

# 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!

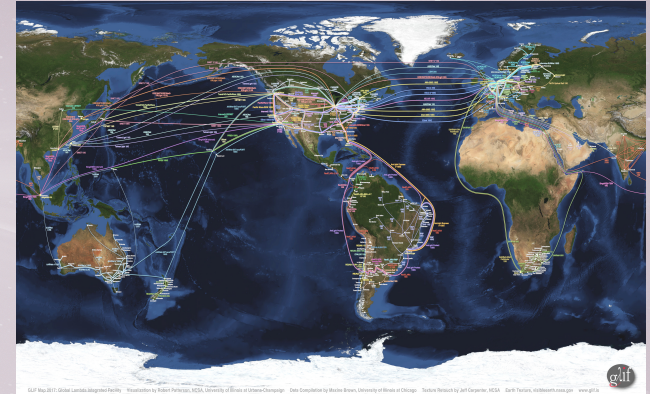


# Mission

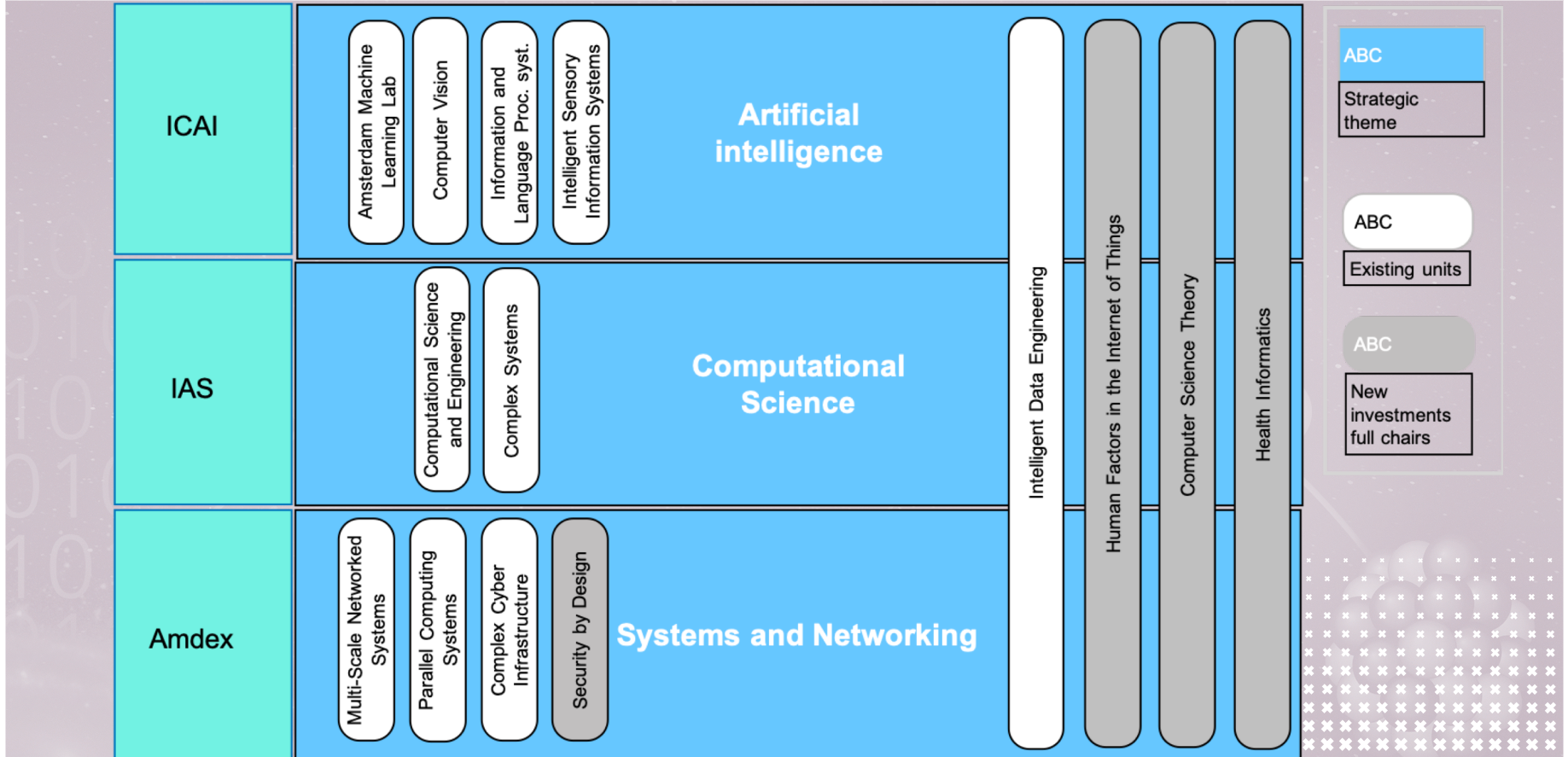
- The Systems and Networking Lab conducts research on leading-edge computer systems of all scales, ranging from global-scale systems and networks to embedded devices.
- Across these multiple scales our particular interest is on extra-functional properties of systems, such as performance, programmability, productivity, security, trust, sustainability and, last but not least, the societal impact of emerging systems-related technologies.

## Broad spectrum of research

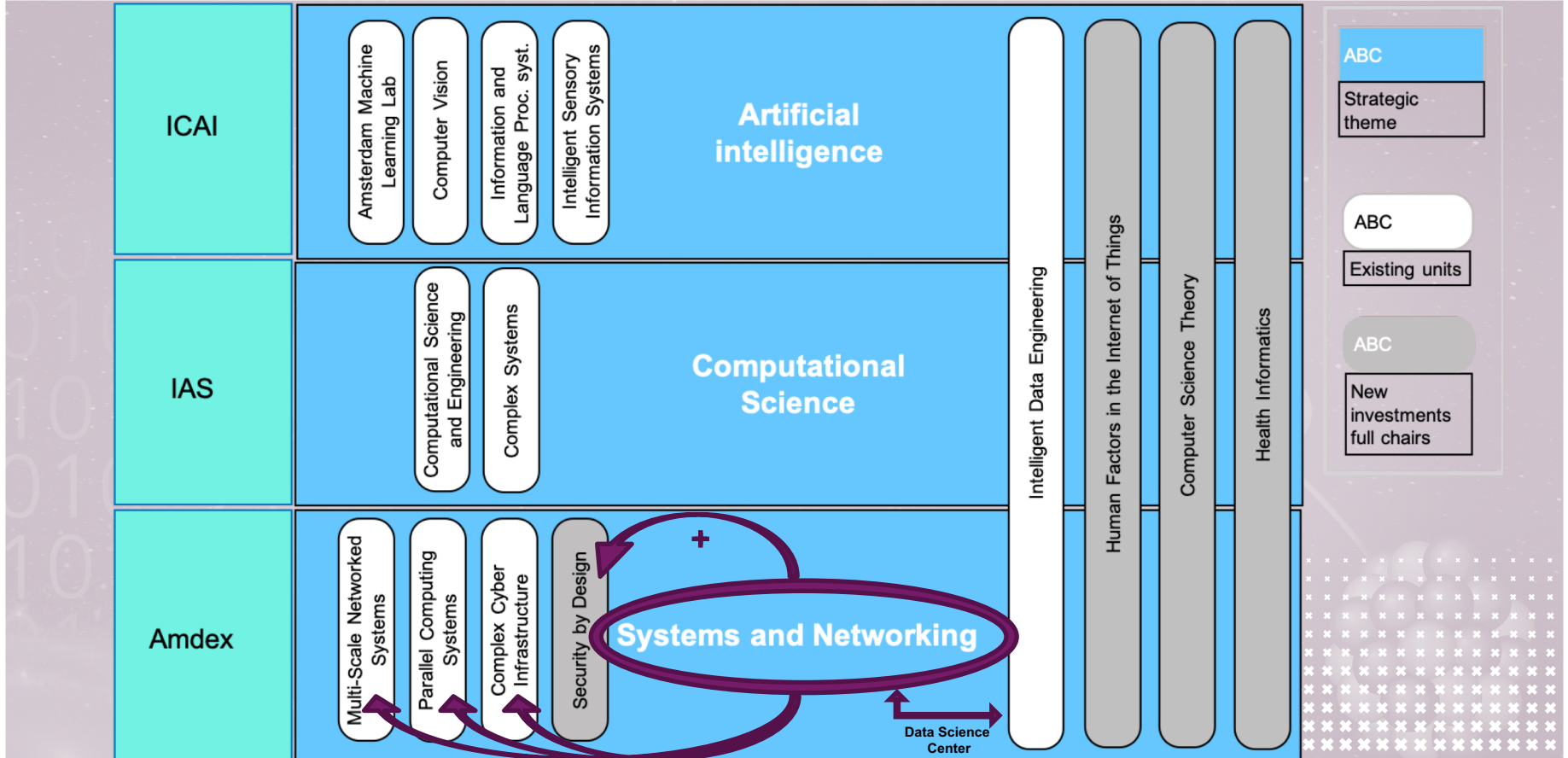
- Advanced networks / Internet architecture
- Network programmability / Overlays / Virtualisation
- Authorisation of Internet resources
- Quality of service for apps on Clouds
- Systems: Embedded / real time / parallel / design
- Performance & Compilers & ExaScale
- Safe Secure Data Sharing / Processing
- Data Sovereignty & Normative Agents & Trust → AmDEX
  
- **Well funded across the themes → ~ 10 Meuro in last 3 years, mostly from NWO / EU and smaller portion from SURF and Industry**



# Position in the Instituut



# Position in the Instituut



- Group leaders: prof.dr.ir. C. de Laat, dr. Paola Grosso, dr. Andy Pimentel
  - 1 full prof (CdL)
  - 4 associate professors (PG, AV, AP, CG)
  - 2+3 assistant professors (AB, ZZ, 3 open positions)
  - 3 part time professors (PA, PK, TvE)
  - 3 endowed professors (RvN, SK, LG)
  - 1 senior researcher (YD)
  - ~ 23 phd students, ~ 6 postdoc's
  - 2 tech support/scientific programmers
  - ~18 guests
  - Total about 65 people ~ 40 fte
  
- Yearly turnover ~ 3 MEuro research + similar education

**Totals  
staffing  
in the  
SNE  
Lab**



Photo: dr. Yuri Demchenko

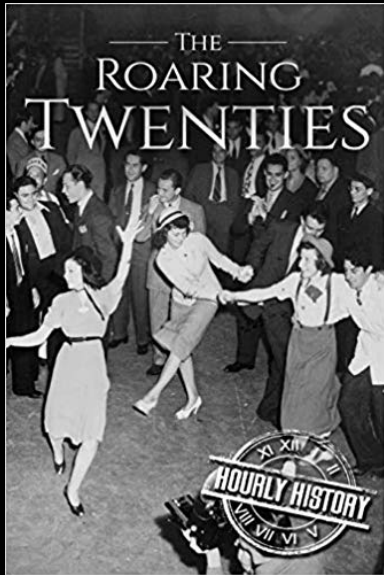


## Research based Education

- The SNE laboratory is closely working together with the SNE Master ([www.os3.nl](http://www.os3.nl)), Software Engineering Master (SE) and the Computer Science Master (CS) programs to disseminate knowledge through education.
- **Both SNE and SE masters score in the top nationwide**
- **SNE master Excellent in accreditation and**



# ICT to support the transformation of Science in the Roaring Twenties

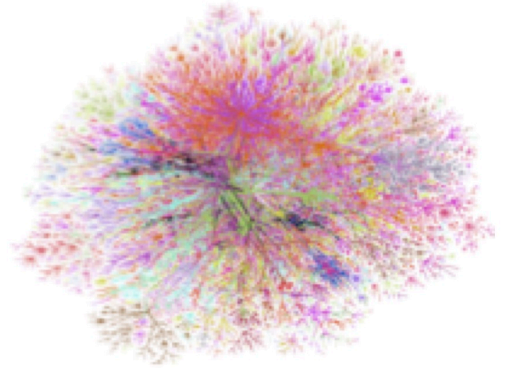


From Wikipedia: The Roaring Twenties refers to the decade of the 1920s in Western society and Western culture. It was a period of **economic prosperity** with a distinctive cultural edge in the United States and Western Europe, particularly in major cities such as Berlin, Chicago, London, Los Angeles, New York City, Paris, and Sydney. In France, the decade was known as the "**années folles**" ('crazy years'), emphasizing the era's **social, artistic and cultural dynamism**. Jazz blossomed, the flapper redefined the modern look for British and American women, and **Art Deco** peaked....

This period saw the large-scale development and use of automobiles, telephones, movies, radio, and electrical appliances being installed in the lives of thousands of Westerners. Aviation soon became a business. Nations saw **rapid industrial and economic growth, accelerated consumer demand**, and introduced significantly new changes in **lifestyle and culture**. The media focused on celebrities, especially sports heroes and movie stars, as cities rooted for their home teams and filled the new palatial cinemas and gigantic sports stadiums. In most major democratic states, women won the right to vote. The **right to vote** made a huge impact on society.



**In most applications, utilization of Big Data often needs to be combined with Scalable Computing.**



*COMPUTING AT DIVERSE SCALES*

*“BIG” DATA*

*Enables dynamic data-driven applications*



*Smart Manufacturing*

*Computer-Aided Drug Discovery*



*Personalized Precision Medicine*

*Smart Cities*



*Disaster Resilience and Response*



*Smart Grid and Energy Management*



*Precision Education*

# Workflows for Data Science Center of Excellence at SDSC

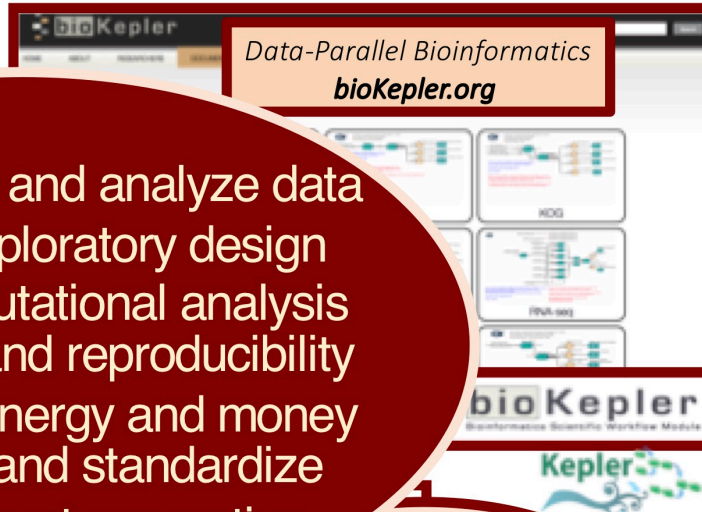


**Goal:** Methodology and tool development to build automated and operational workflow-driven solution architectures on big data and HPC platforms.

