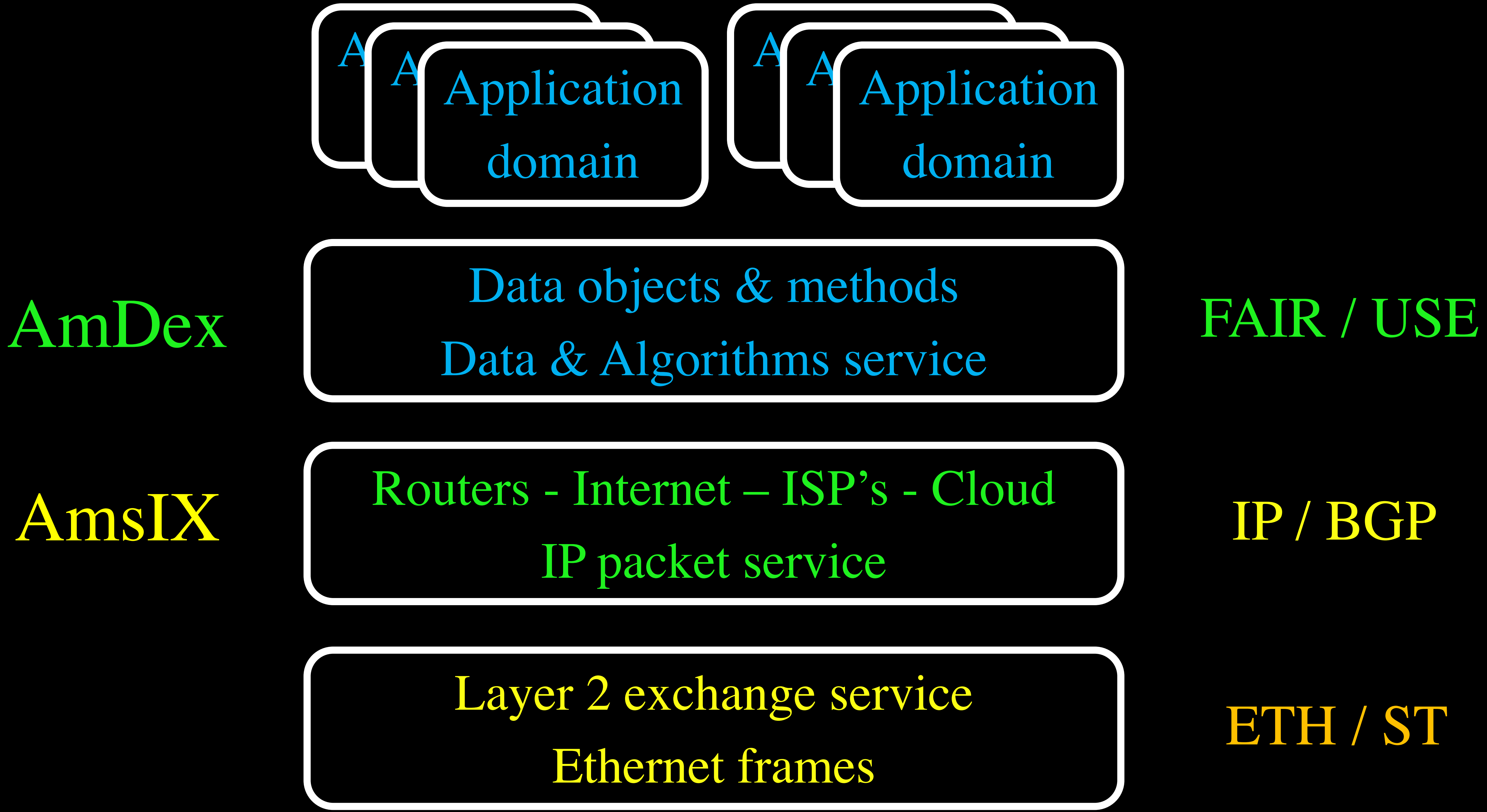
- Strategic:
  - Translate legislation into machine readable policy
  - Define data use policy
  - Trust evaluation models & metrics
- Tactical:
  - Map app given rules & policy & data and resources
  - Bring computing and data to (un)trusted third party
  - Resilience
- Operational:
  - TPM & Encryption schemes to protect & sign
  - Policy evaluation & docker implementations
  - Use VM and SDI/SDN technology to enforce
  - Block chain to record what happened (after the fact!)

IETF: Common Open Policy Service (COPS)

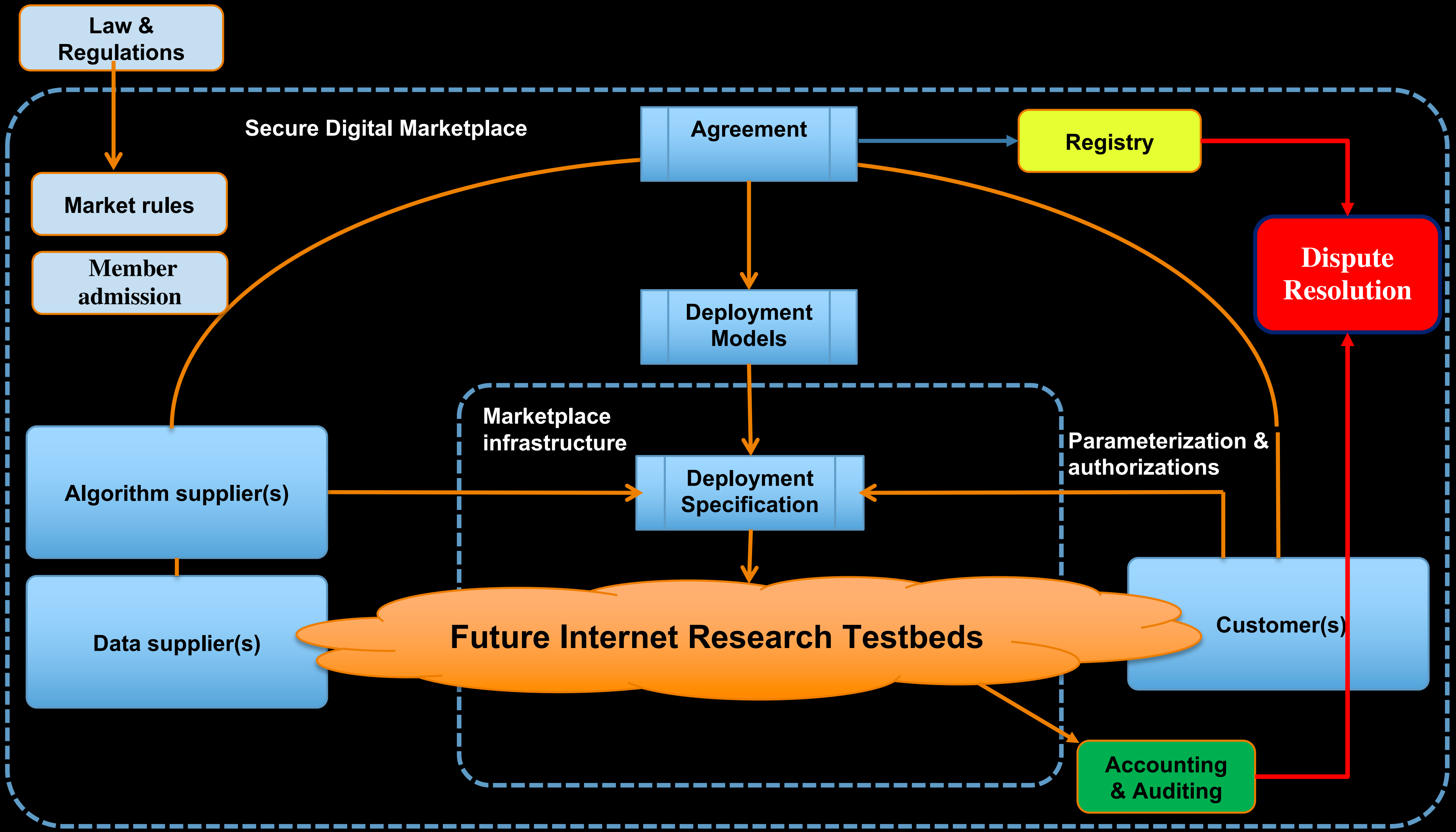
- Rfc 2748, 2753, 4261

```
 ________________                          ____________________
|                 |                        |                    |
|  Network Node   |  Policy Server         |    Network Node    |
|    _____        |      _____             |  _____      _____  |
|   |     |       |     |     |            | |     |    |     | |
|   | PEP |<------|---->| PDP |            | | PEP |<-->| PDP | |
|   |_____|       |     |_____|            | |_____|    |_____| |
|     ^           |                        |                    |
|     |    _____  |                        |____________________|
|     \-->|     | |
|         | LPDP| |
|         |_____| |
|                 |
|_________________|
```

