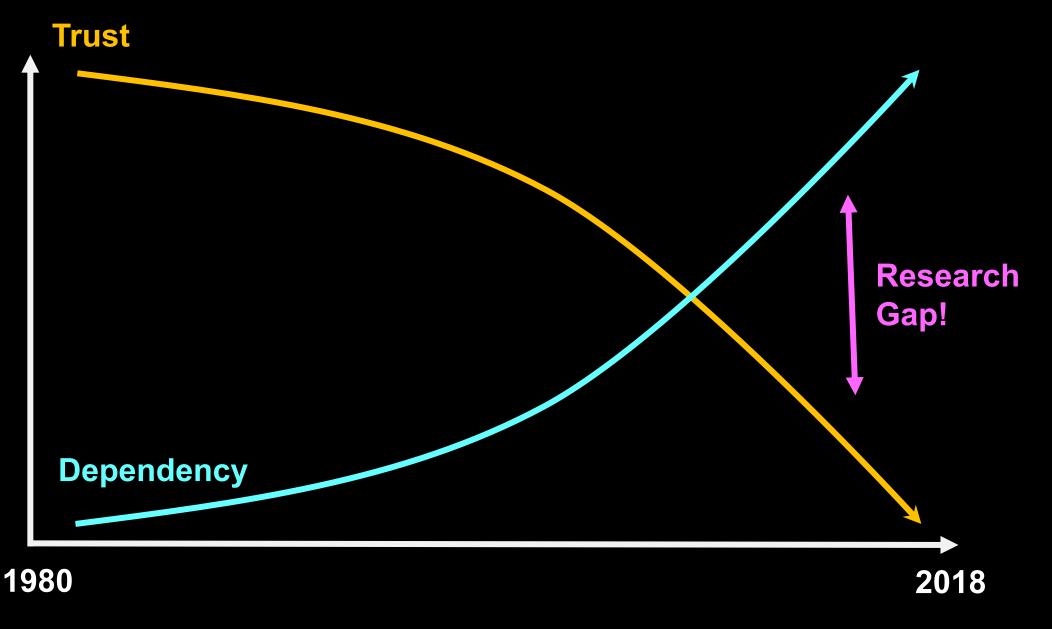
Digital Data Markets: Trusted Data Processing in Untrusted Environments

Cees de Laat

Systems and Networking Laboratory University of Amsterdam

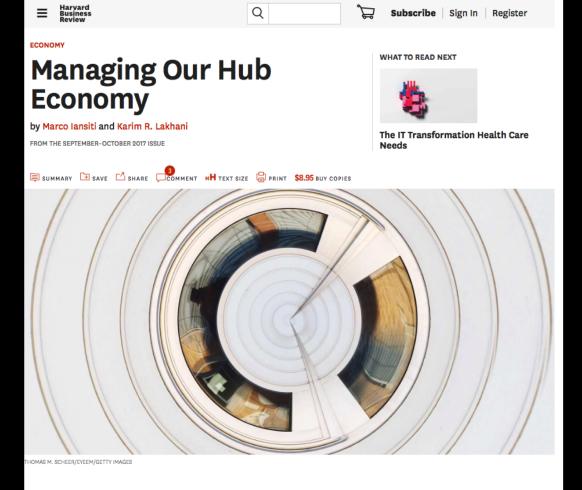
# Fading Trust in Internet



SE

# Harvard Business Review





I. The Problem

The global economy is coalescing around a few digital superpowers. We see unmistakable evidence that a winner-takeall 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. Data value creation monopolies

# Create an equal playing field

 $\rightarrow$ 

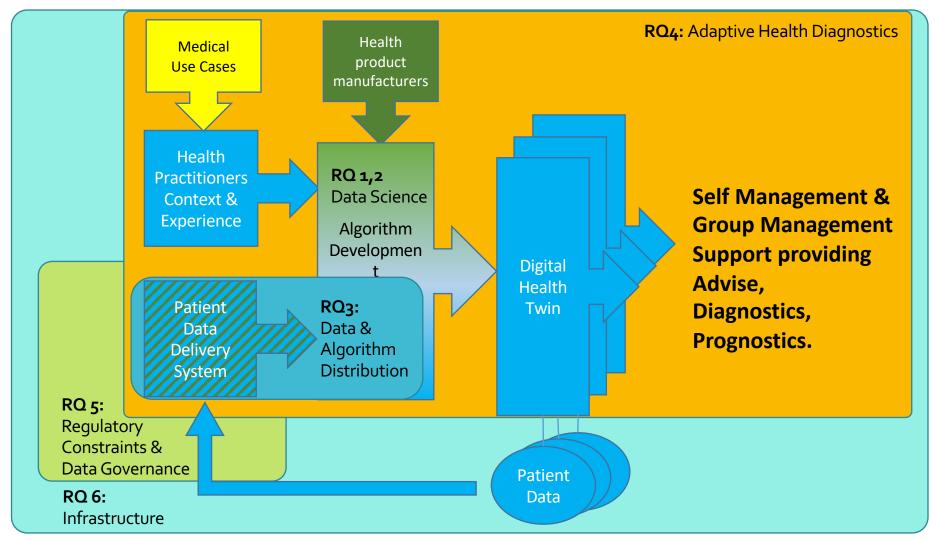
# Sound Market principles

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

## Main problem statement

- Organizations that normally compete have to bring data together to achieve a common goal!
- The shared data may be used for that goal but not for any other!
- Data may have to be processed in untrusted data centers.
  - How to enforce that using modern Cyber Infrastructure?
  - How to organize such alliances?
  - How to translate from strategic via tactical to operational level?
  - What are the different fundamental data infrastructure models to consider?