Real-Time Hazards Management  
[wifire.ucsd.edu](http://wifire.ucsd.edu)

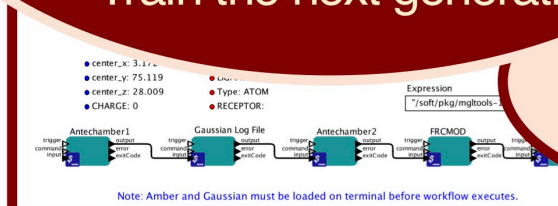


Data-Parallel Bioinformatics  
[bioKepler.org](http://bioKepler.org)



- Find, access and analyze data
- Support exploratory design
- Scale computational analysis
- Fuel reuse and reproducibility
- Save time, energy and money
- Formalize and standardize
- Train the next generation

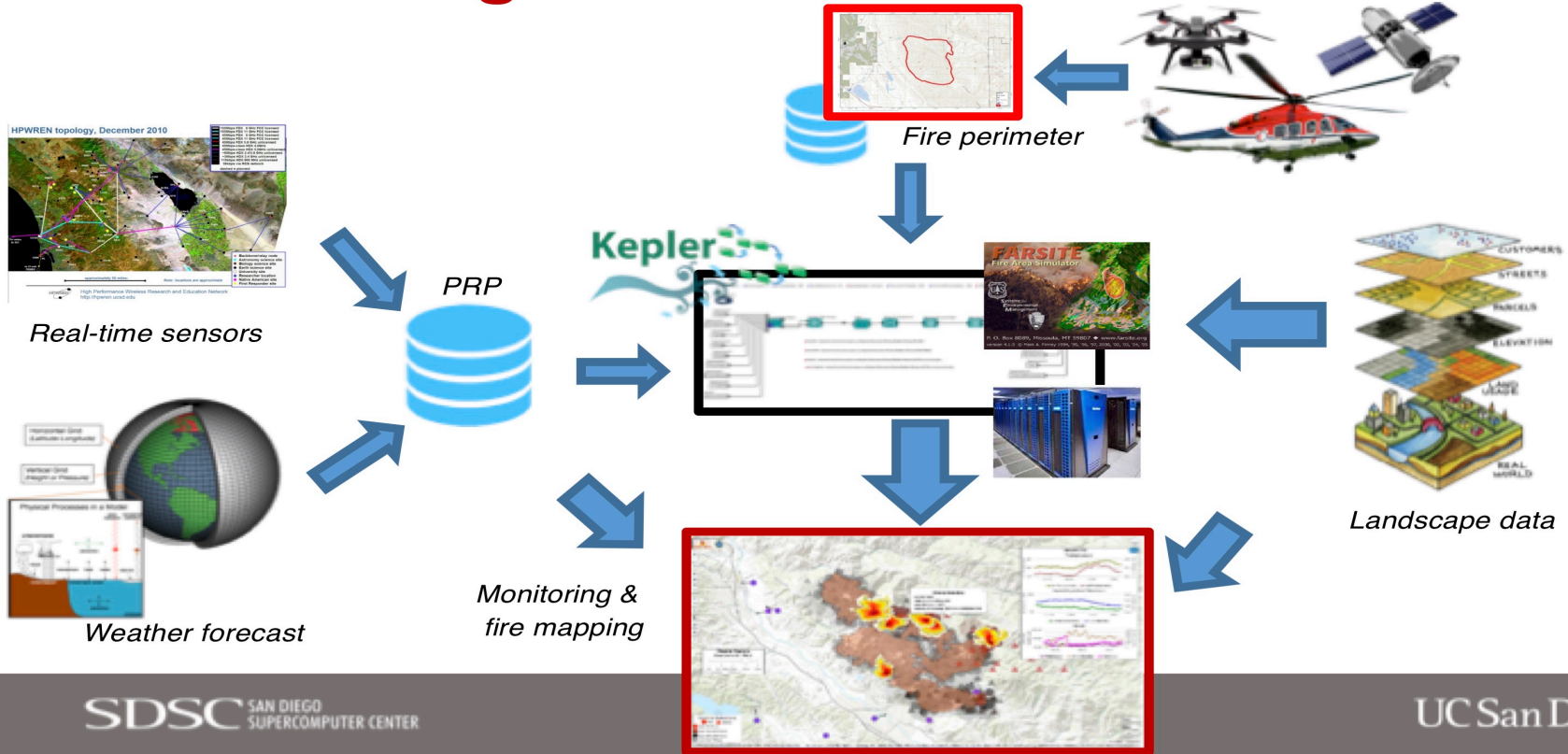
[kepler-project.org](http://kepler-project.org)



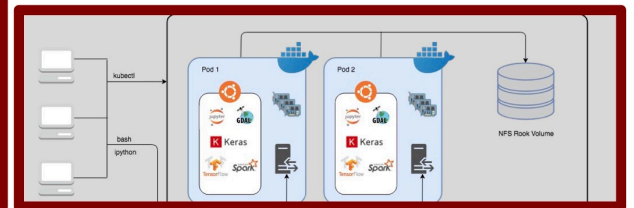
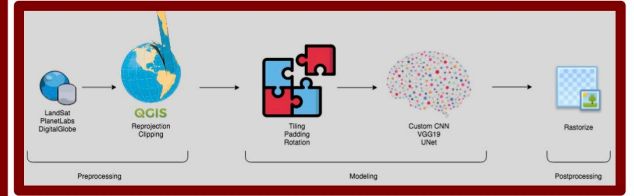
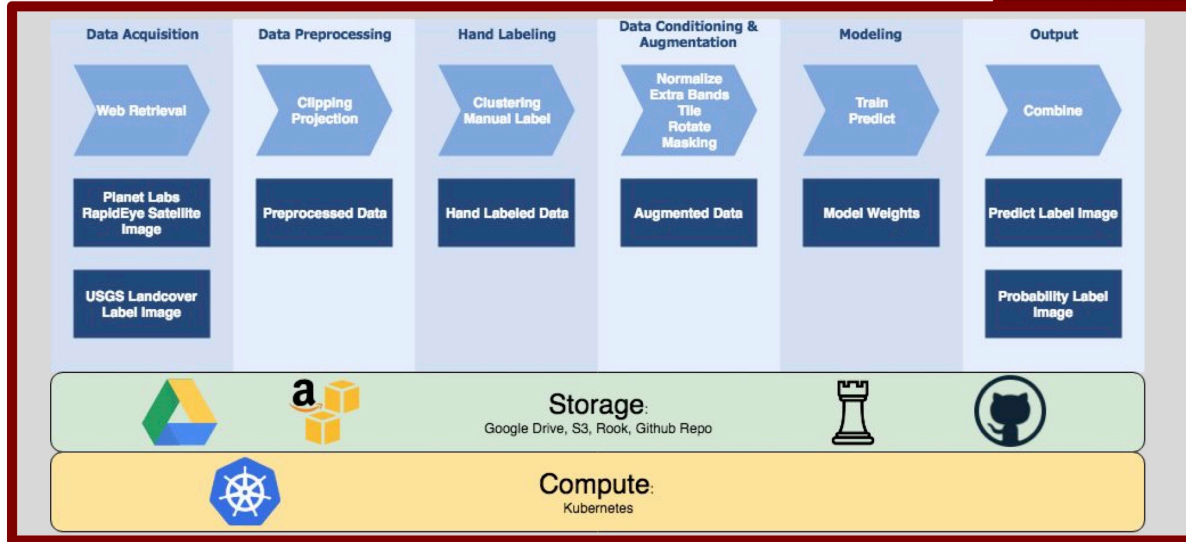
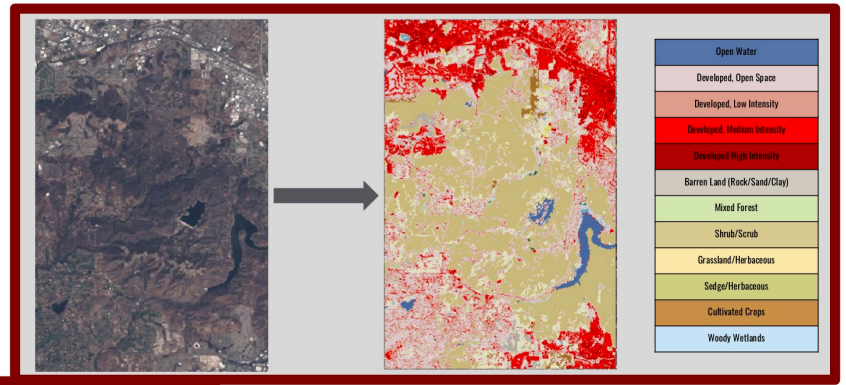
Note: Amber and Gaussian must be loaded on terminal before workflow executes.



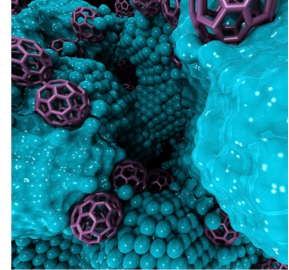
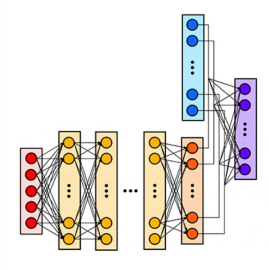
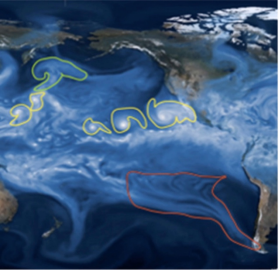
# Fire Modeling Workflows in WIFIRE



# One Piece of the Puzzle: Vegetation Classification using Satellite Imagery



# Scientific Machine Learning & Artificial Intelligence

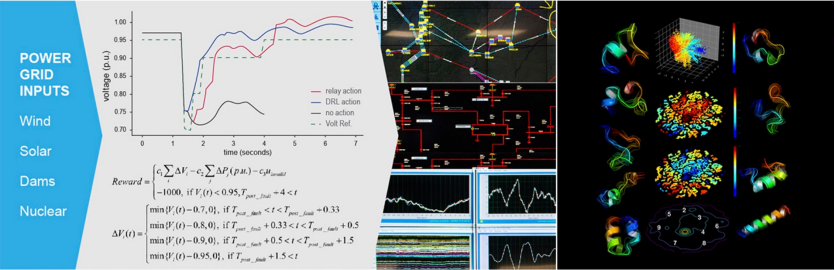


Scientific progress will be driven by

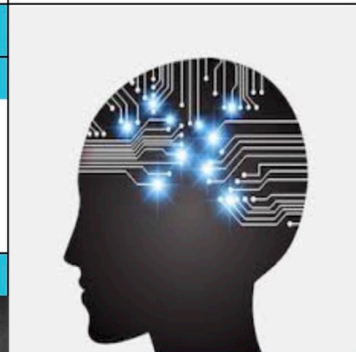
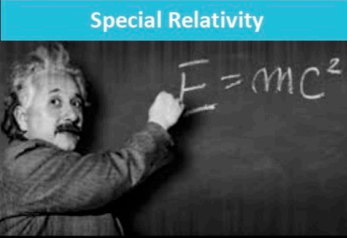
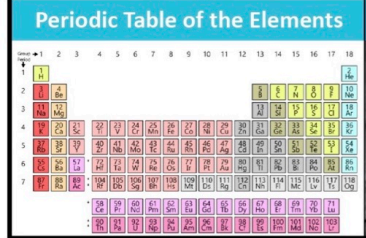
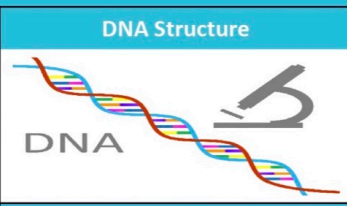
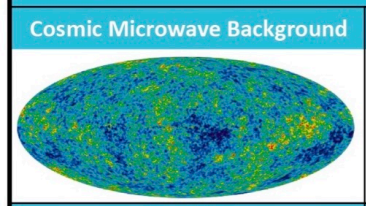
- Massive data: sensors, simulations, networks
- Predictive models and adaptive algorithms
- Heterogeneous high-performance computing

Trend: Human-AI collaborations will transform the way science is done.

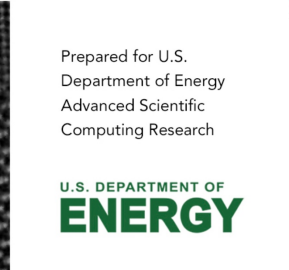
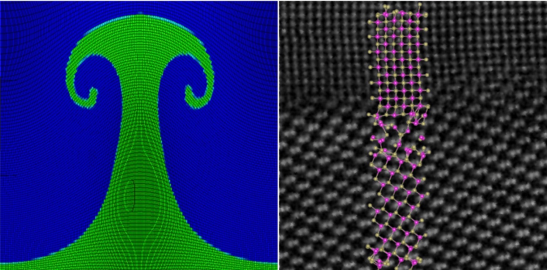
## BASIC RESEARCH NEEDS FOR Scientific Machine Learning Core Technologies for Artificial Intelligence



### EXEMPLARS OF SCIENTIFIC ACHIEVEMENT



Human-AI insights enabled via scientific method, experimentation, & AI reinforcement learning.