☺


Application domain
Application domain
AmDex
Data objects & methods
Data & Algorithms service
FAIR / USE
AmsIX
Routers - Internet – ISP’s - Cloud
IP packet service
IP / BGP
Layer 2 exchange service
Ethernet frames
ETH / ST

Secure Digital Market Place Research



What have we been doing?

- Studying and defining draft Policy
- Working out model & defining Archetypes
- Implementing a proof of concept using several distributed DTN's and dockers on kubernetes.
- Demo at SC18 in Dallas TX, on data harbours.
- Tactical operation of Digital Data Markets
- Modeling and Matching Digital Data Marketplace Policies
- Optimization of degrees of freedom == value

SC16 Demo

DockerMon

Sending docker containers with search algorithms to databases all over the world.

http://sc.delaat.net/sc16/index.html#5

Container-based remote data processing

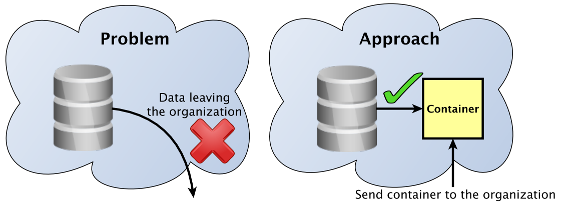
Universiteit van Amsterdam

Łukasz Makowski, Daniel Romão, Cees de Laat, Paola Grosso
System and Networking Research Group, University of Amsterdam

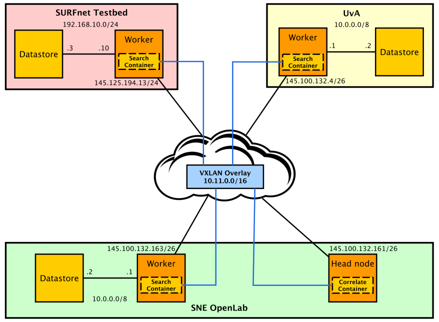
System and Network Engineering

Problem Description

- Scientific datasets are usually made publicly available
  ….but data cannot always leave the organization premises
- On-site data processing can be challenging because of incompatibility of systems or lack of manpower
- Can a container-based system perform remote on-site data processing efficiently?
- What are the networking issues to solve?



Underlay and Overlay



Main features:

- Networked containers
- VXLAN overlay
- Containers that perform data retrieval and computation
- Containers built on-demand
- On-site data processing
- Distributed data source
- Multiple sites with datasets

The Game

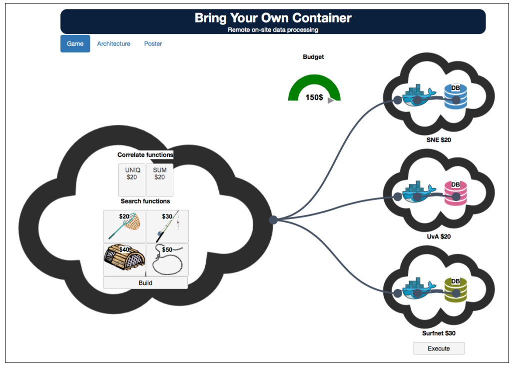
Our SC16 demo is a gamification of the remote dataset processing architecture.

How many different animal species can you find? You have a fixed budget and each function and processing will cost you money!

In our game you will:

- Select a correlate function to combine the results of the different sites.
- Pick different search functions, represented as tools, to find animals in the remote datasets.
- Build containers with the search and correlate functions.
- Execute the containers on the sites of your choice.

Will you have the best score?



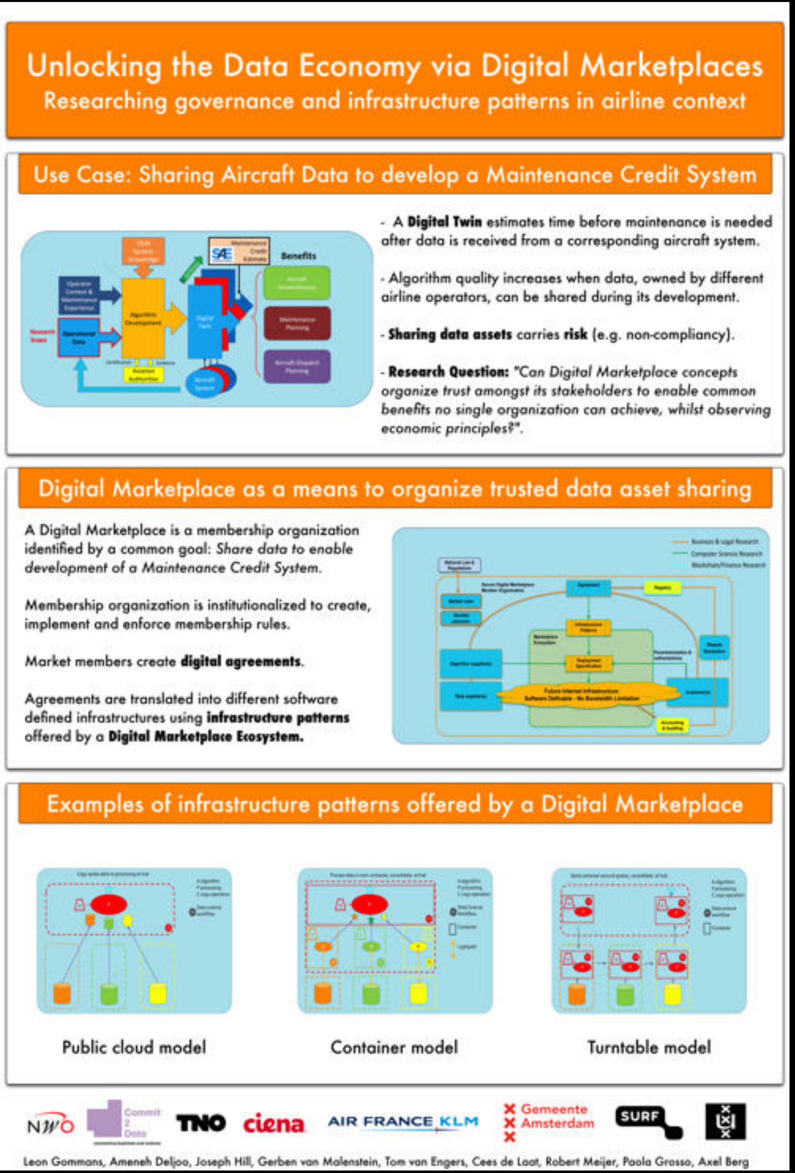
More information:

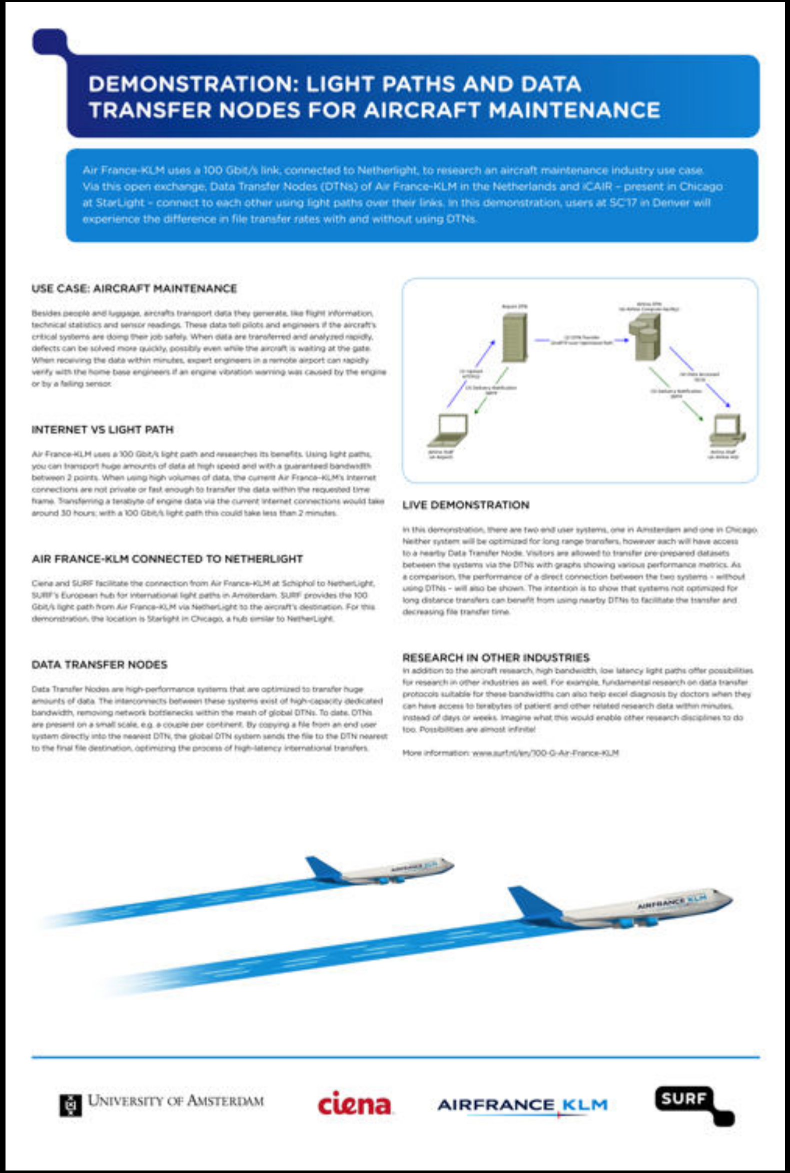
- http://byoc.lab.uvalight.net/info
- http://sne.science.uva.nl/sne/gigaport3
- http://delaat.net/sc

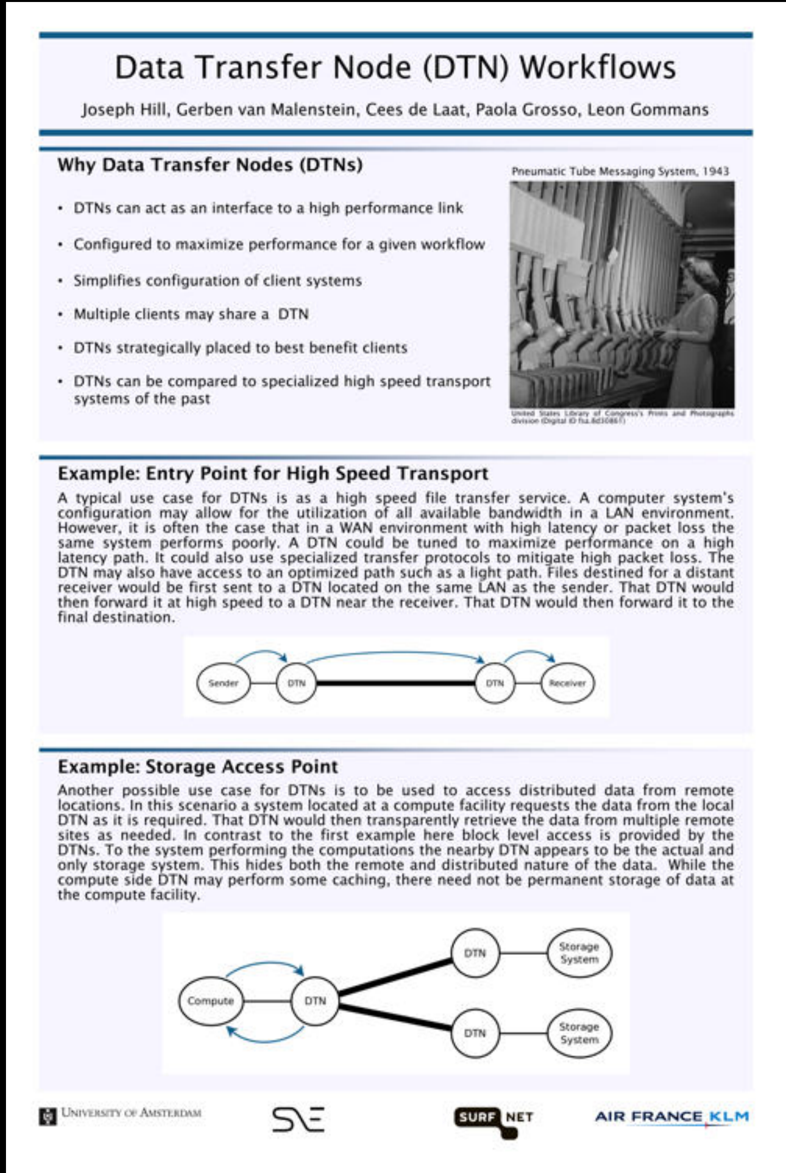
SURF NET

SC17 Posters and proof of concepts & demo's

http://sc.delaat.net/sc17





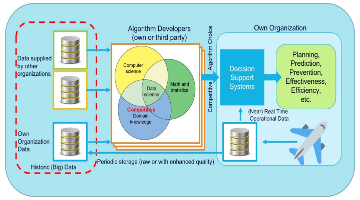


SC18 – Dallas TX

Training AI/ML models using Digital Data Marketplaces

Creating value and competition by enabling access to additional big data owned by multiple organizations in a trusted, fair and economic way

The more data - the better: an aircraft maintenance use-case



- AI/ML algorithm based Decision Support Systems create business value by supporting real-time complex decision taking such as **predicting the need for aircraft maintenance**.
- Algorithm quality increases with the availability of aircraft data.
- Multiple airlines operate the same type of aircraft.
- **Research Question:** *"How can AI/ML algorithm developers be enabled to access additional data from multiple airlines?"*
- **Approach:** *Applying Digital Data Marketplace concepts to facilitate trusted big data sharing for a particular purpose.*

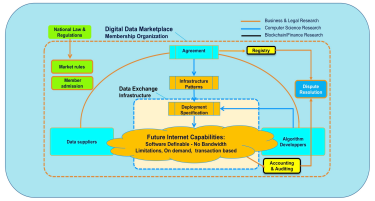
Digital Data Marketplace enabling data sharing and competition

A **Digital Data Marketplace** is a membership organization supporting a common goal: e.g. *enable data sharing to increase value and competitiveness of AI/ML algorithms.*

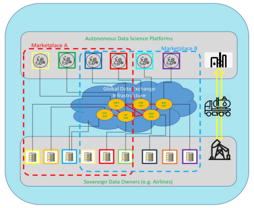
Membership organization is institutionalized to create, implement and enforce membership rules organizing **trust**.