### Health use case Enabling Personal Interventions



### **Big Data Sharing use cases placed in airline context**

**Global Scale** 



City / regional Scale



Campus / Enterprise Scale



NLIP iShare project



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

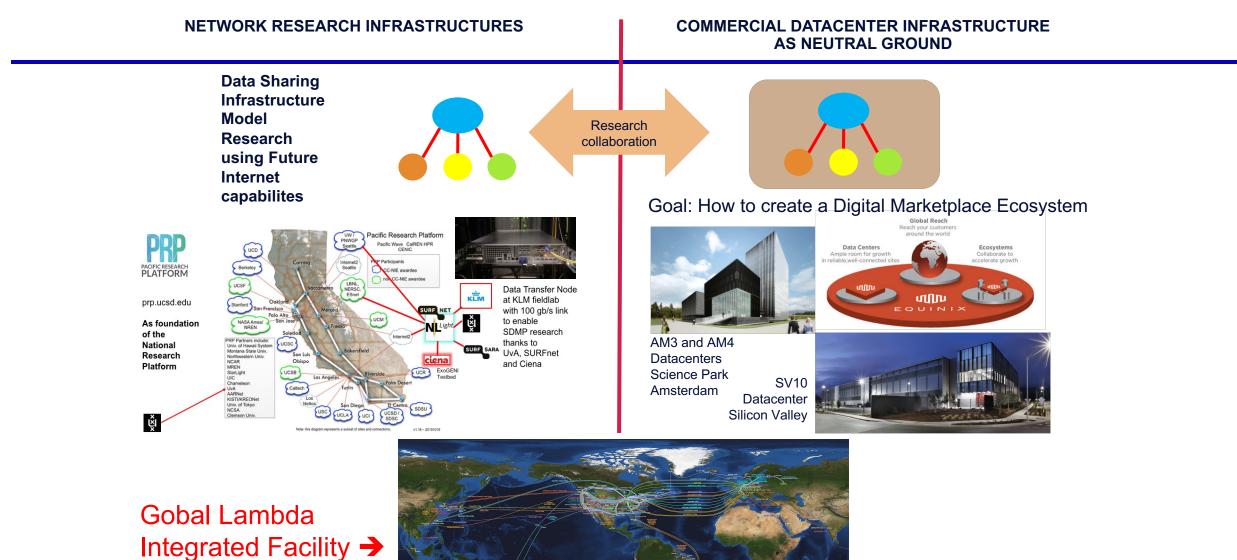


Cybersecurity Big Data NWO COMMIT/ SARNET project 3.5 FTE



AIR FRANCE KLM

### RESEARCH WORKING ALONGSIDE IT INDUSTRY



# 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!)

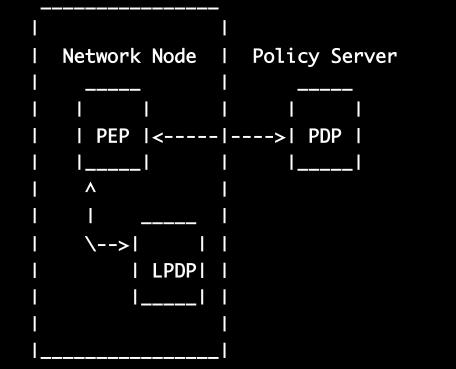


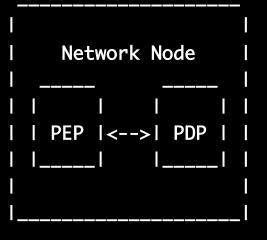
# Data Processing models

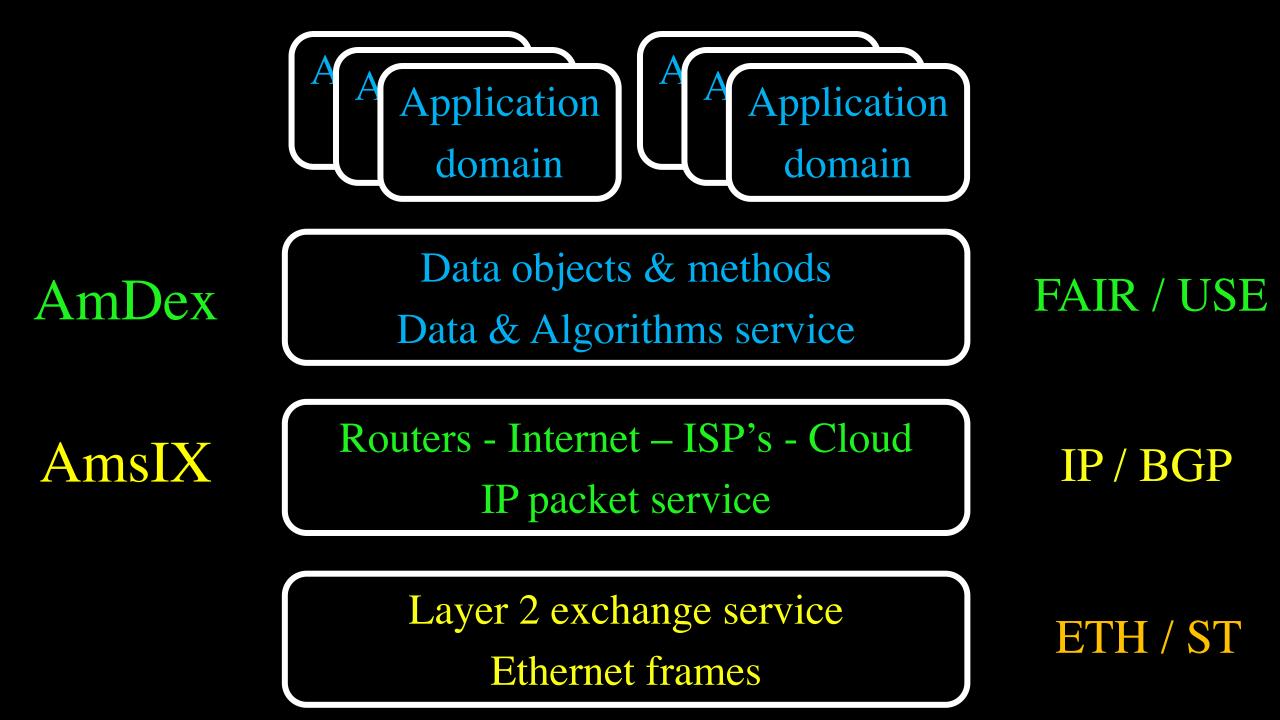
- Bring data to computing
- Bring computing to data
- Bring computing and data to (un)trusted third party
- A mix of all of the above
- Block chain to record what happened
- Block chain for data integrity
- Bring the owner of Data in control!
- Data owner policy + enforcement technology