Prepared for U.S. Department of Energy Advanced Scientific Computing Research

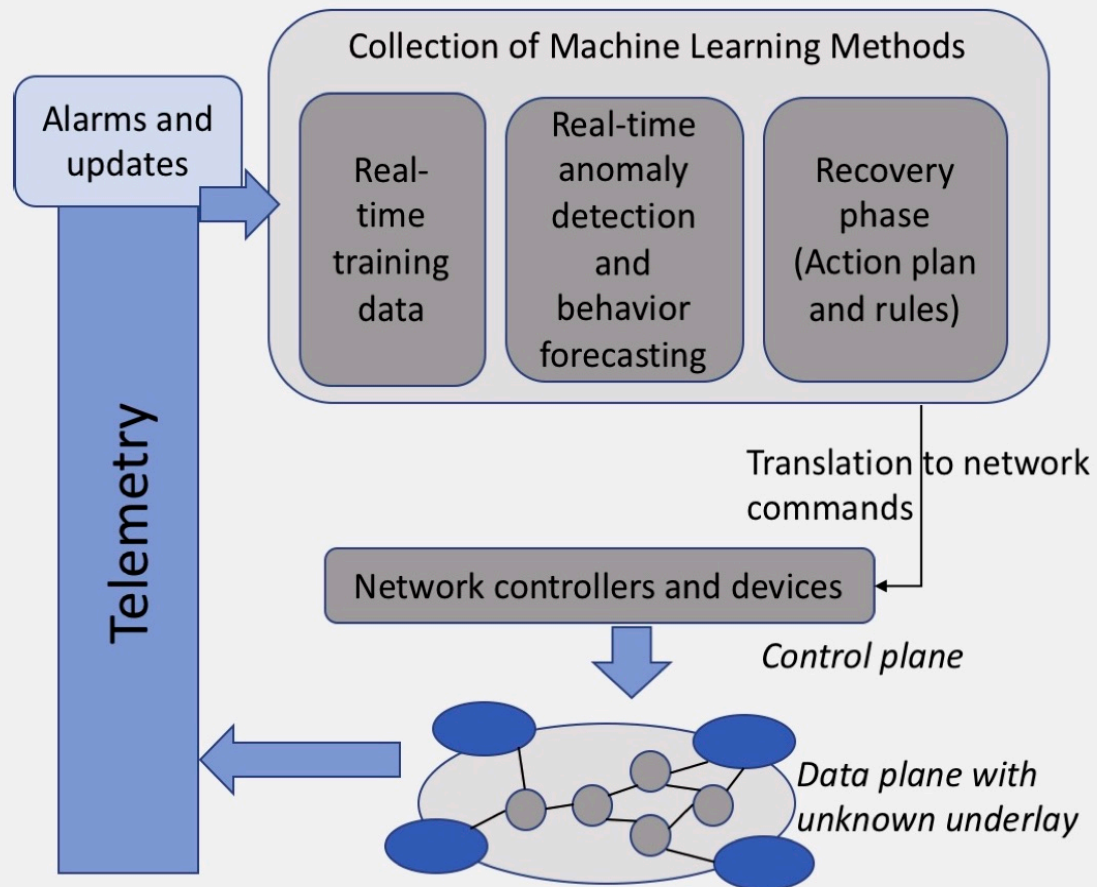


DOE Applied Mathematics Research Program Scientific Machine Learning Workshop (January 2018)

Workshop report:  
<https://www.osti.gov/biblio/1478744>

# Example 1: Optimizing Network Traffic with Machine Learning

Exascale and increasingly complex science applications are exponentially raising demands from underlying DOE networks, such as traffic management, operation scale, and reliability constraints. Networks are the backbone to complex science workflows, ensuring data are delivered securely and on time for important computations to happen. To optimize these distributed workflows, networks are required to understand end-to-end performance needs in advance and be faster, efficient, and more proactive, anticipating bottlenecks before they happen. However, to manage multiple network paths intelligently, various tasks, such as pre-computation and prediction, must be done in near real time. ML provides a collection of algorithms that can add autonomy and assist in decision making to sup-



# Change in computing

- Early days a few big Supercomputers
  - Mostly science domain
- Via grid to commercial cloud
  - AWS, Azure, Google Cloud, IBM, Salesforce
  - The big five: Apple, Alphabet, Microsoft, Facebook and Amazon
  - Computing has transformed into an utility
- Data => Information is the key

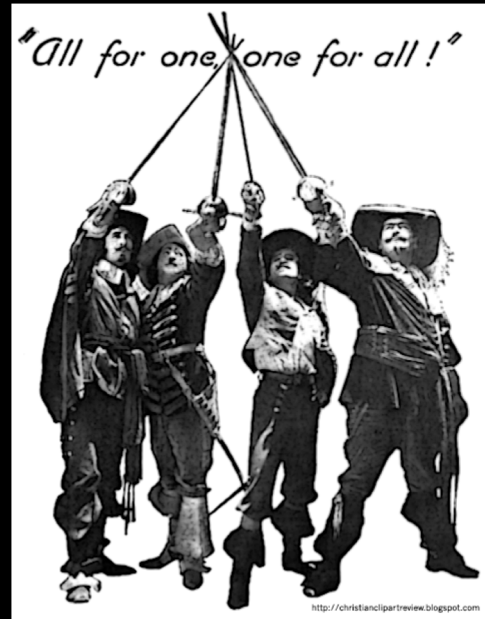


# Main problem statement

- 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?



# 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.

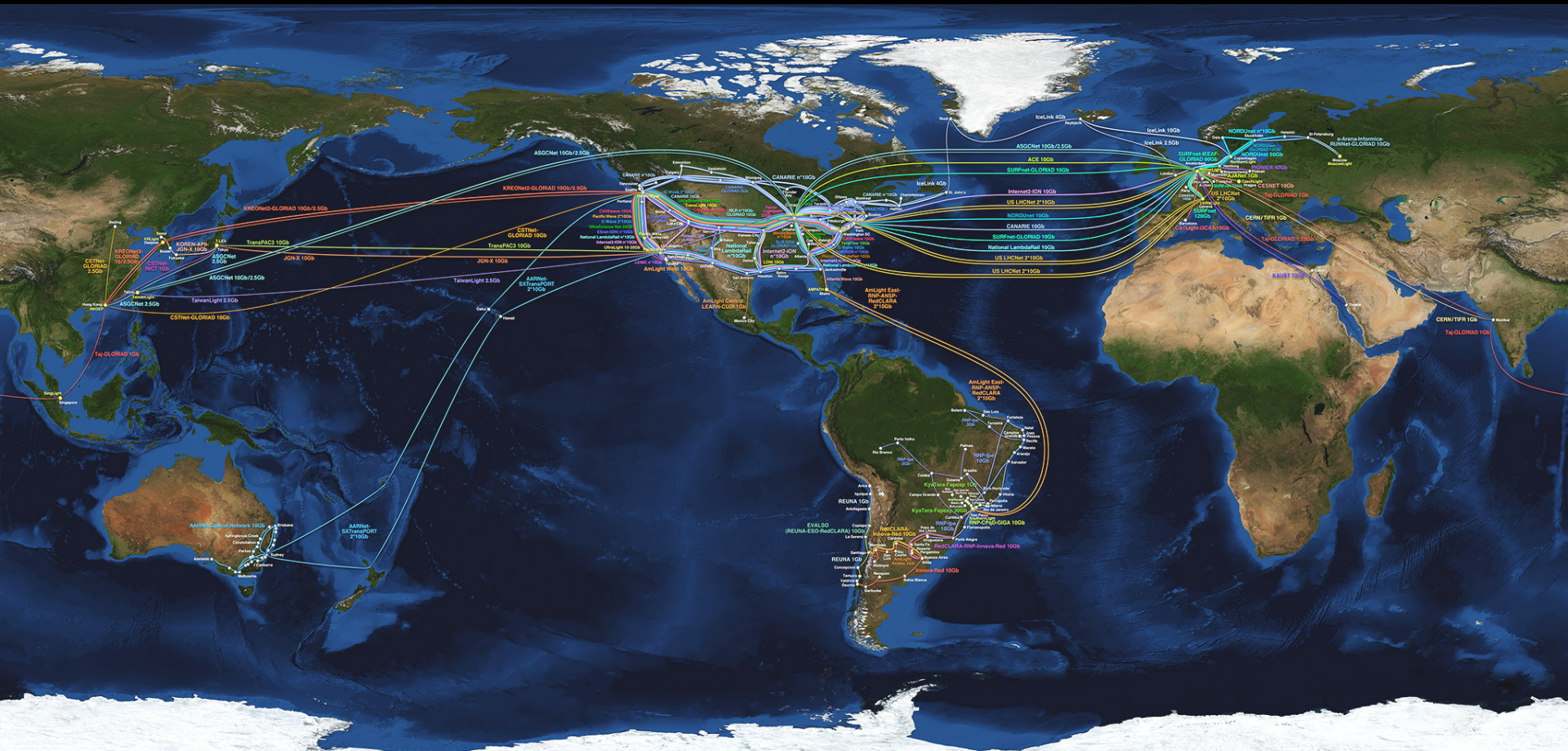
# Now, how do we get and use data?

## 2019 *This Is What Happens In An Internet Minute*



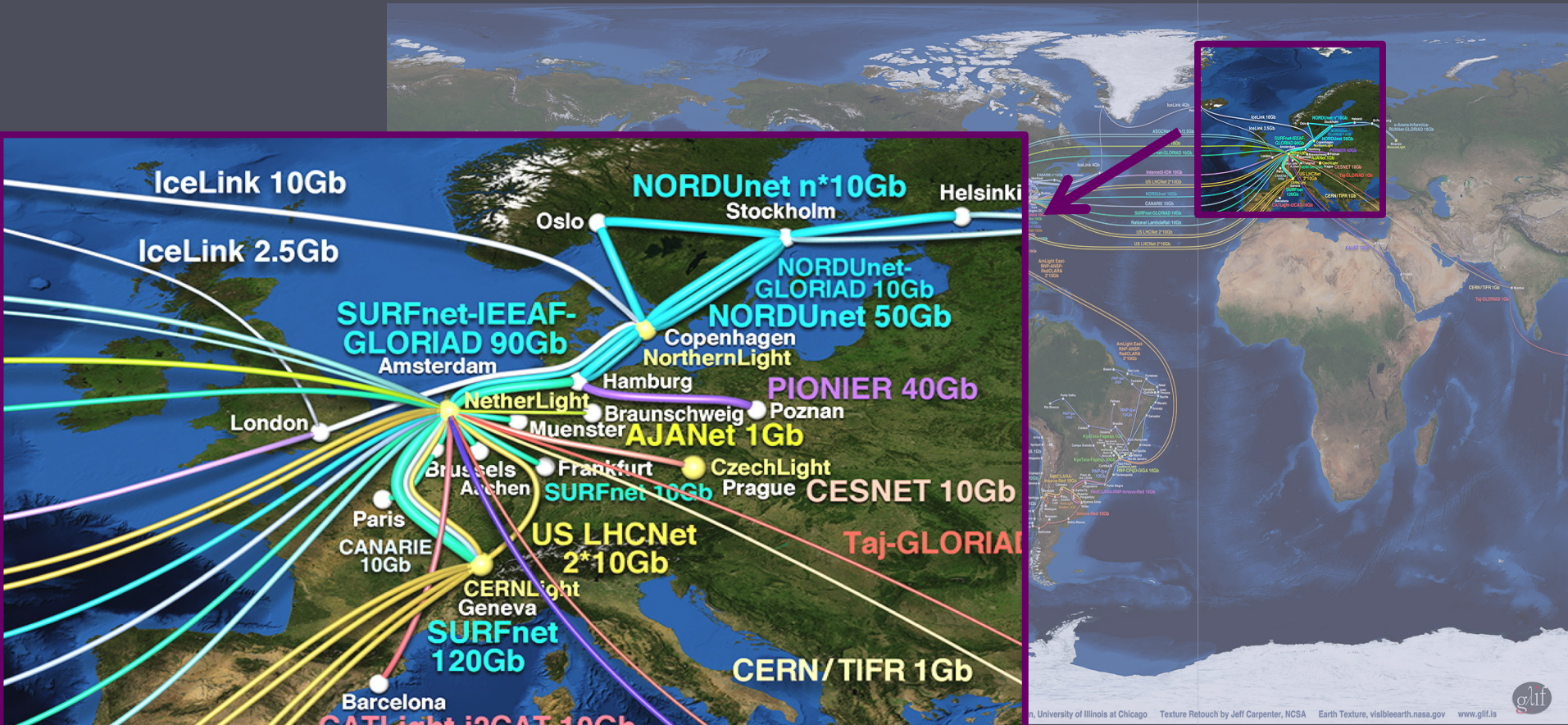
- Move towards streaming
  - Netflix
  - youtube
- Same in science world
  - SKA / LOFAR
  - Light Source
  - Environmental (Marine, Meteorology, ...)
- Data is not always huge
  - Sometimes it is very complex
  - Some example:
    - biodiversity