Market members arrange **digital agreements** to exchange data for a **particular purpose** under specific conditions.

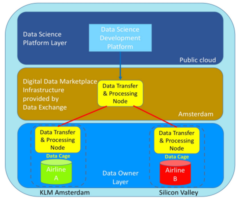
Agreements subsequently drive data science transactions creating processing infrastructures using infrastructure patterns offered by a Data Exchange as **Exchange Patterns**.



Researching Exchange Patterns to support Digital Data Marketplaces



Data Exchange Model



Research Infrastructure

Trust Modelling: What is the optimal infrastructure archetype, describing storage and processing locations and their relationships, which best suit member requirements when considering risk?

Processing Models: What are the implications of distributing data processing across membership organization owned infrastructures in terms of achievable model accuracy and processing performance using federated/distributed models vs centralized models?

Marketplace Reference Architecture: What constitutes a marketplace? Researching needed functions, personas, flows, credentials, contracts & rules, conflict resolution, and much more ...

Research Elements

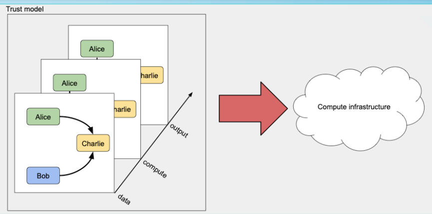
Leon Gommans, Anne Savelkoul, Wouter Kalfsbeek, Dirk van den Herik, David Langerveld, Erik IJzermans, Floris Freeman, Brend Dikkers, Cees de Laat, Tom van Engers, Wouter Los, Paola Grosso, Joseph Hill, Reggie Cushing, Giovanni Sileno, Lu Zhang, Ameneh Deljoo, Thomas Baeck, Willem Koeman, Laurie Strom, Axel Berg, Gerben van Malenstein, Kaladhar Voruganti, Rodney Wilson, Patricia Florissi

Dataharbours: computing archetypes for digital marketplaces

Reginald Cushing, Lu Zhang, Paola Grosso, Tim van Zalingen, Joseph Hill, Leon Gommans, Cees de Laat, Vijaay Doraiswamy, Purvish Purohit, Kaladhar Voruganti, Craig Waldrop, Rodney Wilson, Marc Lyonnais

The problem

How can competing parties share compute and data? The architecture of a digital marketplace is an active research field and has many components to it. Here we investigate a federated computing platform which is molded into different **archetypes** based on **trust** relationships between organizations.



The components

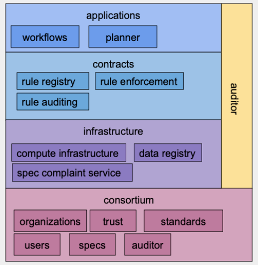
Consortium: is an initial document which brings together organizations that wish to collaborate. It defines static information such as keys to identify parties.

Infrastructure: A single domain organization infrastructure that securely hosts data, compute containers and, optionally, compute infrastructure. We dub this infrastructure a **data harbour**. A harbour implements a set of protocols that allows it to interact with other harbours.

Contracts: Are a set of rules that are shared amongst participating harbours which describe how objects (data, compute) can be traded between harbours and who can process data. In its simplest form is a 7-tuple which binds a user, data object, compute container, contract, consortium, harbour, and expiry date.

An application: Is a distributed pipeline which can make use of several contracts. The combination of application and contract defines the archetype of the computation i.e. how data and compute are moved to effect computation.

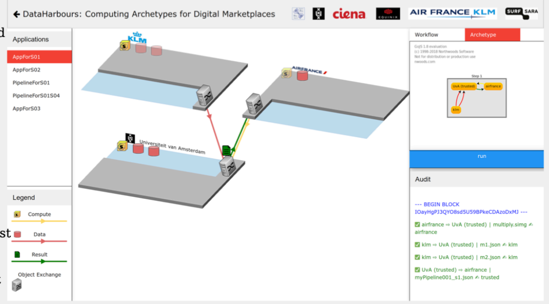
Auditor: A trusted entity that collects audit trails for use in litigation of policy violations.



In action

Federated computing on 3 distributed data harbours. Here we illustrate one archetype where KLM and Airfrance do not trust each other and employ a trusted 3rd party to send the data and compute for processing.

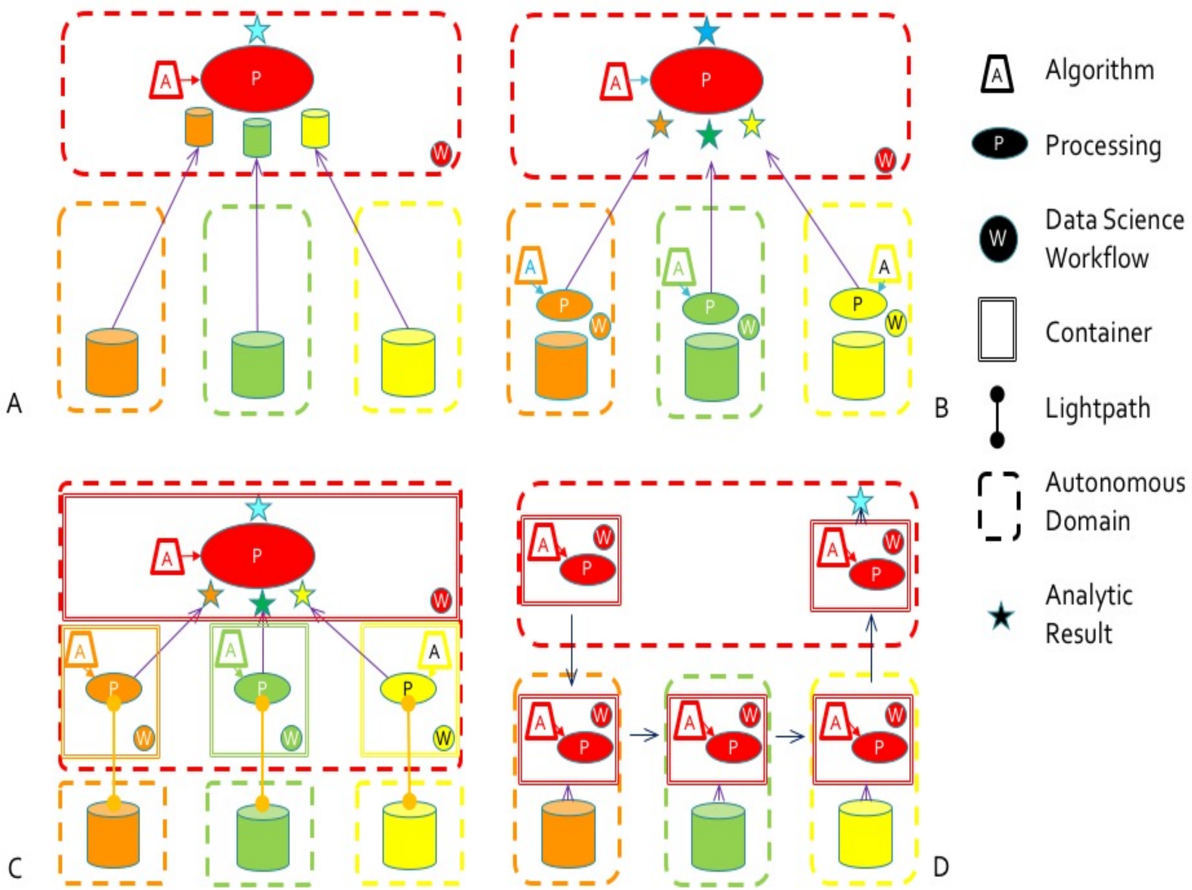
For the scenario to succeed the different harbours need to effect several transactions which are governed by contractual rules.