# IETF: Common Open Policy Service (COPS)

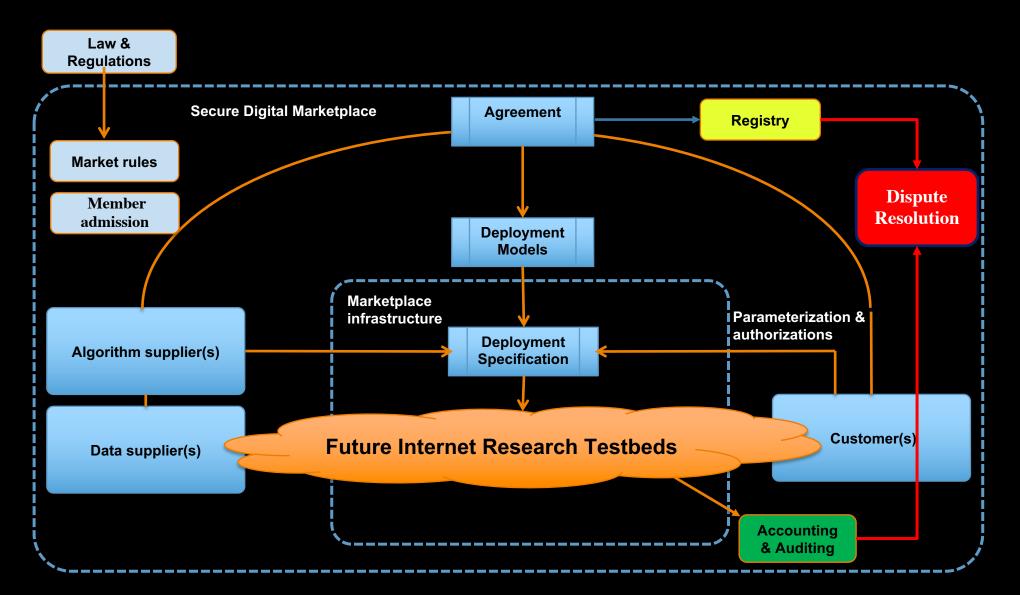
• Rfc 2748, 2753, 4261







### **Secure Digital Market Place Research**



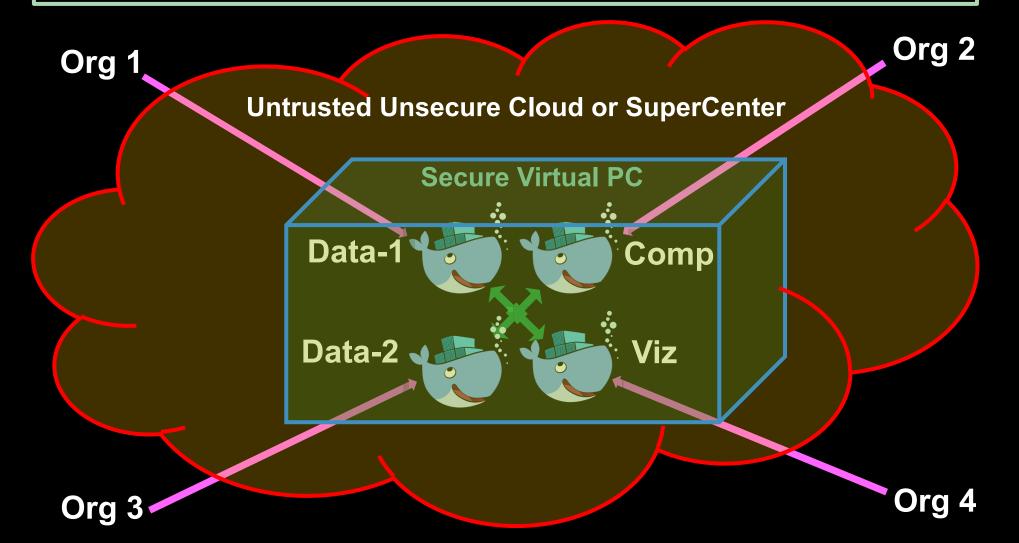




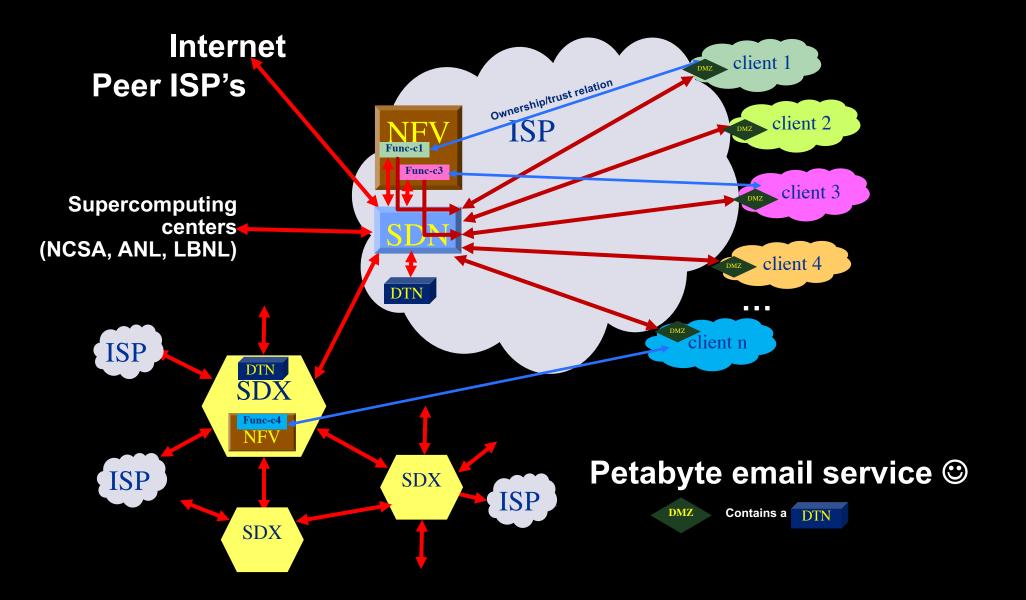
## Secure Policy Enforced Data Processing



- Bringing data and processing software from competing organisations together for common goal
- Docker with encryption, policy engine, certs/keys, blockchain and secure networking
- Data Docker (virtual encryped hard drive)
- Compute Docker (protected application, signed algorithms)
- Visualization Docker (to visualize output)



### Networks of ScienceDMZ's & SDX's



SĒ

### 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.
- Working on a demo for SC18 in Dallas TX, 11-16 Nov.
- Tactical operation of Digital Data Markets
- 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

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



**Problem Description** Scientific datasets are usually made publicly available ....but data cannot always leave the Problem Approach organization premises On-site data processing can be challenging because of incompatibility of systems or Data leaving lack of manpower ne organizati Can a container-based system perform remote on-site data processing efficiently? Send container to the organization What are the networking issues to solve? **Underlay and Overlay** Main features: Networked containers