# The GLIF – LightPaths around the World

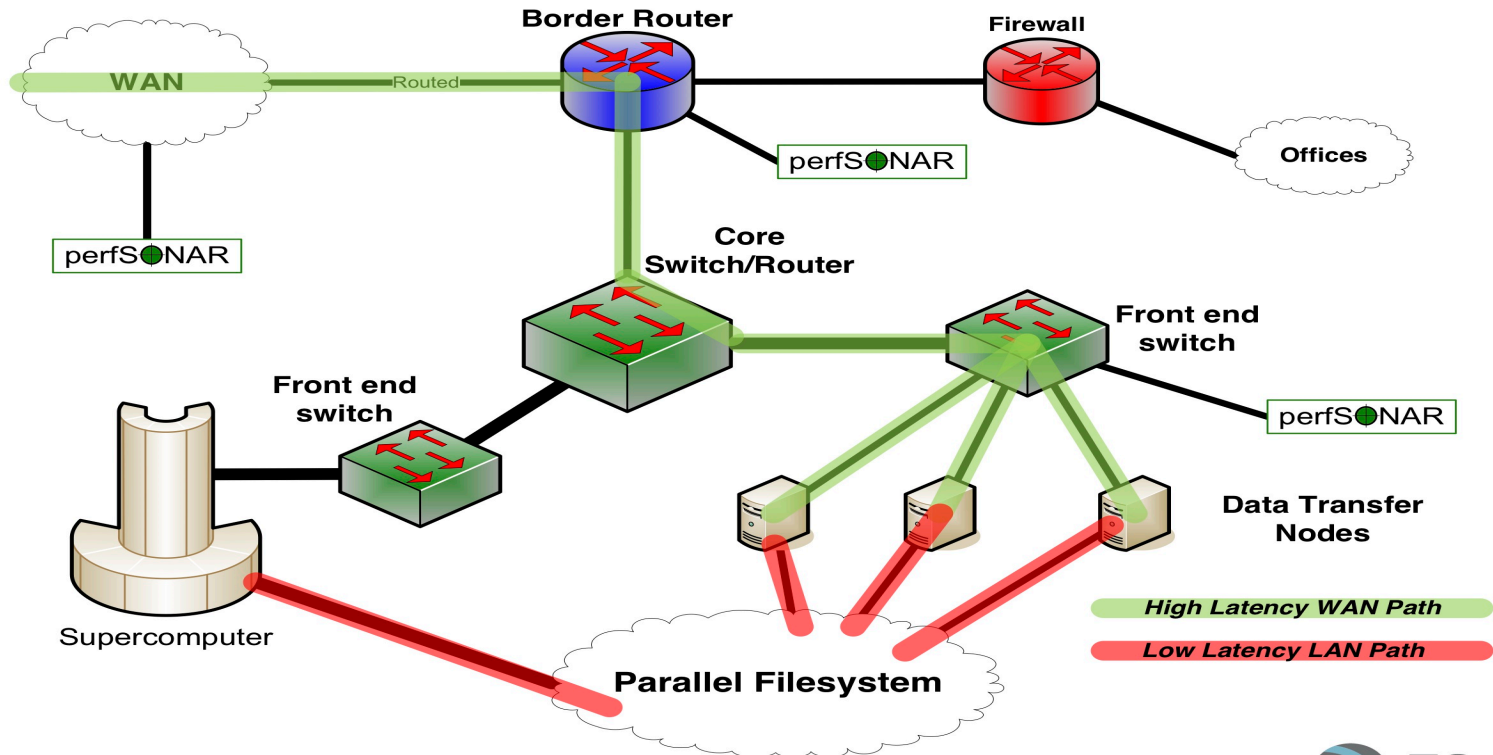


# Amsterdam is a major hub in The GLIF

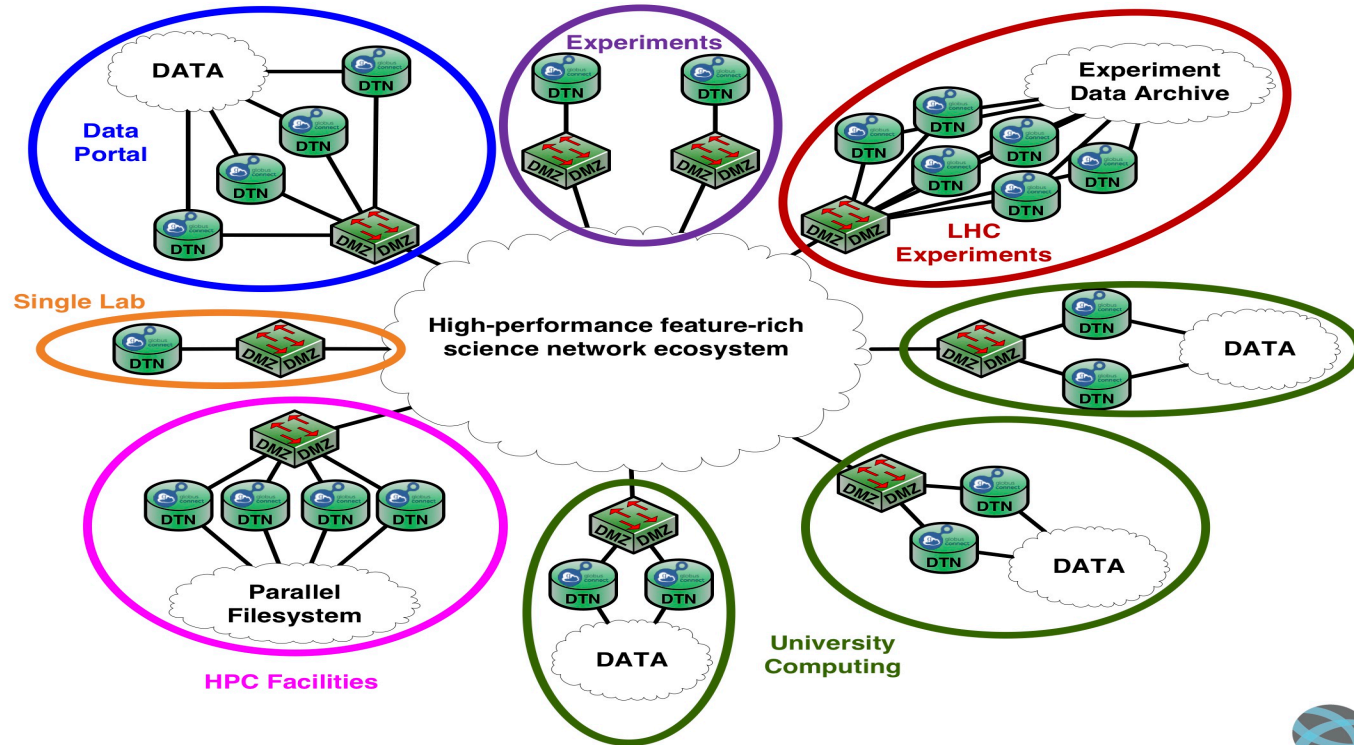
F Dijkstra, J van der Ham, P Grosso, C de Laat, "A path finding implementation for multi-layer networks", Future Generation Computer Systems 25 (2), 142-146.