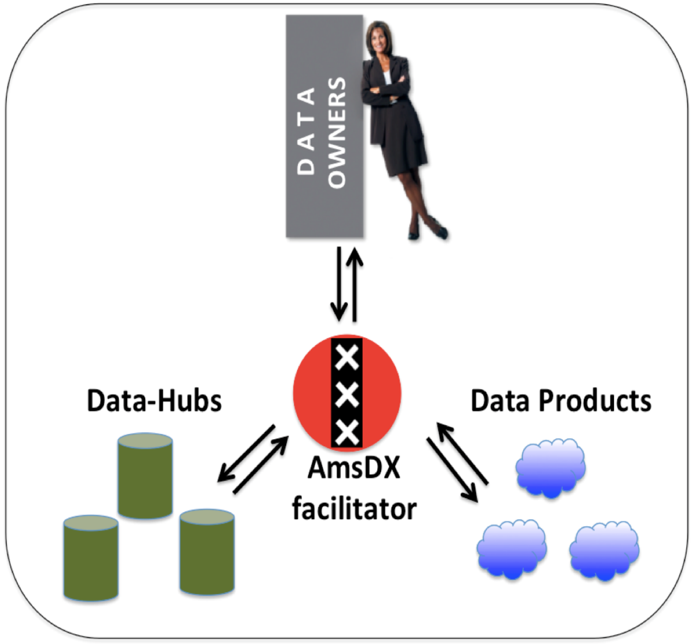


The transaction **protocol** involves first identifying both parties are who they say they are through pub/priv key challenges and secondly, that at least a **contract** rule is matched to allow the transaction. Important steps of the transactions are **audit** logged i.e. signed and published to and audit log collector.