VXLAN overlav 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 Bring Your Own Contai 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? http://byoc.lab.uvalight.net/info

http://sne.science.uva.nl/sne/gigaport3
http://delaat.net/sc

More information:



# SC17 Posters and proof of concepts & demo's

#### http://sc.delaat.net/sc17

#### Unlocking the Data Economy via Digital Marketplaces Researching governance and infrastructure patterns in airline context

#### Use Case: Sharing Aircraft Data to develop a Maintenance Credit System



- A Digital Twin estimates time before maintenance is needed after data is received from a corresponding aircraft system.

Algorithm quality increases when data, owned by different airline operators, can be shared during its development.

Sharing data assets carries risk (e.g. non-compliancy).

Research Question: "Can Digital Marketplace concepts organize trust amongst its stakeholders to enable common benefits no single organization can achieve, whilst observing economic principles?".

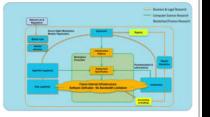
#### Digital Marketplace as a means to organize trusted data asset sharing

A Digital Marketplace is a membership organization identified by a common goal: Share data to enable development of a Maintenance Credit System.

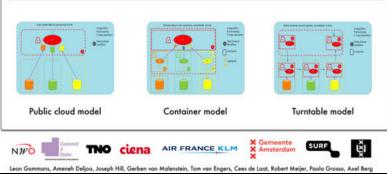
Membership organization is institutionalized to create, implement and enforce membership rules.