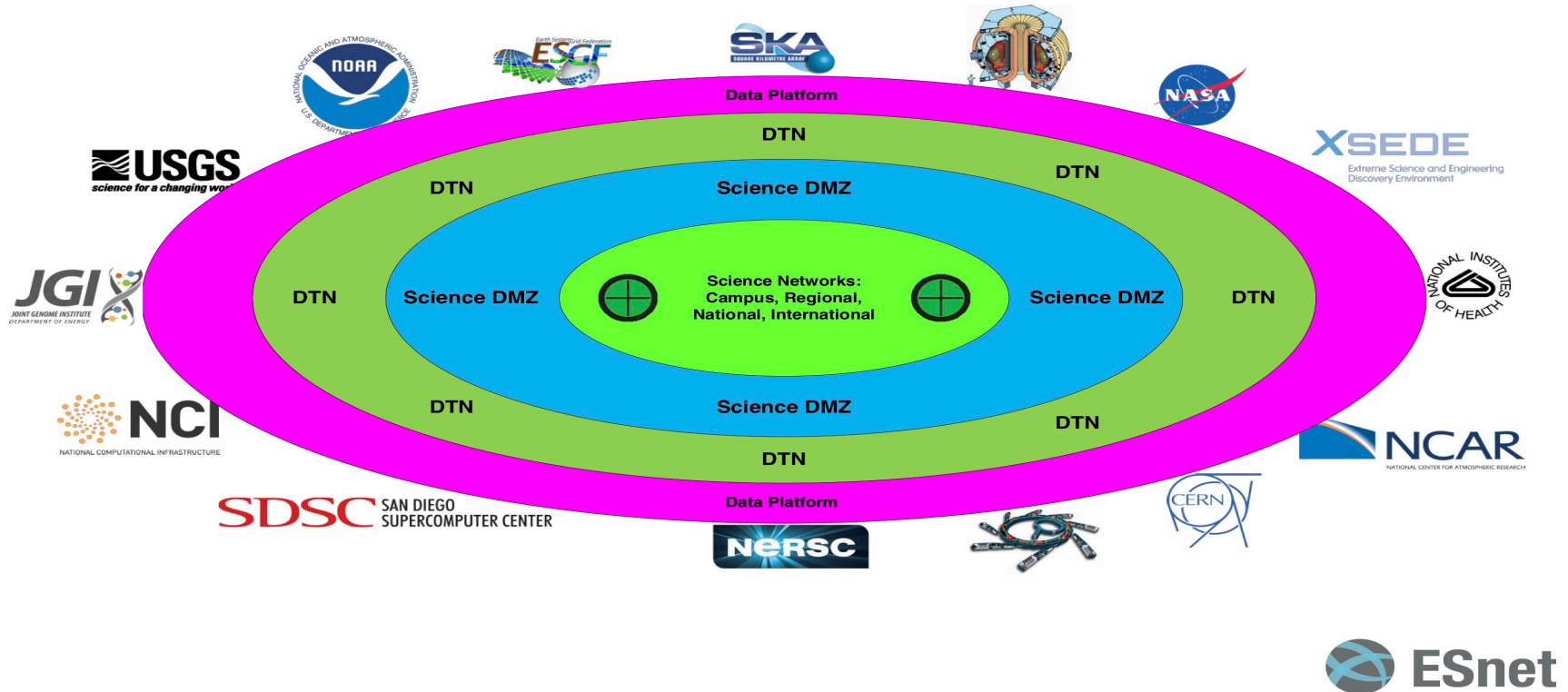
# Science DMZ – HPC Center DTN Cluster



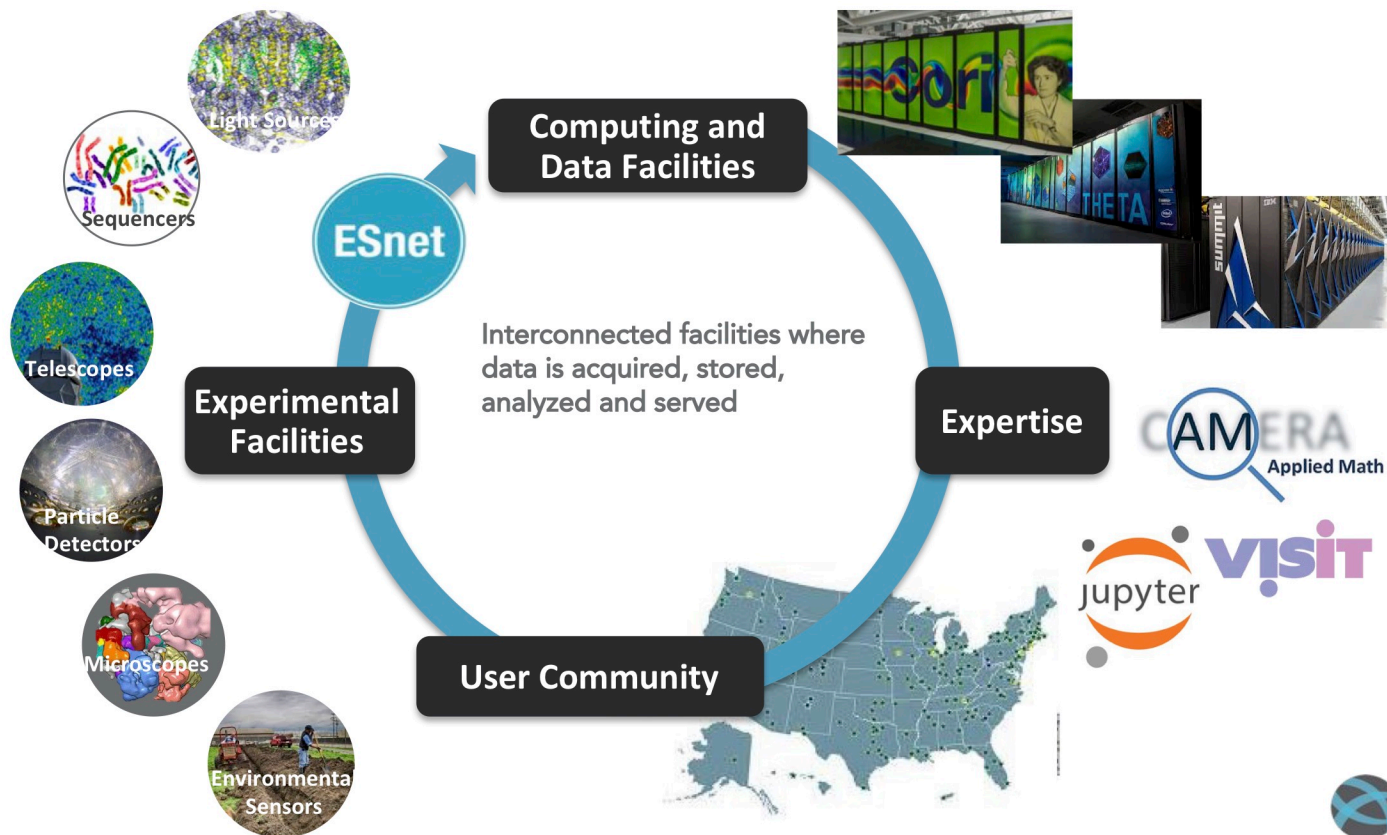
# Science DMZs for Science Applications



# Data Ecosystem – Concentric View

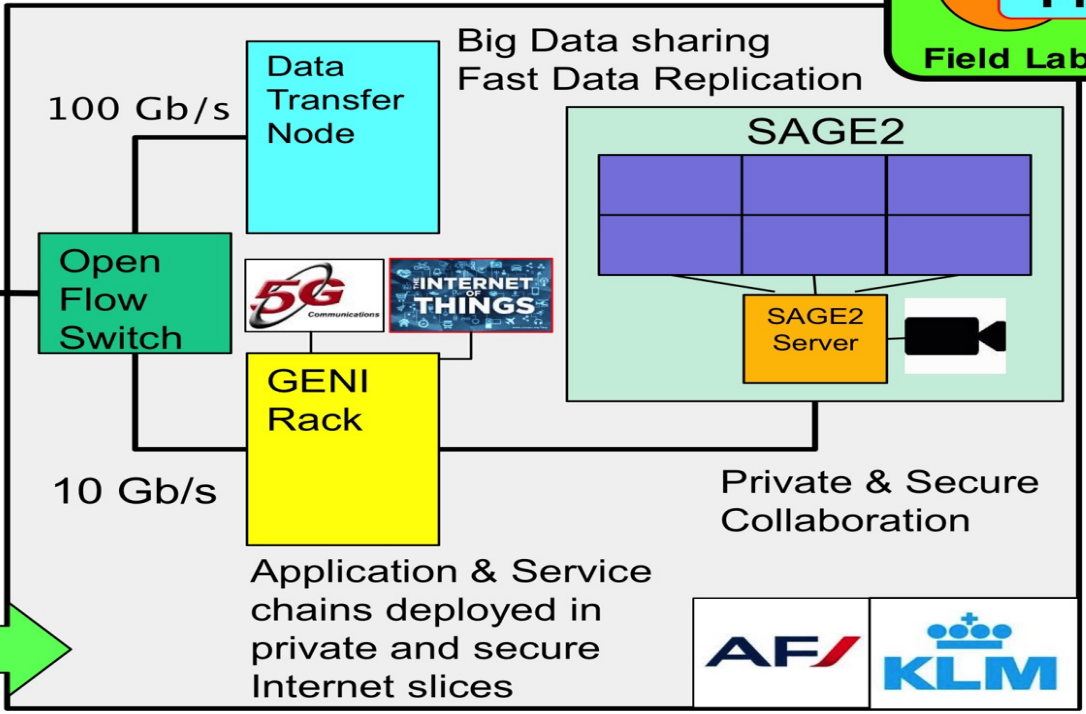
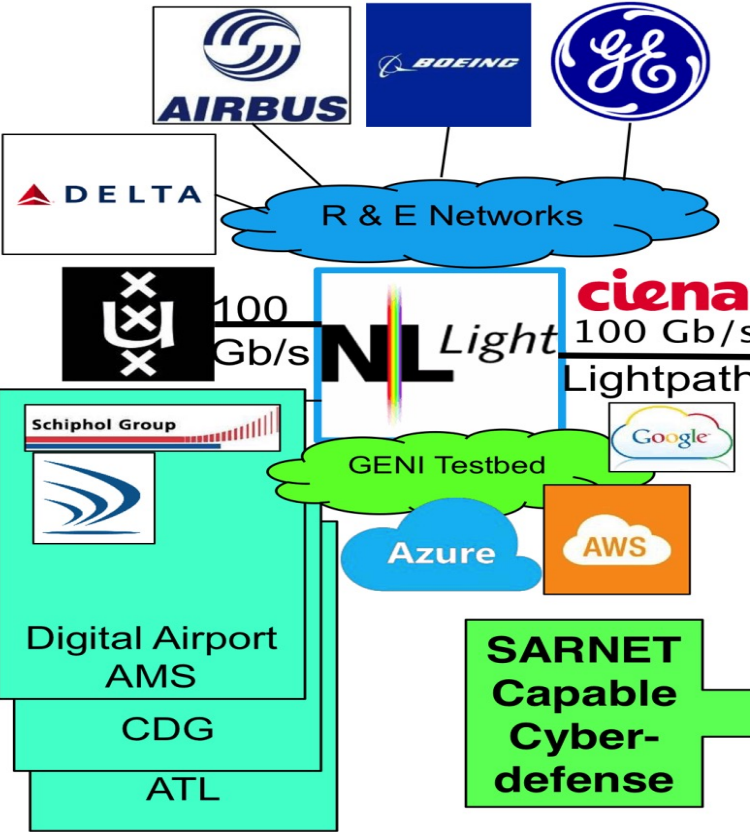


# Superfacility Model for Productive, Reproducible Science





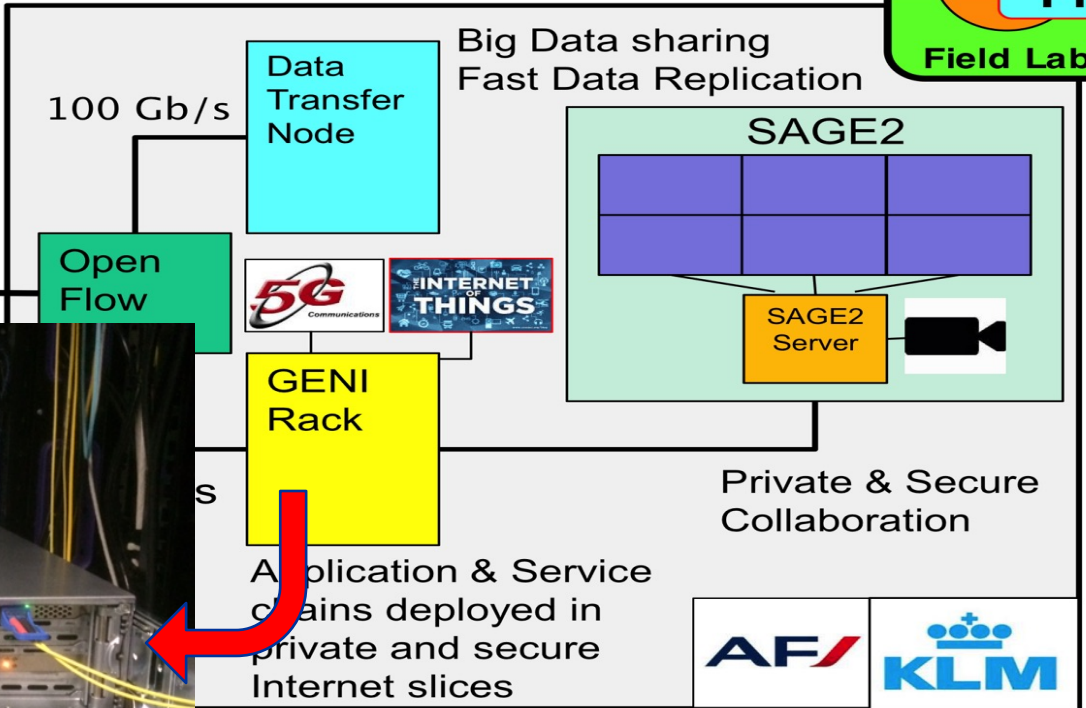
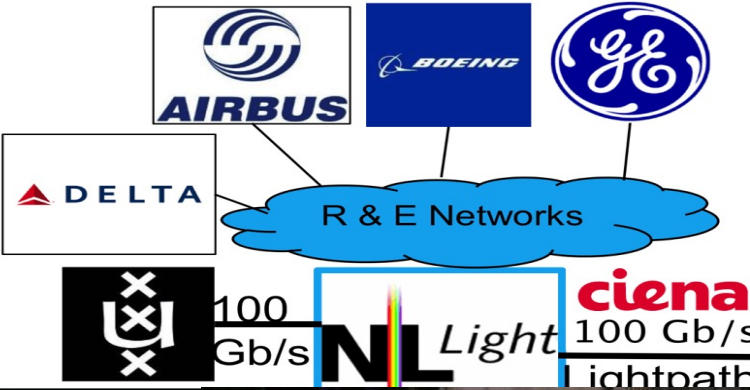
# Ambition to put capabilities into fieldlab



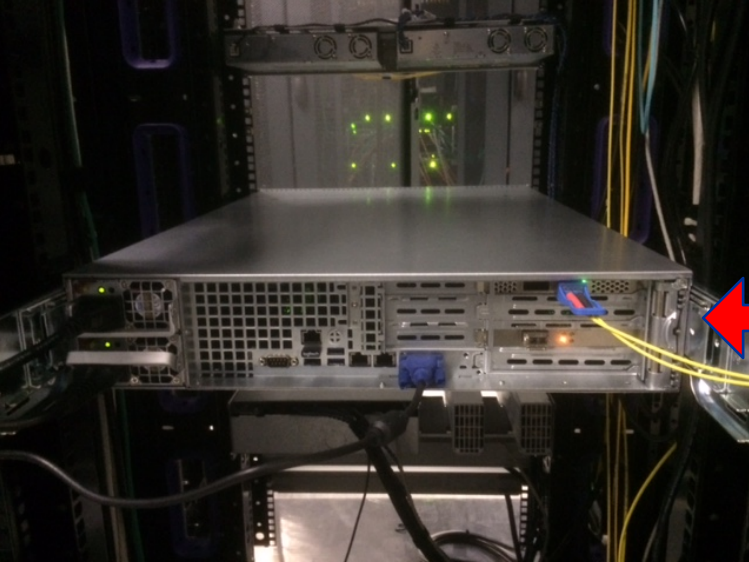
Re-enforcing ICT preconditions:  
Each envisaged site has similar elements



# Ambition to put capabilities into fieldlab



- Schiphol Group
- Digital Air
- AMS
- CDC
- AT

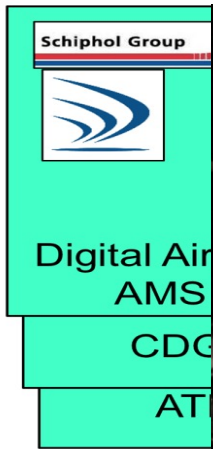
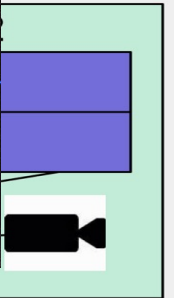
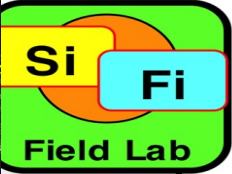


enabling ICT preconditions:  
visaged site has similar elements



# AF/KLM FieldLab

## Ambition to p



GENI Rack

Application & Service chains deployed in private and secure Internet slices

Private & Secure Collaboration



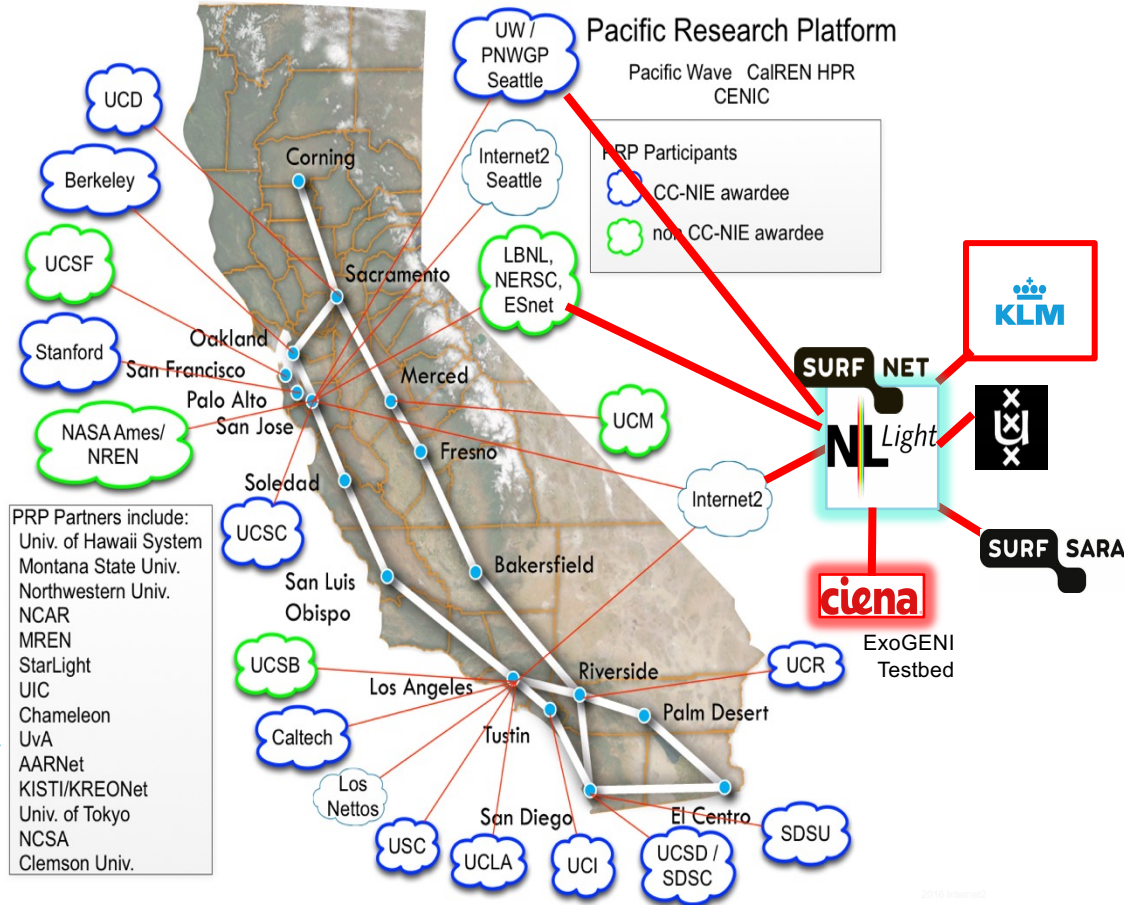
enriching ICT preconditions:  
visaged site has similar elements