INFRASTRUCTURE PATTERN EXAMPLES

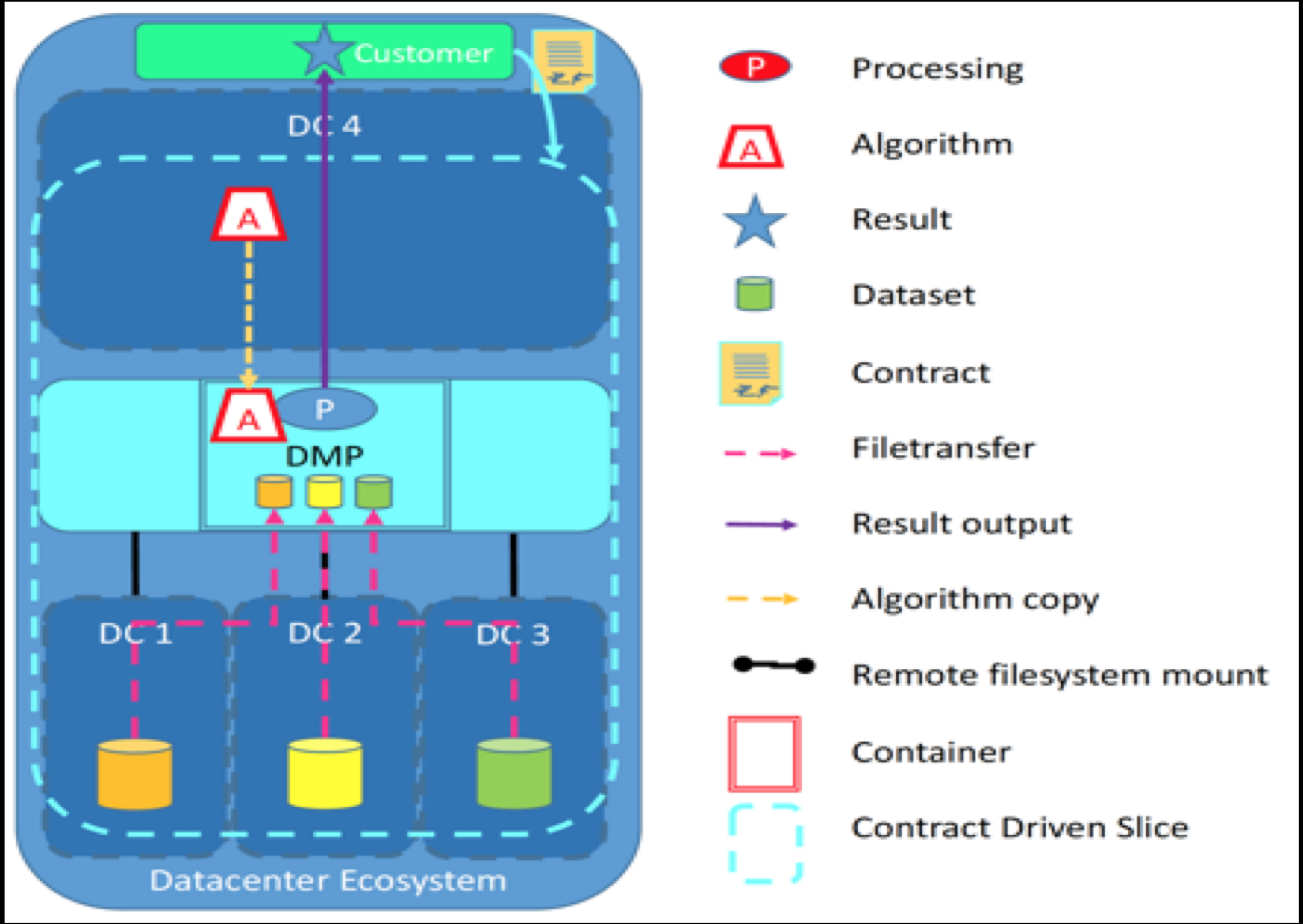
OFFERED BY A DATA EXCHANGE TO MARKETPLACES TO CHOOSE FROM





Open Data Market

DMP archetypes and their representation

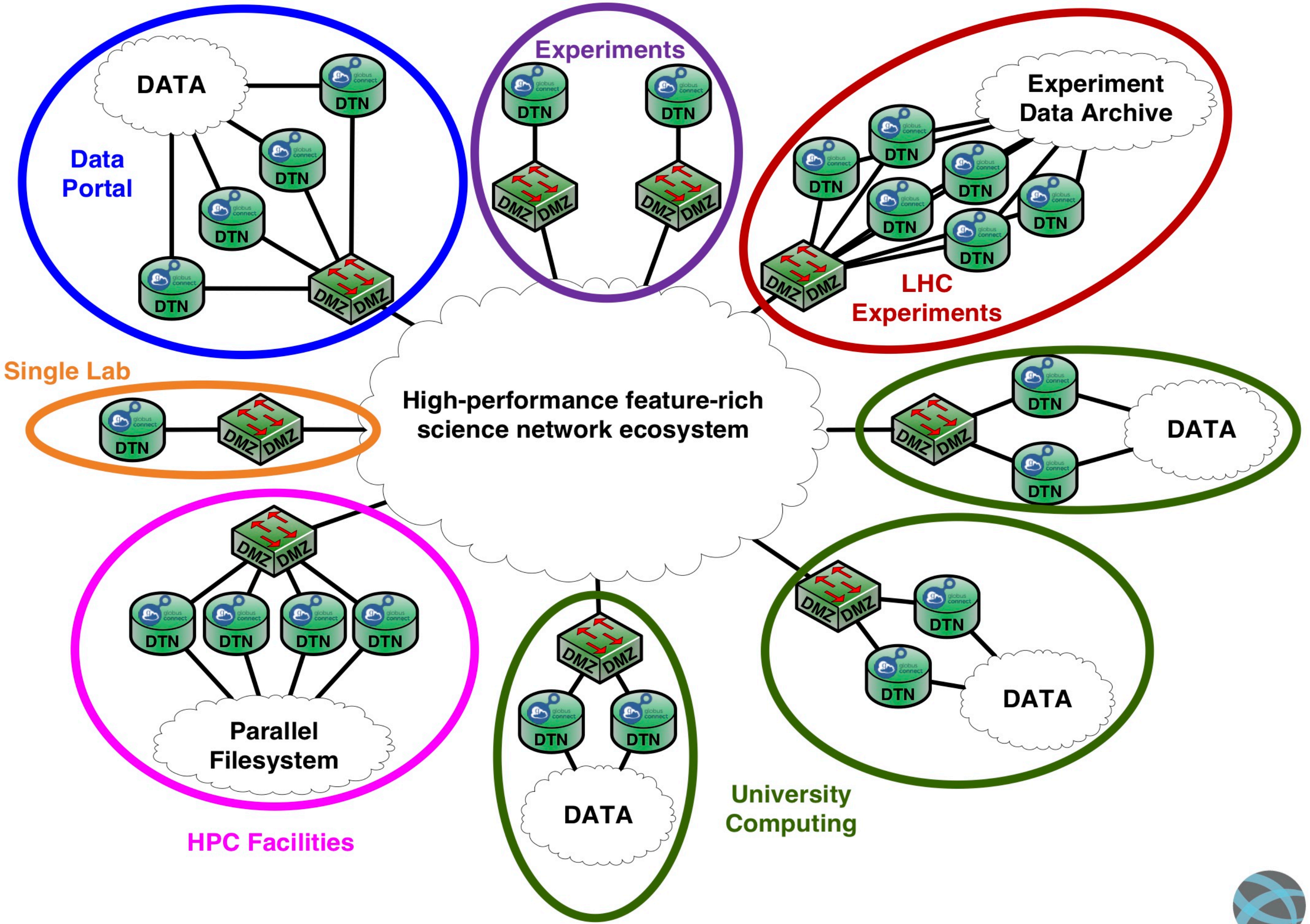


On the left one of the many collaboration models within a DMP. We call this archetype. One DMP can support multiple archetypes depending on the contracts between partners.

To match application/user requests to the archetype we need to model the archetype on the left in generic ways.

To be presented next week at eScience conference.