Market members create digital agreements.

Agreements are translated into different software defined infrastructures using infrastructure patterns offered by a Digital Marketplace Ecosystem.



#### Examples of infrastructure patterns offered by a Digital Marketplace



#### DEMONSTRATION: LIGHT PATHS AND DATA TRANSFER NODES FOR AIRCRAFT MAINTENANCE

It StarLight - connect to each other using light paths over their links. In this demonstration, users at SC17 in Denver will experience the difference in file transfer rates with and without using OTNs.

#### USE CASE: AIRCRAFT MAINTENANCE

Besides people and luggage, anonafts transport data they generate, like flight information technical statistics and sensor readings. These data tell pilots and engineers if the aircraft's critical systems are doing their job safely. When data are transferred and analyzed rapidly defects can be solved more quickly, possibly even while the aircraft is waiting at the gate. When receiving the data within minutes, expert engineers in a remote airport can rapidly verify with the home base engineers if an engine vibration warring was caused by the engine or by a failing sensor.

#### INTERNET VS LIGHT PATH

Air France-HLM uses a 300 Gbit/s light gath and researches its benefits. Using light gaths you can transport huge amounts of data at high speed and with a guaranteed bandwolth between 2 points. When using high volumes of data, the current Air France-KUM's Internet connections are not misate or fast ensuch to transfer the data within the new estert time. frame. Transferring a tenatiyte of engine data via the current internet connections would take around 30 hours: with a 300 Gbit/s light path this could have less than 2 minutes.

#### AIR FRANCE-KLM CONNECTED TO NETHERLIGHT

Ciene and SURF facilitate the connection from Air France-KLM at Schiphol to NetherLight. SUTTY's European hub for International light paths in Amsterdam SUTT provides the 100 Gbit/s light path from Air France-KUM via NetherLight to the avoraft's destination. For this demonstration, the location is Starlight in Chicago, a hub similar to NatherLight.

#### DATA TRANSFER NODES

Data Transfer Nodes are high-performance systems that are optimized to transfer huge amounts of data. The interconnects between these systems exist of high-capacity dedicated bandwidth, removing network bottlenecks within the mesh of global DTNs. To date. DTNs are present on a small scale, e.g. a couple per continent. By copying a file from an end user system directly into the nearest DTN, the global DTN system sends the file to the DTN nearest to the final file destination, optimizing the process of high-latency international transfers.





#### ciena

#### AIRFRANCE KLM

#### Data Transfer Node (DTN) Workflows

Joseph Hill, Gerben van Malenstein, Cees de Laat, Paola Grosso, Leon Gommans

#### Why Data Transfer Nodes (DTNs)

- DTNs can act as an interface to a high performance link
- · Configured to maximize performance for a given workflow
- · Simplifies configuration of client systems
- · Multiple clients may share a DTN
- · DTNs strategically placed to best benefit clients
- · DTNs can be compared to specialized high speed transport systems of the past

# Pneumatic Tube Messaging System, 1943

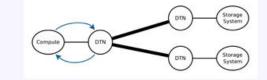
#### Example: Entry Point for High Speed Transport

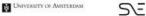
A typical use case for DTNs is as a high speed file transfer service. A computer system's configuration may allow for the utilization of all available bandwidth in a LAN environment. However, it is often the case that in a WAN environment with high latency or packet loss the same system performs poorly. A DTN could be tuned to maximize performance on a high latency path. It could also use specialized transfer protocols to mitigate high packet loss. The DTN may also have access to an optimized path such as a light path. Files destined for a distant receiver would be first sent to a DTN located on the same LAN as the sender. That DTN would then forward it at high speed to a DTN near the receiver. That DTN would then forward it to the final destination.



#### Example: Storage Access Point

Another possible use case for DTNs is to be used to access distributed data from remote locations. In this scenario a system located at a compute facility requests the data from the local DTN as it is required. That DTN would then transparently retrieve the data from multiple remote sites as needed. In contrast to the first example here block level access is provided by the DTNs. To the system performing the computations the nearby DTN appears to be the actual and only storage system. This hides both the remote and distributed nature of the data. While the compute side DTN may perform some caching, there need not be permanent storage of data at the compute facility.









to a rearby Data Transfer Node. Visitors are allowed to transfer pre-prepared datasets between the systems via the DTNs with graphs showing various performance metrics. As a comparison. The performance of a direct connection between the two systems - without using DTNs - will also be shown. The intention is to show that systems not optimized for long distance transfers can benefit from using nearby DTNs to facilitate the transfer and decreasing file transfer time.

DESEARCH IN OTHER INDUSTRIES.

in addition to the aircraft research, high bandwidth, low latency light paths offer poss research in other industries as well. For example, fundamental research on data transfer protocols suitable for these bandwidths can also help excel diagnosis by doctors when they can have access to terabytes of patient and other related research data within minutes instead of days or weeks. Imagine what this would enable other research disciplines to do truy Doceabilities are almost infinited

Non information was sufficien/200 G-Ar-Engres-KLM

# 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.

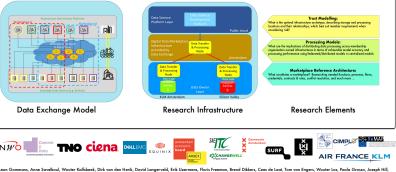
Membership organization is institutionalized to create, implement and enforce membership rules organizing trust.