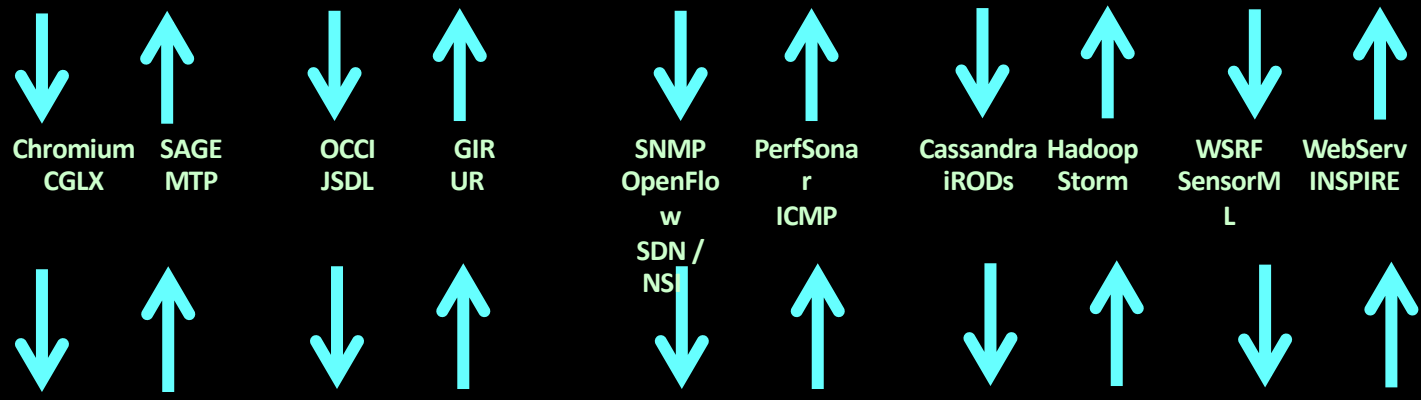
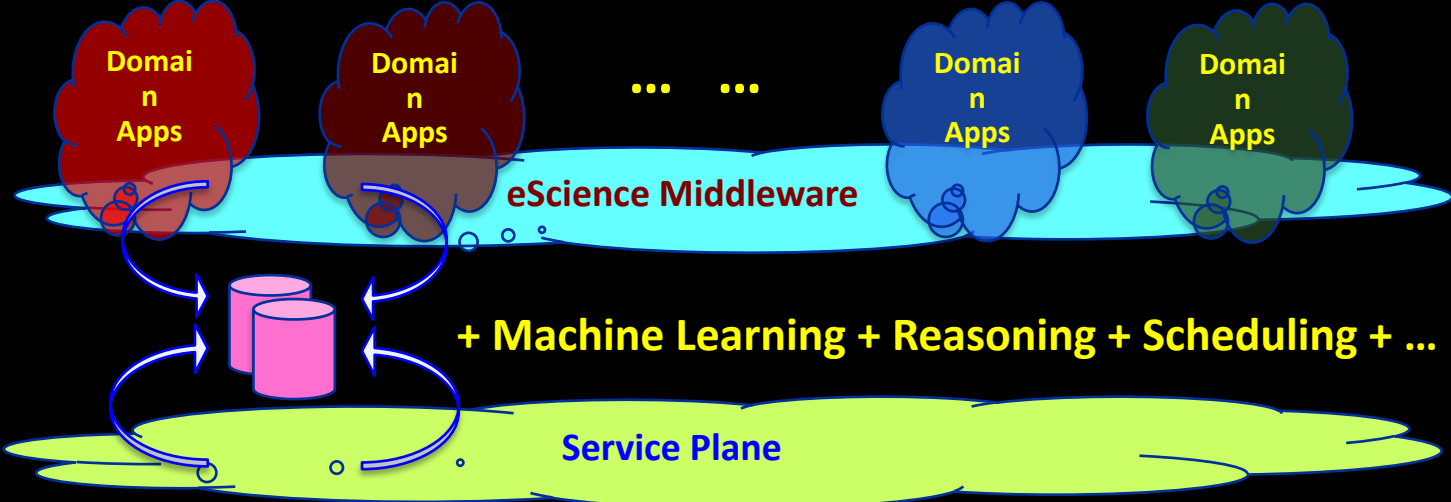
# Pacific Research Platform testbed involvement

**Research goal:**  
Explore value of academic network research capabilities that enable innovative ways & models to share big data assets

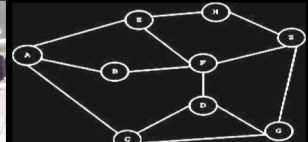


PRP Partners include:  
Univ. of Hawaii System  
Montana State Univ.  
Northwestern Univ.  
NCAR  
MREN  
StarLight  
UIC  
Chameleon  
UvA  
AARNet  
KISTI/KREONet  
Univ. of Tokyo  
NCSA  
Clemson Univ.





GRID/Cloud Computing



# The Big Data Challenge

Doing Science

ICT to enable Science

Wisdom

AI

Knowledge to act

Analytics  
Decision Support

Information

Web/OWL

Data  
a.o. from ESFRI's

Docker, VM,  
XML, RDF, rSpec, SNMP



# The Big Data Challenge

Doing Science

ICT to enable Science

Wisdom

AI

Scientists live here!

Interdisciplinary Science App Store

Analytics library / Github / etc

Knowledge to act

Analytics Decision Support

MAGIC DATA CARPET

curation - description - trust - security - policy - integrity

Information

Web/OWL

Data

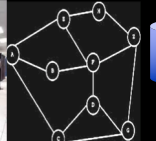
a.o. from ESFRI's

Docker, VM, XML, RDF, rSpec, SNMP

DSC eScience

RDM/DANS

ICT/SURF



# SARNET: Security Autonomous Response with programmable NETWORKS

Marc Lyonais, Leon Gommans, Rodney Wilson, Lydia Meijer, Frank Fransen Tom van Engers, Paola Grosso, Gauravdeep Shami, Cees de Laat, Ameneh Deljoo, Ralph Koning, Ben de Graaff, Gleb Polevoy, Stojan Travanovski.



# Big Data: real time ICT for logistics Data Logistics 4 Logistics Data (dl4ld)

Lydia Meijer (PI), Cees de Laat (Co-PI), Leon Gommans, Tom van Engers, Paola Grosso, Kees Nieuwenhuis.



# EPI: Enabling Personalized Interventions

Cees de Laat(PI), Sander Klous (PL), Leon Gommans, Tom van Engers, Paola Grosso, Henri Bal, Anwar Osseyran, Aki Harma, Douwe Biesma, Peter Grünwald, Floortje Scheepers, Gertjan Kaspers.





# SARNET: Security Autonomous Response with programmable NETWORKS

Marc Lyonais, Leon Gommans, Rodney Wilson, Lydia Meijer, Frank Fransen Tom van Engers, Paola Grosso, Gauravdeep Shami, Cees de Laat, Ameneh Deljoo, Ralph Koning, Ben de Graaff, Gleb Polevoy, Stojan Travanovski.



## Big Data: real time ICT for logistics Data Logistics 4 Logistics Data (dl4ld)

Lydia Meijer (PI), Cees de Laat (Co-PI), Leon Gommans, Tom van Engers, Paola Grosso, Kees Nieuwenhuis.



## EPI: Enabling Personalized Interventions

Cees de Laat (PI), Sander Klous (PL), Leon Gommans, Tom van Engers, Paola Grosso, Henri Bal, Anwar Osseyran, Aki Harma, Douwe Biesma, Peter Grünwald, Floortje Scheepers, Gertjan Kaspers.

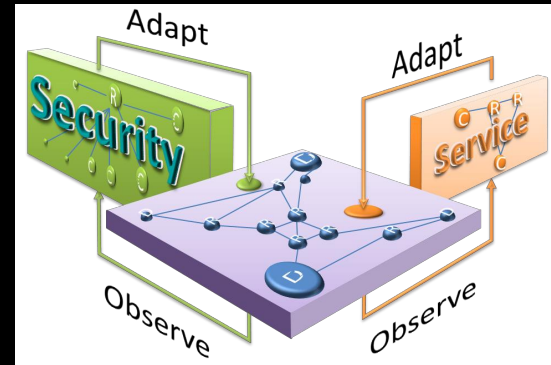


# Cyber security program SARNET

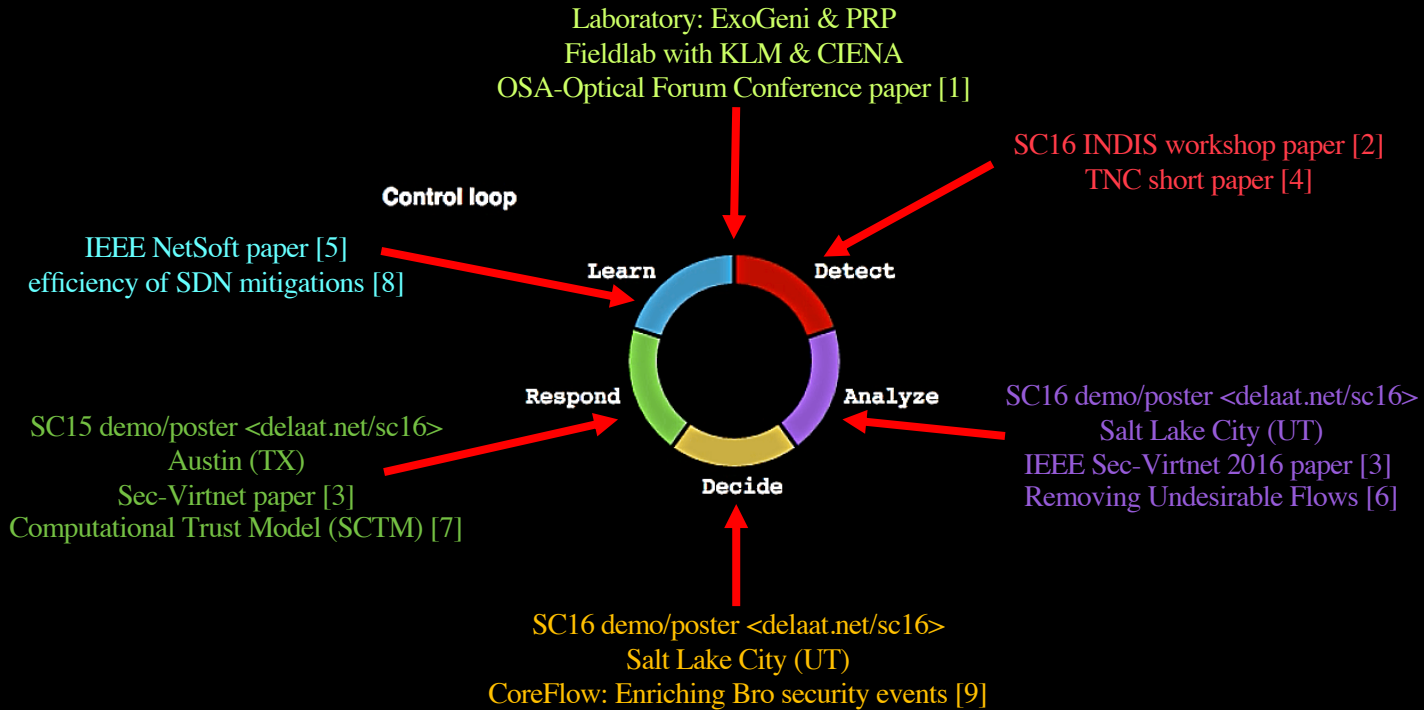
Research goal is to obtain the knowledge to create ICT systems that:

- model their state (situation)
- discover by observations and reasoning if and how an attack is developing and calculate the associated risks
- have the knowledge to calculate the effect of counter measures on states and their risks
- choose and execute one.

In short, we research the concept of networked computer infrastructures exhibiting SAR: Security Autonomous Response.



# SARNET Publications (subset)



1. Paper: R. Koning, A. Deljoo, S. Trajanovski, B. de Graaff, P. Grosso, L. Gommans, T. van Engers, F. Fransen, R. Meijer, R. Wilson, and C. de Laat, "Enabling E-Science Applications with Dynamic Optical Networks: Secure Autonomous Response Networks", OSA Optical Fiber Communication Conference and Exposition, 19-23 March 2017, Los Angeles, California.
2. Paper: Ralph Koning, Nick Buraglio, Cees de Laat, Paola Grosso, "CoreFlow: Enriching Bro security events using network traffic monitoring data.", Special section on high-performance networking for distributed data-intensive science, SC16", Future Generation Computer Systems, <accepted for publication>
3. Paper: Ralph Koning, Ben de Graaff, Cees de Laat, Robert Meijer, Paola Grosso, "Analysis of Software Defined Networking defenses against Distributed Denial of Service attacks", The IEEE International Workshop on Security in Virtualized Networks (Sec-VirtNet 2016) at the 2nd IEEE International Conference on Network Softwarization (NetSoft 2016), Seoul Korea, June 10, 2016.
4. Short paper: Nick Buraglio, Ralph Koning, Cees de Laat, Paola Grosso, "Enriching network and security events for event detection", Conference proceedings TNC2017, <https://tnc17.naant.org/core/presentation/30>
5. Paper: Ralph Koning, Ben de Graaff, Robert Meijer, Cees de Laat, Paola Grosso, "Measuring the effectiveness of SDN mitigations against cyber attacks", IEEE Conference on Network Softwarization (NetSoft 2017 - SNS 2017), Bologna, Italy, July 3-7, 2017.
6. Paper: Gleb Polevoy, Stojan Trajanovski, Paola Grosso and Cees de Laat, "Removing Undesirable Flows by Edge Deletion.", COCOA'2018 conference, December 15 - 17, 2018, Atlanta, Georgia, USA, Springer-Verlag.
7. Paper: Ameneh Deljoo, Tom van Engers, Leon Gommans, Cees de Laat, "Social Computational Trust Model (SCTM): A Framework to Facilitate Selection of Partners". In: Proceedings of 2018 IEEE/ACM Innovating the Network for Data-Intensive Science (INDIS), Dallas, TX, USA, 2018
8. Paper: R. Koning, B. de Graaff, G. Polevoy, R. Meijer, C. de Laat, P. Grosso, "Measuring the efficiency of SDN mitigations against attacks on computer infrastructures", Future Generation Computer Systems 91, 144-156.
9. Ralph Koning, Nick Buraglio, Cees de Laat, Paola Grosso, "CoreFlow: Enriching Bro security events using network traffic monitoring data.", Special section on high-performance networking for distributed data-intensive science, SC16", Future Generation Computer Systems

# SARNET: Security Autonomous Response with programmable NETWORKS

Marc Lyonnais, Leon Gommans, Rodney Wilson, Lydia Meijer, Frank Fransen Tom van Engers, Paola Grosso, Gauravdeep Shami, Cees de Laat, Ameneh Deljoo, Ralph Koning, Ben de Graaff, Gleb Polevoy, Stojan Travanovski.



## Big Data: real time ICT for logistics Data Logistics 4 Logistics Data (dl4ld)

Lydia Meijer (PI), Cees de Laat (Co-PI), Leon Gommans, Tom van Engers, Paola Grosso, Kees Nieuwenhuis.