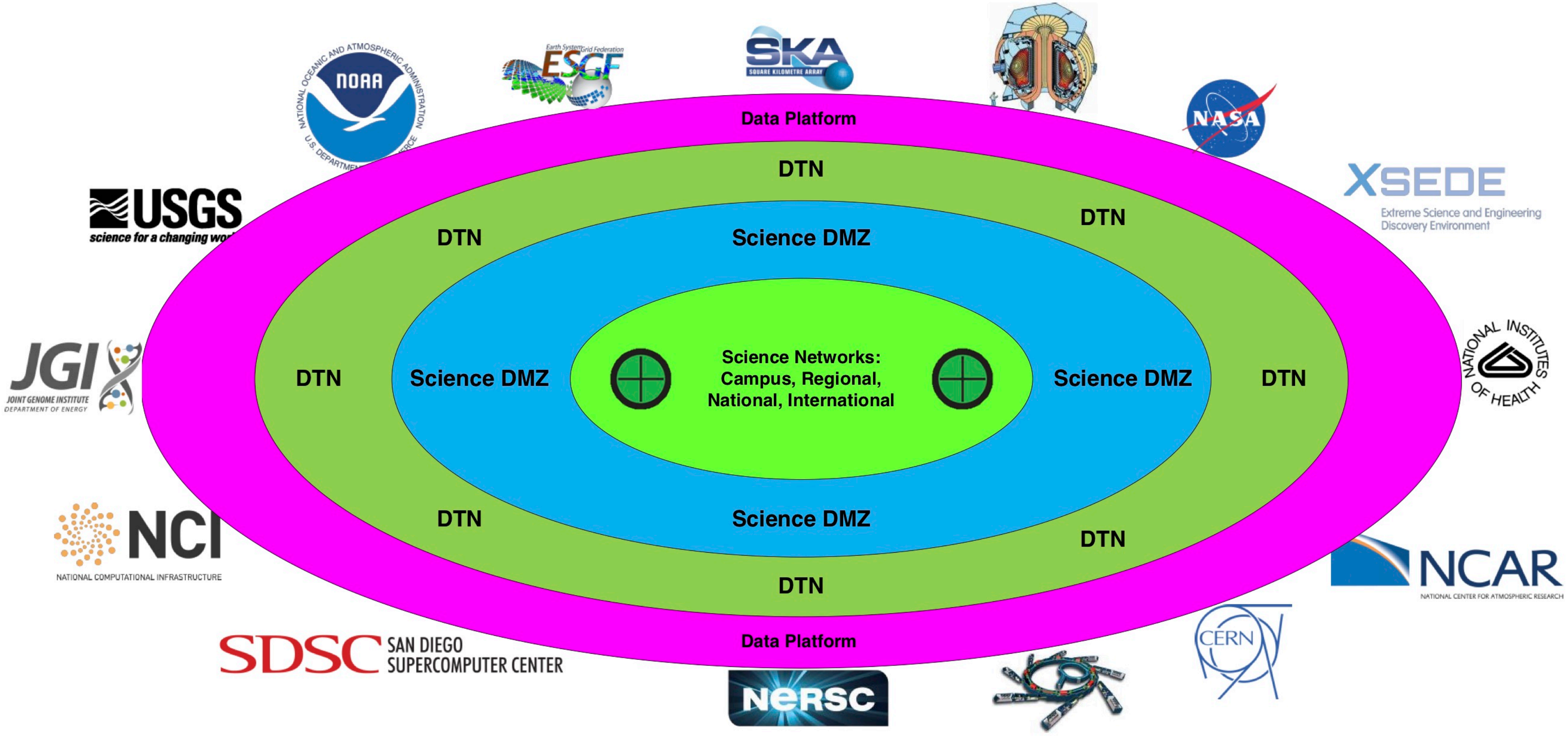
Science DMZs for Science Applications



ESnet

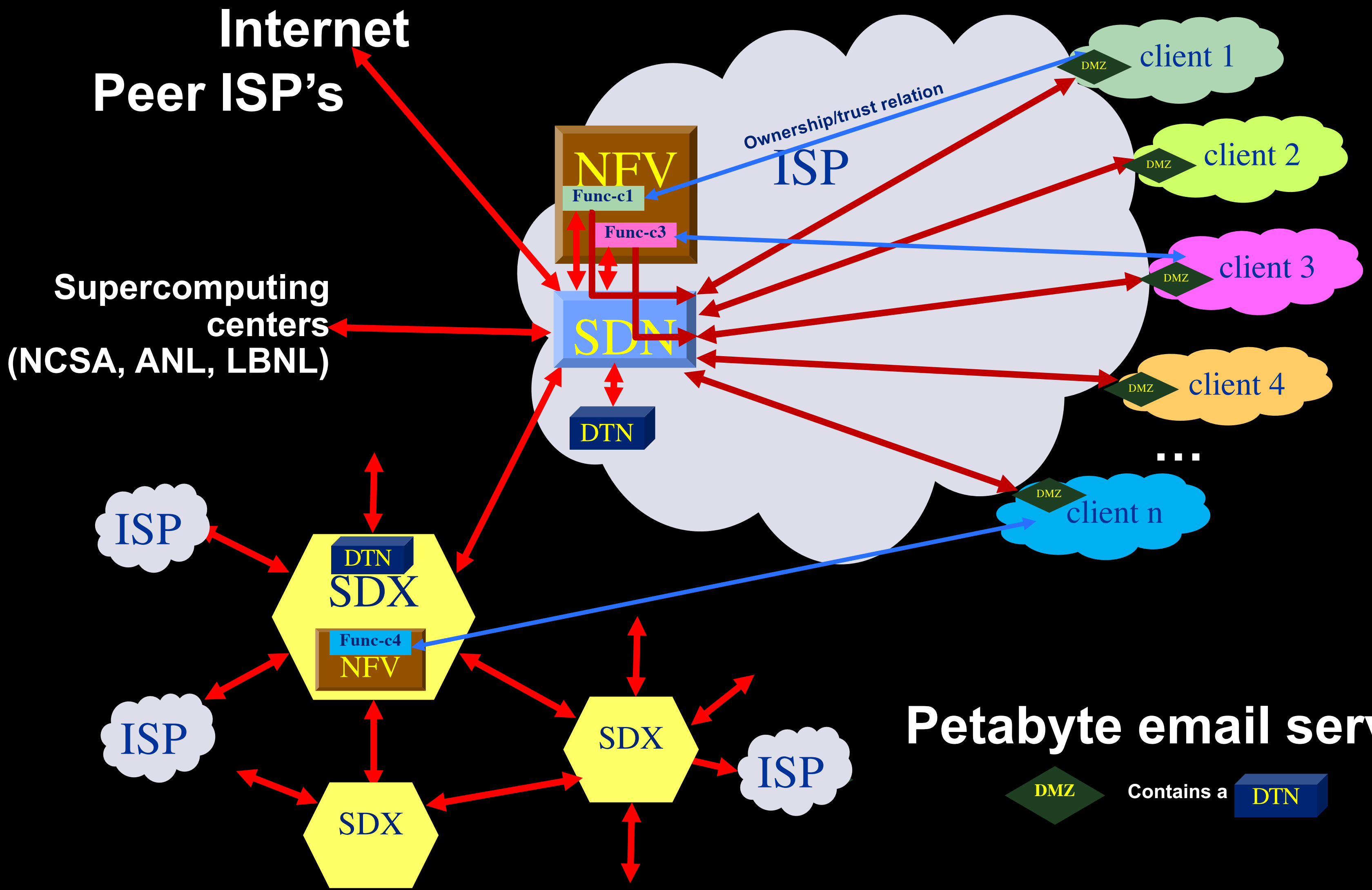
Courtesy Eli Dart, ESnet

Data Ecosystem – Concentric View



Courtesy Eli Dart, ESnet

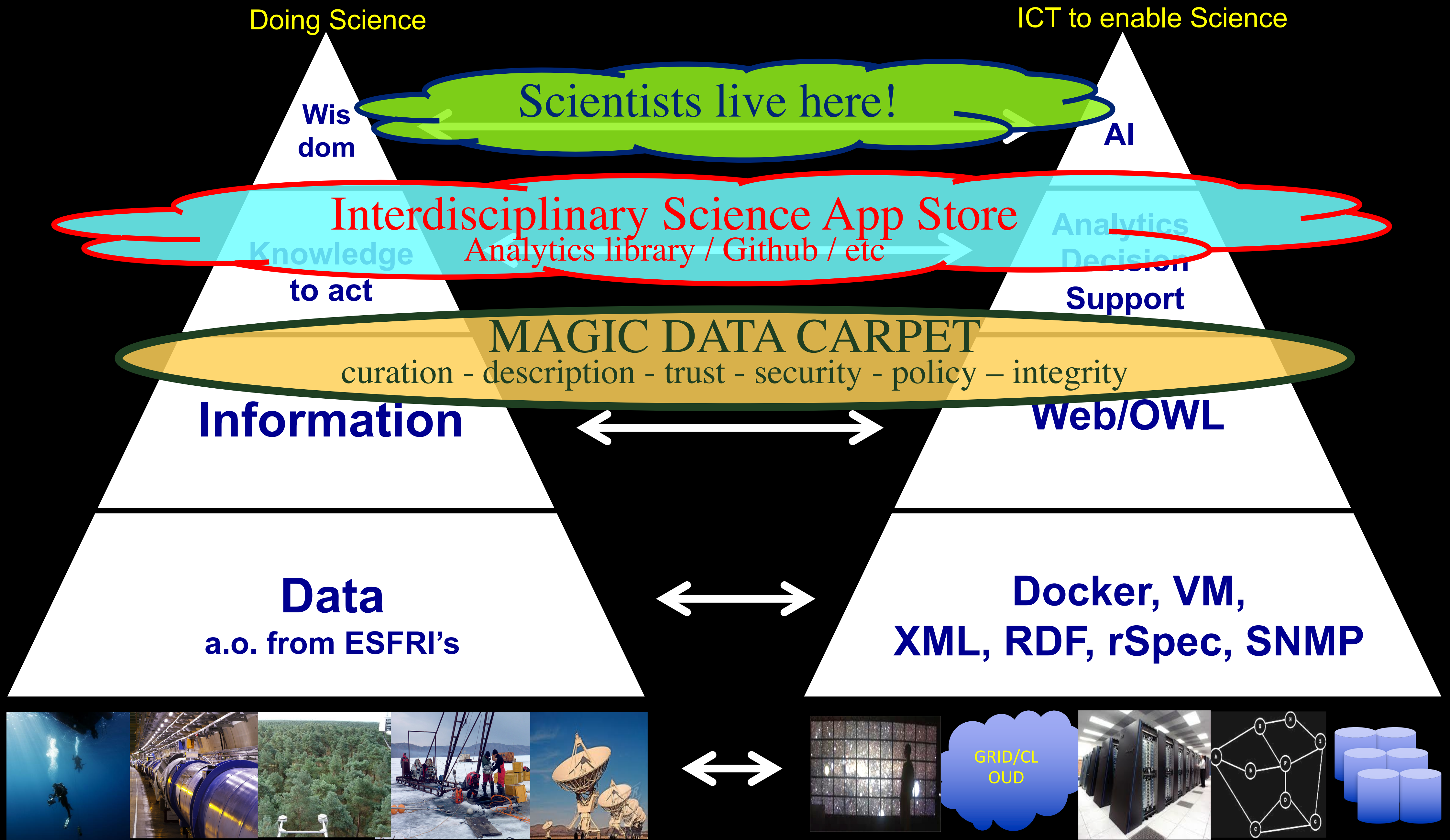
Networks of ScienceDMZ's & SDX's



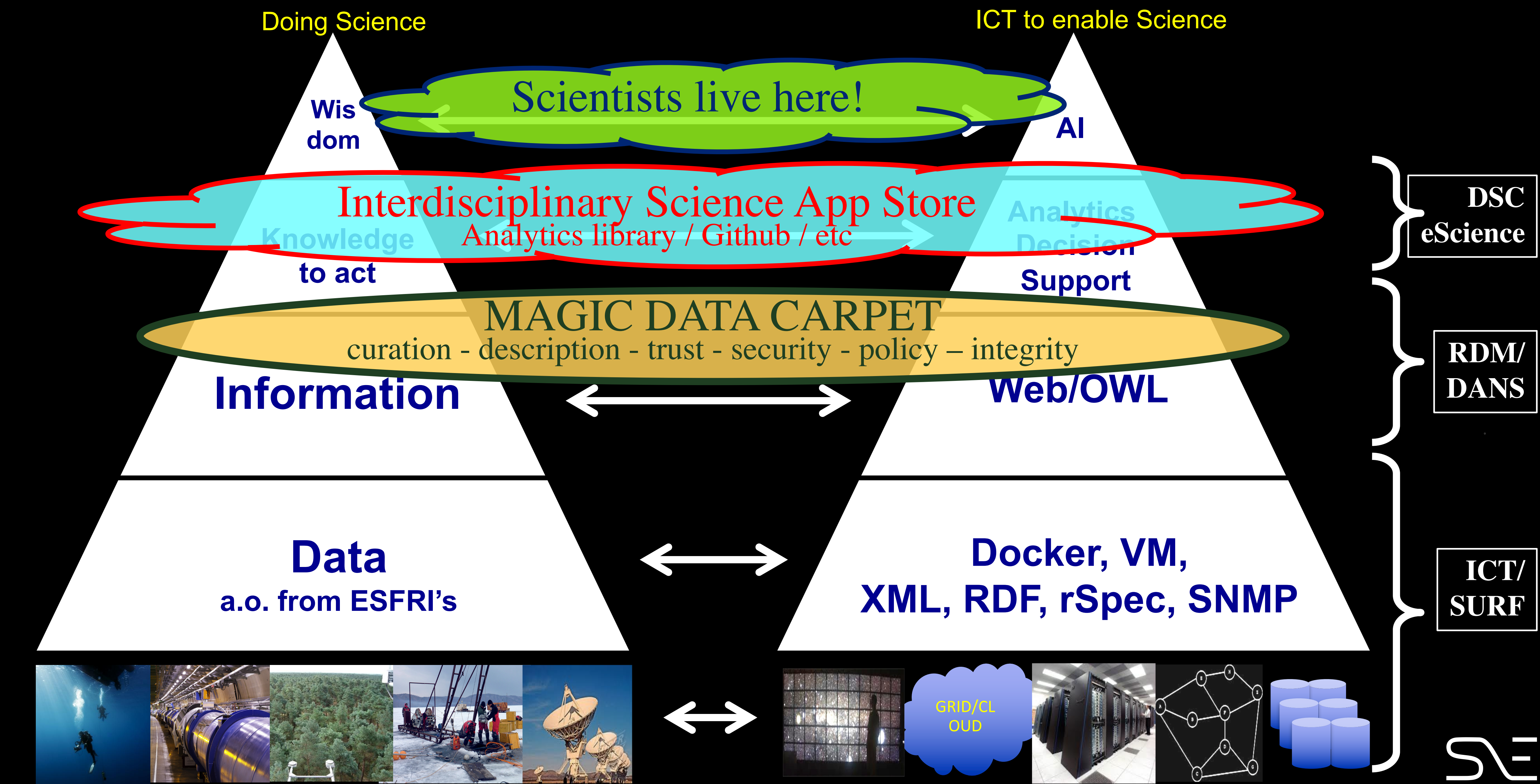
Petabyte email service ☺

DMZ Contains a DTN

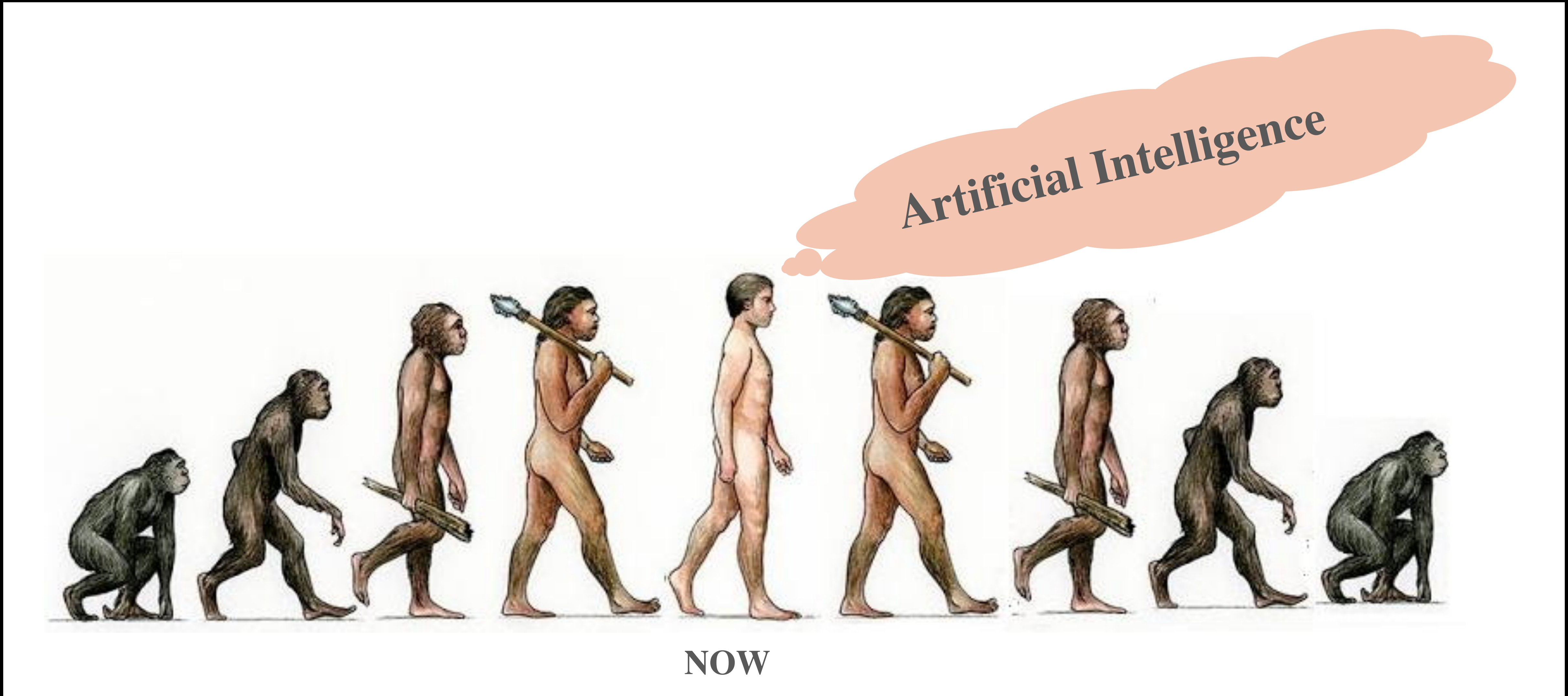
The Big Data Challenge



The Big Data Challenge



AI forking off



Conclusions, Info, Acknowledgements, Q&A

- Data hindered by risk of unexpected use, lack of trust
- Using market principles, enforcement and determining incentives and value in the data life cycle to make data flow
- More information:
  - http://delaat.net/dl4ld http://delaat.net/epi
  - https://www.esciencecenter.nl/project/secconnet
  - https://towardsamdex.org