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



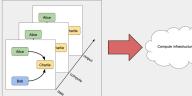
Savelkoul, Wouter Kalfsbeek, Dirk van den Herik, David Langerveld, Erik Uzermans, Floris Freeman, Brend Dikkers, Cees de Laat, Tom van I , Giovanni Sileno, Lu Zhang, Ameneh Deljoo, Thomas Baeck, Willem Koeman, Laurie Strom, Axel Berg, Gerben van Malenstein, Kaladhar V

#### 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



workflows

rule registry

rule auditing

spec complaint service

proanizations trust

users specs auditor

planner

rule enforcemen

standards

NWO

contracts

infrastructure

compute infrastructure data registry

#### 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. applications

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.

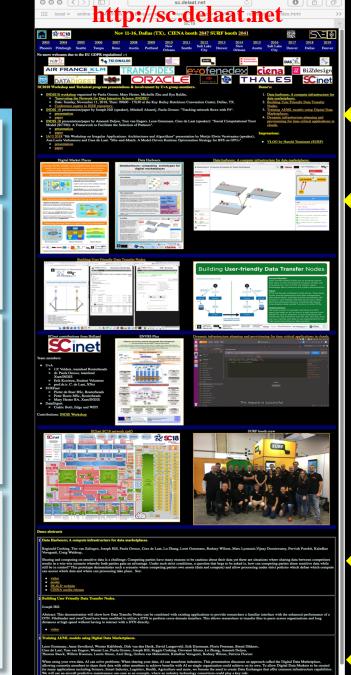
Contracts: Are a set of rules that are shared amongst participating harbours which describe how objects (data, compute) can be traded between harbours and who can process data. In its simplest form is a 7-tuple which binds a user. data object, compute container, contract, consortium, harbour, and expiry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

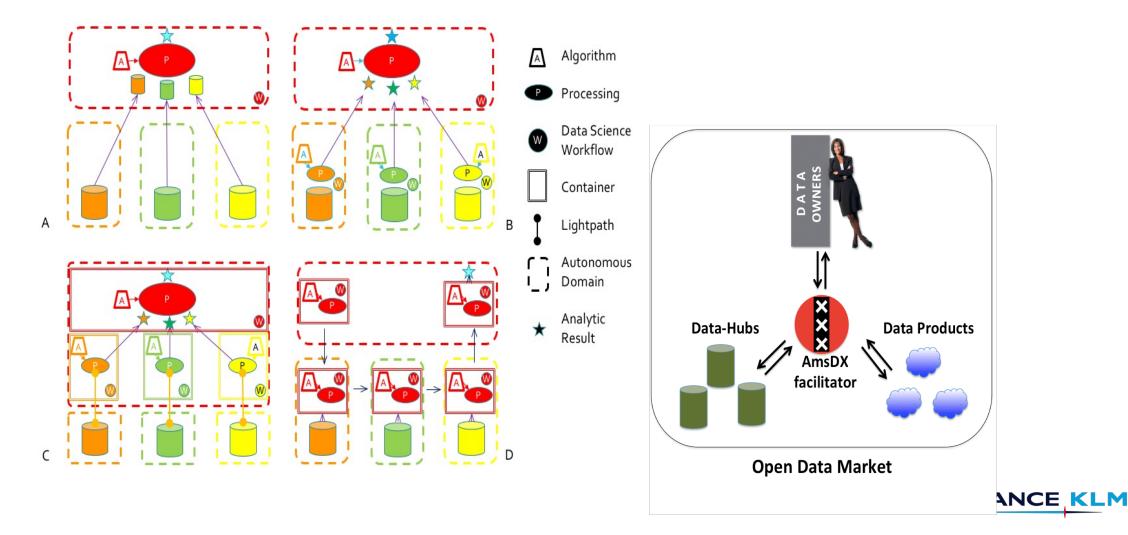
#### In action DataHarbours: Computing Archetypes for Digital Marketplac Federated computing on 3 distributed data harbours. Here we illustrate one archetype where KLM and Airfrance do not trust each other and employ a trusted 3rd party to send the data and compute for processing. **6 - 1** For the scenario to succeed the different harbours need to effect several transactions which are Legend governed by contractual rules. Comput The transaction protocol involves first - Deta 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.