## EPI: Enabling Personalized Interventions

Cees de Laat(PI), Sander Klous (PL), Leon Gommans, Tom van Engers, Paola Grosso, Henri Bal, Anwar Osseyran, Aki Harma, Douwe Biesma, Peter Grünwald, Floortje Scheepers, Gertjan Kaspers.



# Harvard Business Review



Harvard Business Review

ECONOMY

## Managing Our Hub Economy


by **Marco Iansiti** and **Karim R. Lakhani**

FROM THE SEPTEMBER–OCTOBER 2017 ISSUE

WHAT TO READ NEXT

The IT Transformation Health Care Needs

SUMMARY SAVE SHARE COMMENT H TEXT SIZE PRINT \$8.95 BUY COPIES



THOMAS M. SCHEER/EVEEM/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.

Data value creation  
monopolies



Create an equal  
playing field



Sound Market  
principles

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

# Data Sharing: 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 or Algorithms may have to be processed in foreign data centers.
  - How to organize alliances?
  - How to translate from strategic via tactical to operational level?
  - How to enforce policy using modern Cyber Infrastructure?
  - What are the different fundamental data infrastructure models to consider?

# Big Data Sharing use cases placed in airline context



Global Scale



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

National Scale



Cargo Logistics Data  
(C1) DaL4LoD  
(C2) Secure scalable  
policy-enforced  
distributed data  
Processing  
(using blockchain)



Cybersecurity Big Data  
NWO **COMMIT/**  
**SARNET** project  
3.5 FTE

Campus /  
Enterprise Scale

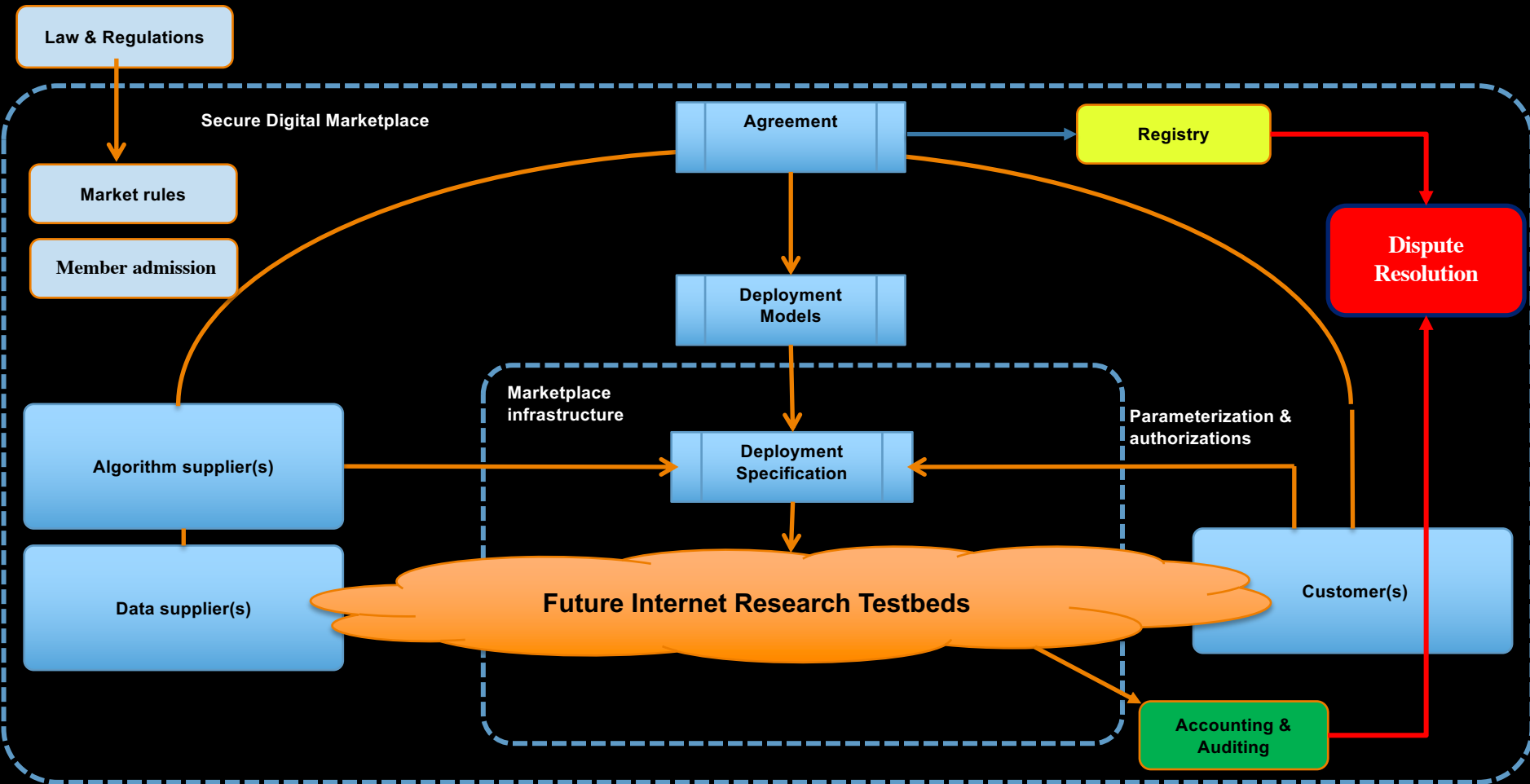
NLIP **iShare** project



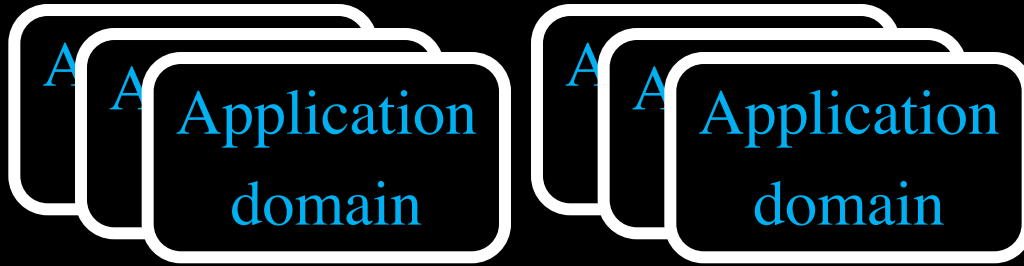
**iSHARE**  
powered by NLIP



# Secure Digital Market Place Research







AmDex

Data objects & methods  
Data & Algorithms service

FAIR / USE

AmsIX

Routers - Internet – ISP's - Cloud  
IP packet service

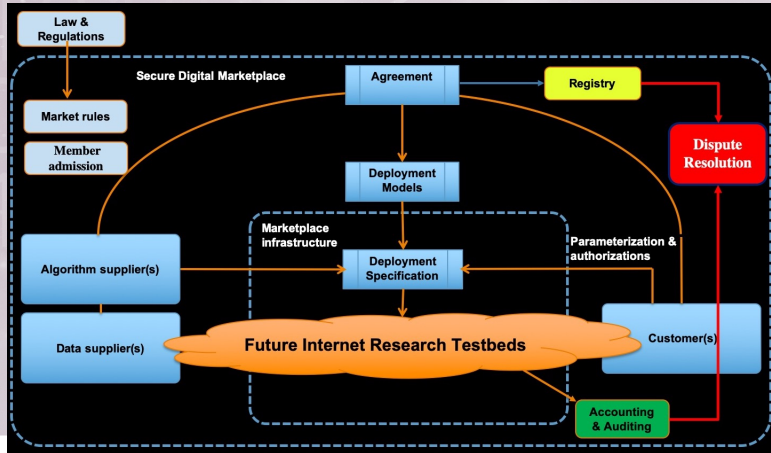
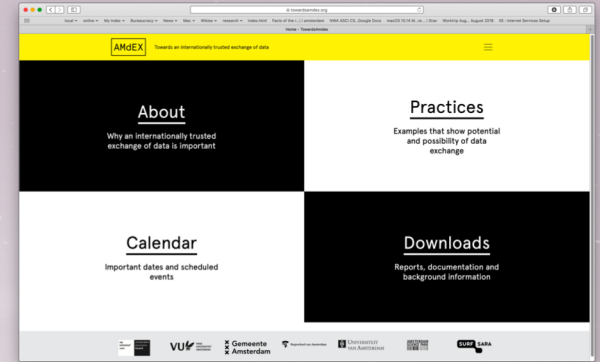
IP / BGP

Layer 2 exchange service  
Ethernet frames

ETH / ST

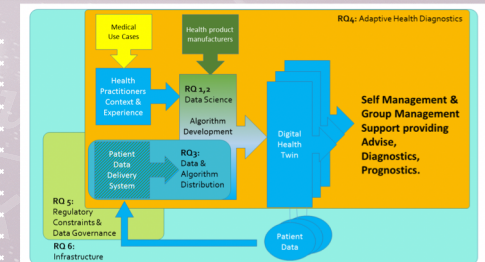
## AMdEX.eu

- Competing organisations, share data for common benefit
- Trust, Risk, data ownership & control
  - Industry: AF-KLM, Health, etc
  - Science: European Open Science Cloud
  - Society: Amsterdam Economic Board



Aircraft Maintenance AF-KLM

Health:  
Enabling  
Personal  
Interventions



# SARNET: Security Autonomous Response with programmable NETWORKS

Marc Lyonnais, Leon Gommans, Rodney Wilson, Lydia Meijer, Frank Fransen Tom van Engers, Paola Grosso, Gauravdeep Shami, Cees de Laat, Ameneh Deljoo, Ralph Koning, Ben de Graaff, Gleb Polevoy, Stojan Travanovski.



## Big Data: real time ICT for logistics Data Logistics 4 Logistics Data (dl4ld)

Lydia Meijer (PI), Cees de Laat (Co-PI), Leon Gommans, Tom van Engers, Paola Grosso, Kees Nieuwenhuis.

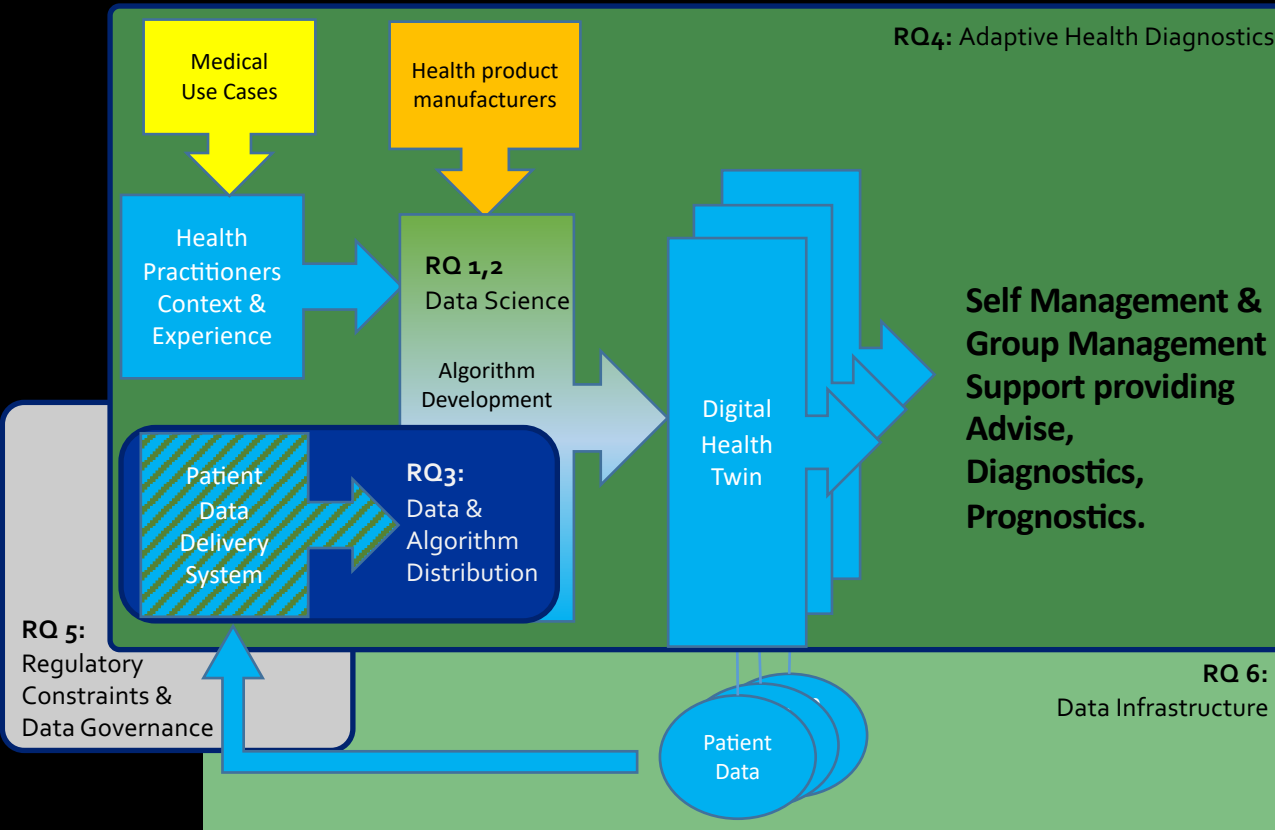


## EPI: Enabling Personalized Interventions

Cees de Laat(PI), Sander Klous (PL), Leon Gommans, Tom van Engers, Paola Grosso, Henri Bal, Anwar Osseyran, Aki Harma, Douwe Biesma, Peter Grünwald, Floortje Scheepers, Gertjan Kaspers.



# 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.

# Research questions

- RQ1: **Dynamically Analyzing Interventions** based on Small Groups: how can we determine, based on as little data as possible, whether an intervention does or does not work for a small group or even an individual patient?
- RQ2: **Dynamically Personalizing the Group**: how can we identify effective intervention strategies and optimize personalization strategies applicable for different patient and lifestyle profiles via dynamic (on-line) clustering of patients? Can those clusters be adapted as new data about patients and results of interventions come in and as other data may be removed or modified?
- RQ3: **Data and Algorithm Distribution**: what are the consequences of a distributed, multi-platform, multi-domain, multi-data-source big data infrastructure on the machine learning algorithms and what are potential consequences on performance?
- RQ4: Adaptive health diagnosis leading to optimized intervention: how can we **enhance self- / joint management** by dynamically integrating updated models generated from machine learning from various data sources in state of the art health support systems that based on personal health records, knowledge of health modes and effective interventions?
- RQ5: **Regulatory constraints** and data governance: how can we create scalable solutions that meet legal requirements and consent or medical necessity-based access to data for allowed data processing and preventing breaches of these rules by embedded compliance, providing evidence trails and transparency, thus building trust in a sensitive big data sharing infrastructure?
- RQ6: **Infrastructure**: how can the various requirements from the use-cases be implemented using a single functional ICT-infrastructure architecture?

# SC16 Demo

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

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

UNIVERSITEIT VAN AMSTERDAM **Container-based remote data processing** Łukasz Makowski, Daniel Romão, Cees de Laat, Paola Grosso  
System and Networking Research Group, University of Amsterdam

**S&E**  
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?

**Problem**: Data leaving the organization (marked with a red X).

**Approach**: Send container to the organization (marked with a green checkmark).

### 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?

**Bring Your Own Container**

Score: 0/1000 Budget: 1000

Search functions: BIO, EAT, SLEEP, WALK

Containers: 1000

**More information:**

- <http://byoc.lab.uvlight.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>

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

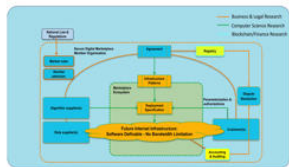
### Digital Marketplace as a means to organize trusted data asset sharing

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

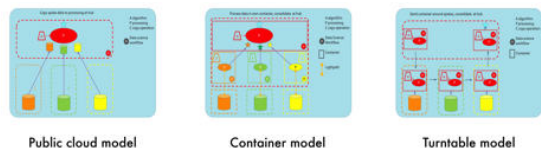
Membership organization is institutionalized to create, implement and enforce membership rules.

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

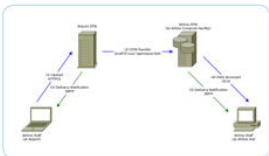


## DEMONSTRATION: LIGHT PATHS AND DATA TRANSFER NODES FOR AIRCRAFT MAINTENANCE

Air France-KLM uses a 100 Gbit/s link, connected to Netherlight, to research an aircraft maintenance industry use case. Via this open exchange, Data Transfer Nodes (DTNs) of Air France-KLM in the Netherlands and CAIR – present in Chicago at 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 DTNs.

### USE CASE: AIRCRAFT MAINTENANCE

Receivers analyze and log engine aircraft 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 along their job safety. When data are transferred and analyzed rapidly, defects can be caught more quickly, thereby avoid the aircraft's 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 warning was caused by the engine or by a falling sensor.



### INTERNET VS LIGHT PATH

Air France-KLM uses a 100 Gbit/s light path and measures its benefits. Using light paths, you can transport huge amounts of data at high speed and with a guaranteed bandwidth between 2 points. When using high volumes of data, the current Air France-KLM internet connections are not capable or fast enough to transfer the data within the required time frame. Transferring a terabyte of engine data via the current internet connection would take around 30 hours, with a 100 Gbit/s light path this could take less than 2 minutes.

### AIR FRANCE-KLM CONNECTED TO NETHERLIGHT

Cair and SURF facilitate the connection from Air France-KLM at Schiphol to Netherlight, SURF's European hub for international light paths in Amsterdam. SURF provides the 100 Gbit/s light path from Air France-KLM via Netherlight to the aircraft's destination. For this demonstration, the location is Starlight in Chicago, a full connect to Netherlight.

### DATA TRANSFER NODES

Data Transfer Nodes are high-performance systems that are optimized to transfer huge amounts of data. The interconnects between these systems are 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.

### LEIVE DEMONSTRATION

In this demonstration, there are two end user systems, one in Amsterdam and one in Chicago. Nether system will be optimized for long range transfers, however each will have access to a nearby Data Transfer Node. Nodes are allowed to transfer pre-measured data 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 the transfer time.

### RESEARCH IN OTHER INDUSTRIES

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

More information: [www.auftrijvler/100-G-Air-France-KLM](http://www.auftrijvler/100-G-Air-France-KLM)



## Data Transfer Node (DTN) Workflows

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

### 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



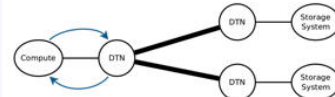
### Example: Entry Point for High Speed Transport

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.

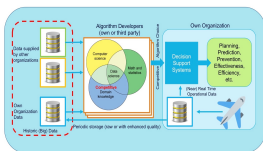


# SC18 – Dallas TX

## 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 decision making 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.

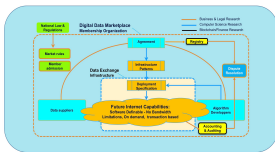
## Digital Data Marketplace enabling data sharing and competition

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.

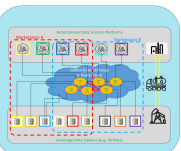
Membership organization is institutionalized to create, implement and enforce membership rules organizing trust.

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

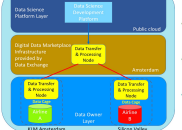
Agreements subsequently drive data science transactions creating processing infrastructures using infrastructure patterns offered by a Data Exchange Platform.



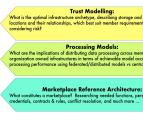
## Researching Exchange Patterns to support Digital Data Marketplaces



Data Exchange Model



Research Infrastructure



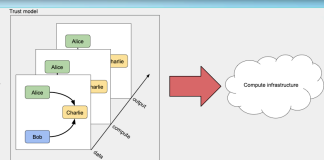
Research Elements

## Dataharbours: computing archetypes for digital marketplaces

Reginald Cushing, Lu Zhang, Paola Grosso, Tim van Zalingen, Joseph Hill, Leon Gommans, Cees de Laat, Vijaya 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

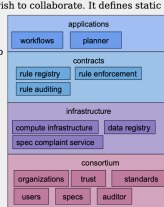
**Consortium:** is an initial document which brings together organizations that wish to collaborate. It defines static information such as keys to identify parties.

**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.

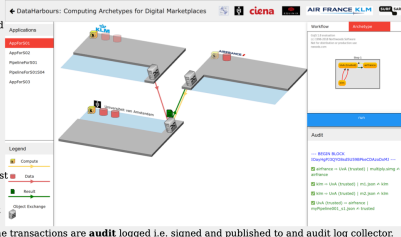


### In action

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/private 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.

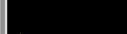


**Exhibitors:** AIR FRANCE KLM, TRANSFIDES, BIZZdesign, THALES, SCinet, ORACLE, Equinix, Gemalto, Cisco, SAP, IBM, Microsoft, Amazon, Google, Facebook, LinkedIn, Twitter, YouTube, Instagram, etc.

**Content Sections:**

- Building User-friendly Data Transfer Nodes:** Focuses on enhancing data transfer efficiency and user experience.
- Building Information Systems and Applications for the critical applications:** Discusses the integration of data into critical business systems.
- Building User-friendly Data Transfer Nodes:** Details the architecture and implementation of data transfer nodes.

**Footer:** Includes logos for various partners and sponsors, along with contact information for the booth.





# SC19 – Denver CO

<https://sc.delaat.net>



SC2019 Workshop and Technical program presentations & involvement by UvA group members.

- Innovating the Network for Data Intensive Science (INDIS) workshop organized by Paola Grosso, Michallo Zhu and Ilya Baldin.
  - Sunday Nov 17, 9:00 - 17:30, room 501
    - SCinet Architecture and Technical Challenge introduction", by JP Velders and Cees de Laast as part of the Architecture, Technical Challenge, XNET and NRE talk in INDIS
    - "Hop recording and forwarding state logging...Two implementations for path tracking in P4", Joseph Hill (presenter), Silke Krossen and Paola Grosso
- Technology Challenge, Jury chaired by Cees de Laast. The goal of the challenge is to demonstrate that both networking and high-performance computing resources are essential elements of the cyber infrastructure required to advance modern data-driven scientific applications. Join us for this exciting new program.
  - Tuesday, November 19, 10:30-12:00 and 13:15-13:45
    - Technology Challenge Overview and Jury Introduction
    - 5G Cellular Broadband Radio Service (CBRS) Proof-of-Concept for Scientific Applications - University of Utah Center for High Performance Computing, Murray School District and the Utah Education and Telehealth Network
    - Dynamic Network-Centric Multi-Cloud Platform (DyNano) for Real-Time Weather Forecasting Workflows - RENGUONG CHAPPEL Hill, USC/ISI, UMass Amherst and Rutgers University
    - Real-Time Analysis of Streaming Symptom Data - Argonne National Laboratory, Northwestern University, Starlight, Northern Illinois University, University of Chicago
  - Thursday, November 21, 10:30
    - Technology Challenge Recognition Ceremony
- BOF "Global Research Platform (GRP): Creating Worldwide Advanced Services and Infrastructure for Science," organized by Maxine Brown, Buseung Cho, Cees de Laast, Paola Grosso, Francois Lee, Marek Michalewicz.
  - Tuesday, November 19, 5:15 - 6:45pm, room 710
    - "Globally Distributed Secure Data Exchange Fabric", by Paola Grosso and Cees de Laast

## A secure network overlay for tracking and enforcement of data transaction rules.

Ralph Koning, Reginald Cushing Lu Zhang, Cees de Laast, Paola Grosso, University of Amsterdam

Competing companies can, together, generate value from collaborating on data and compute. Examples include airlines industry, ports, healthcare.

Clearly this poses a challenge of how to facilitate such collaborations through technology. Here we look at one piece of the puzzle (i.e. setting up distributed multi-domain infrastructures between such parties to facilitate the running of applications.

**Motivation**  
Multi-Domain distributed applications need to share data and compute under different policies.

**Challenges**  
- Map data sharing policies to infrastructure  
- Build an infrastructure that facilitates these applications.  
- Control sharing of data and compute.  
- Audit activity of the network.  
- Minimize risk of policy/security breaches.

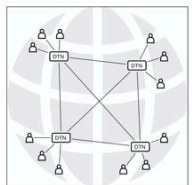
**Control functions**  
- Securing bucket-to-bucket communication through transaction specific VPNs  
- Bucket node key address used as VPN keys  
- Opening connection endpoints on demand. Bucket containers have no network interface. Interfaces are only created and attached per signed transaction.

**Overlay**  
- Nodes on the network are addressed using their public key  
- Nodes include: domain controllers, data buckets, auditors, application verifiers, users  
- Keys create chains of trust and permission through cryptographic signature trails.  
- Applications are decomposed to a set of transactions.  
- Transactions drive the overlay.

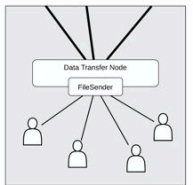
**Network of Auditors**  
- Auditor nodes on the network provide a signing and verification layer that is checked by the control layer.  
- Auditors sign network actions based on their internal policy.  
- Auditors are independent of each other.  
- The more number of signatures an action gets (e.g. transactions) the more confident the control layer is.  
- Auditors cross-verify each other's logs to minimize log tampering.

**In short...**  
- Overlay allows for a distributed infrastructure.  
- Key-based addressing allows for node signature trails and trust chains.  
- Network of auditors provides robust signing/verification of network transactions.  
- Control layer enforces security using inputs from auditors and minimizes attack vectors on data transfers.

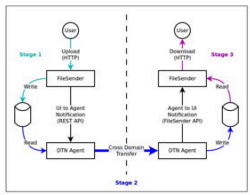
## User Friendly Data Transfers with DTNs



- A worldwide network of DTNs supports users by allowing long distance transfers to be done by specialized systems on optimized paths.
- Data transfer domains are determined based on geographical region or organization.
- DNS service records are used to find the appropriate DTN for a recipient's domain with support for user specific entries.
- DNS service records are used to find the appropriate DTN for a recipient's domain with support for user specific entries.
- Between domains the DTNs provide an API that allows for peer DTNs to choose from a selection of protocols designed for high speed data transfers.

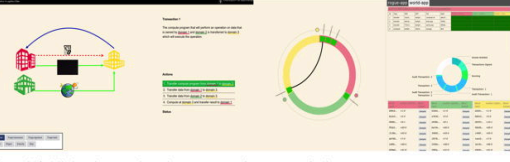


- Within a domain the DTN provides an API that is utilized by user facing data transfer applications such as FileSender.
- Users continue to use the tools they are familiar with without any need to consider the location of the recipient.
- Data transfers between domains are performed faster with both the sender and the recipient using their preferred data transfer application.
- There is no change in how data transfers between users within the same domain are performed.



- The transfer is carried out in three stages
  - Stage 1: User Upload
  - Stage 2: Cross Domain Transfer
  - Stage 3: User Download
- Notifications signal the transition between stages
- FileSender need not be run on the same system as the DTN Agent

Proof of Concept, see <https://dl4ld.nl/>



This research is funded by the Dutch Science Foundation in the Comm2Data program (grant no. 628.001.001) and by Equinox.

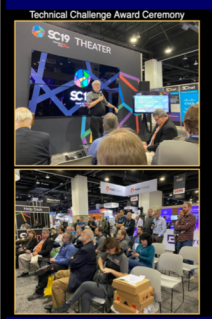


### A secure network overlay for DDM

### A secure network overlay for tracking and enforcement of data transaction rules.

**Motivation**  
Multi-Domain distributed applications need to share data and compute under different policies.

**Challenges**  
- Map data sharing policies to infrastructure.  
- Build an infrastructure that facilitates these applications.  
- Control sharing of data and compute.  
- Audit activity of the network.  
- Minimize risk of policy/security breaches.



### DTN assisted FileSender

### User Friendly Data Transfers with DTNs

A secure network overlay for  
monitoring and enforcement of  
multi-domain applications.

# 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/seconnet>
  - <https://towardsamdex.org>