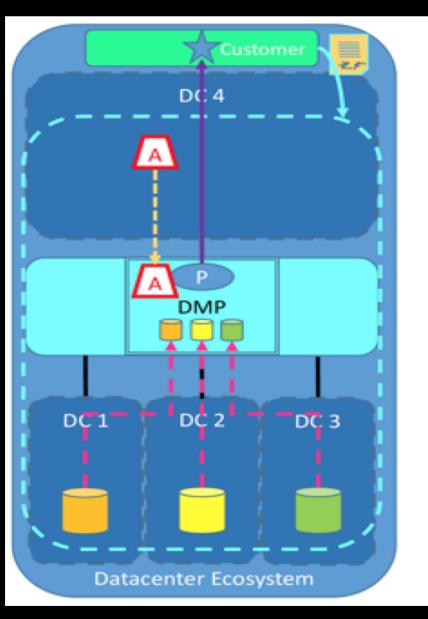


### **INFRASTRUCTURE PATTERN EXAMPLES**

#### OFFERED BY A DATA EXCHANGE TO MARKETPLACES TO CHOOSE FROM



# DMP archetypes and their representation



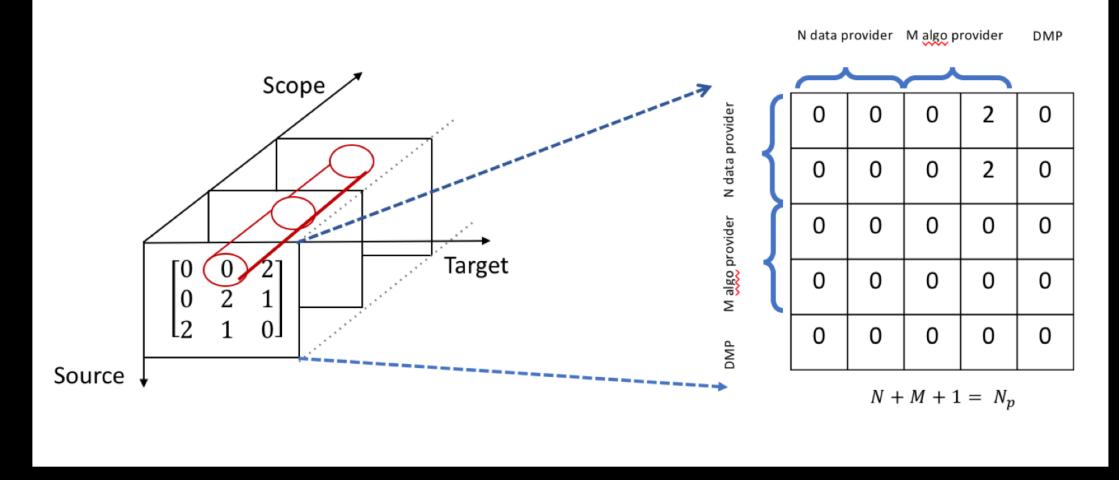


On the left one of the many collaboration models within a DMP. We call this archetype. One DMP can support multiple archetypes depending on the contracts between partners.

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

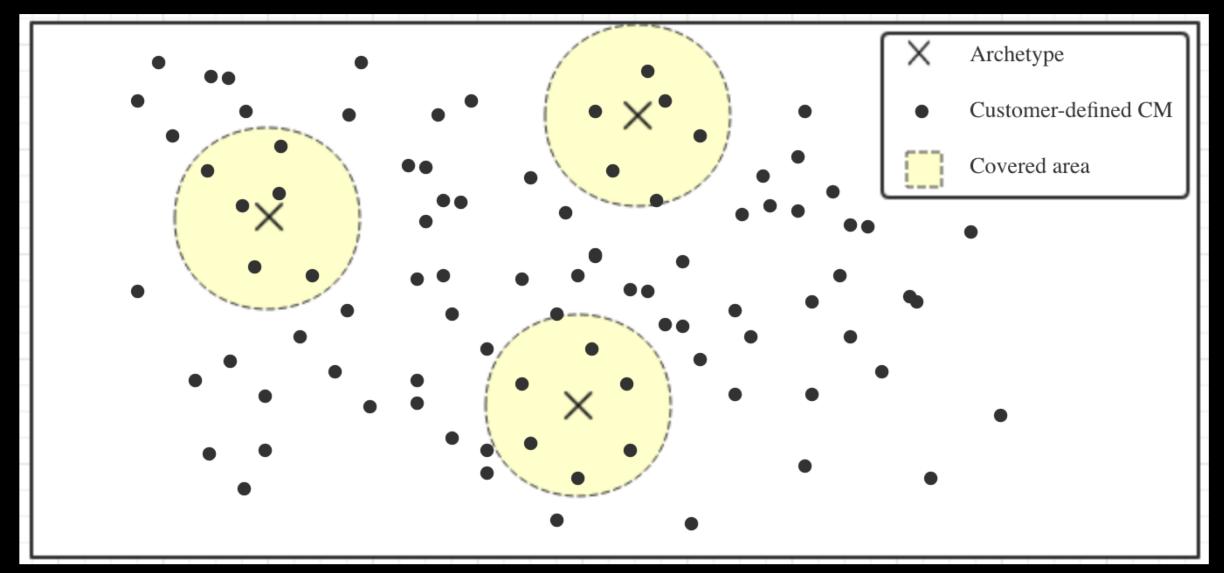
S

# Models for archetypes



Parties in the DMP collaborate across a number of scopes: data, computing and output. They share data, they share algorithms and they can share results. The matrix on the left represent the level of collaboration between two parties in each of the scope. In the previous slide we had four parties plus the DMP exchange, so we have a 5x5 matrix.

# Matching requests



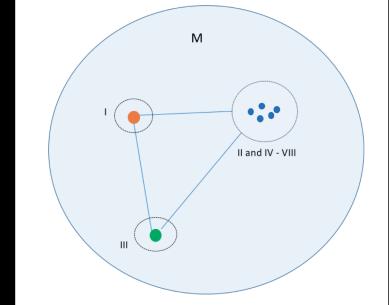
Our work is to match a customer application to the 'closest archetype'.

# Dimensions of DMP

### Coverage

### Extensibility

How well can we satisfy users request with the available archetypes?



### Precision

Flexibility

How well the archetype database of a given DMP fits a request by customer?

How easily the requests from potential customers could be satisfied?

Can a DMP provide more archetypes to user?

We defined four metrics to determine the 'richness' of a DMP.

- More information:
  - -<u>http://delaat.net/dl4ld</u> and <u>http://delaat.net/epi</u>
  - <u>https://towardsamdex.org</u>
- Contributions from:
  - Leon Gommans, Wouter Los, Paola Grosso, Yuri Demchenko, Lydia Meijer, Tom van Engers, Reggie Cushing, Ameneh Deljoo, Sara Shakeri, Lu Zhang, Joseph Hill, Lukasz Makowski, Ralph Koning, Gleb Polevoy, Tim van Zalingen, and many others!





### **Data Hub System Applicability**

### Industry

- Cross Cutting Field lab
- Innovation with SURF

### Science

- European Open Science Cloud
- FAIR model
  - Findable Accessible Interpretable Reusable





### Society

- Smart Cities & Arena
- Streaming Data Decision Support



# **Experimental Setup**

### Data Transfer nodes at UvA, KLM and Equinix Running Kubernetes with a number of dockers (pods) see below.

tim@uva-kube-04:~\$ kubectl get pods -o wide							
NAME	READY	STATUS	RESTARTS	AGE	IP	NODE	NOMINATED NODE
be2-deployment-c87646848-wt815	1/1	Running	Θ	77m	192.168.5.39	eqx-kube-03	<none></none>
mq1	1/1	Running	Θ	13d	192.168.1.2	uva-kube-02	<none></none>
oex.airfrance	1/1	Running	Θ	10d	192.168.4.16	eqx-kube-02	<none></none>
oex.klm	1/1	Running	Θ	10d	192.168.3.5	eqx-kube-01	<none></none>
oex.trusted	1/1	Running	Θ	13d	192.168.1.3	uva-kube-02	<none></none>
planner1	1/1	Running	Θ	10d	192.168.4.15	eqx-kube-02	<none></none>

- be2 is the backend for the website.
  - It serves the static pages and passes new information and input to the planner.
- mq1 is the message queue that each oex writes logs to.
- oex.\* each is one zone or 'object exchange server'.
  - It is responsible for handling requests from others and sending requests to other parties.
  - It should do so in accordance with a preselected contract/archetype.
- planner1 handles requests from be2 when selecting an archetype (and passes it on to the oex's) and when an application or pipeline is started (and passes it on to the oex).

DataHarbours: Computing archetypes for digital marketplaces

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

### SAE Use Case envisaged research collaboration

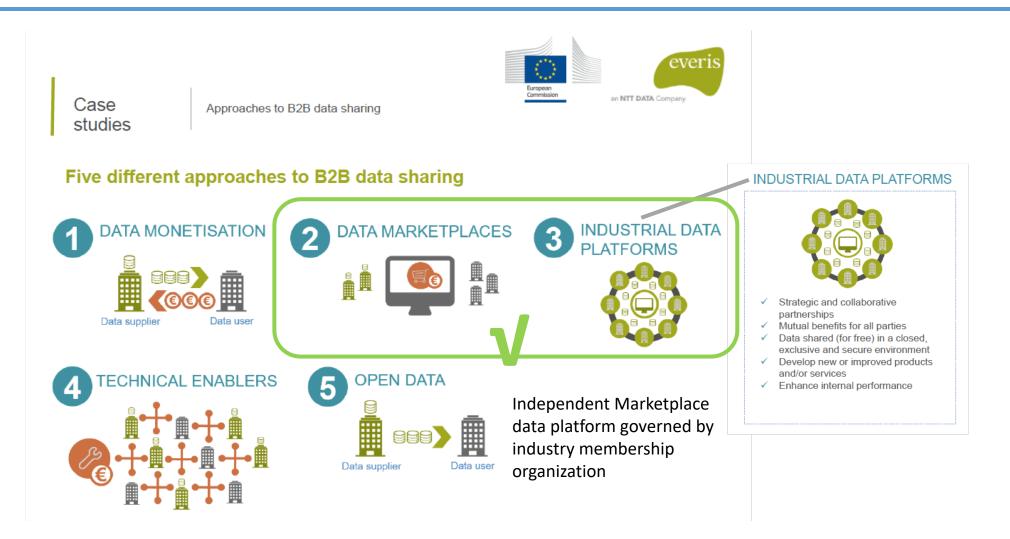


SSE System and Network Engineering

### AIR FRANCE KLM

### data Sharing approach: Combine 2&3

#### MANAGED BY AN INDEPENDENT INDUSTRY MEMBERSHIP ORGANIZATION



https://publications.europa.eu/en/publication-detail/-/publication/8b8776ff-4834-11e8-be1d-01aa75ed71a1/language-en