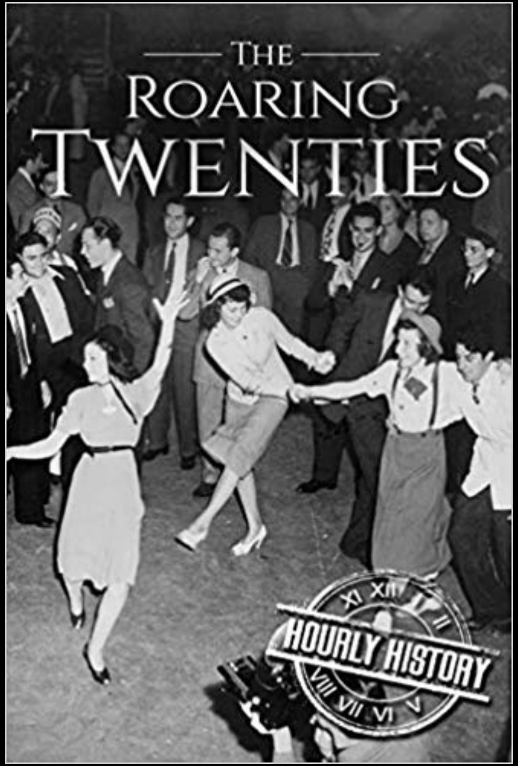
ICT to support the transformation of Science in the Roaring Twenties

Cees de Laat

Systems and Networking Laboratory
University of Amsterdam
Affiliate Lawrence Berkeley National Laboratory

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.

AIM

- Observe how the art of Science is transforming with AI & ML.
- Understand how the ICT world looks like in 2030.
- Understand what hinders Science, Industry, Society to progress.
- What is needed to obtain EU leadership
  - Why?
  - Where?
  - What?

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

Computer-Aided Drug Discovery *Smart Cities* *Disaster Resilience and Response*

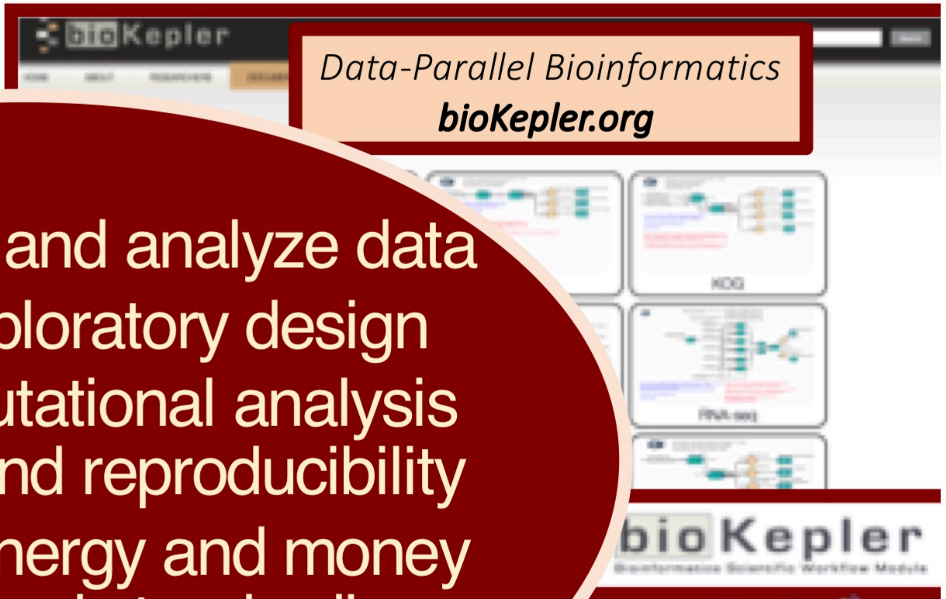
Smart Manufacturing *Personalized Precision Medicine* *Smart Grid and Energy Management*

Workflows for Data Science Center of Excellence at SDSC

WorDS.sdsc.edu

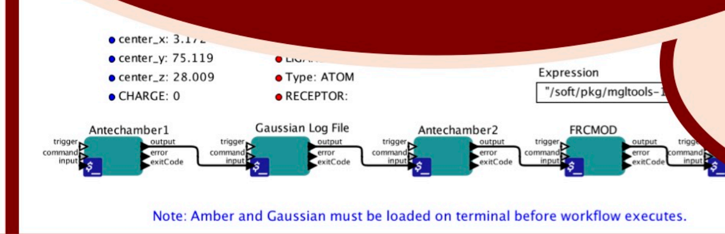
Real-Time Hazards Management
wifire.ucsd.edu

Data-Parallel Bioinformatics
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

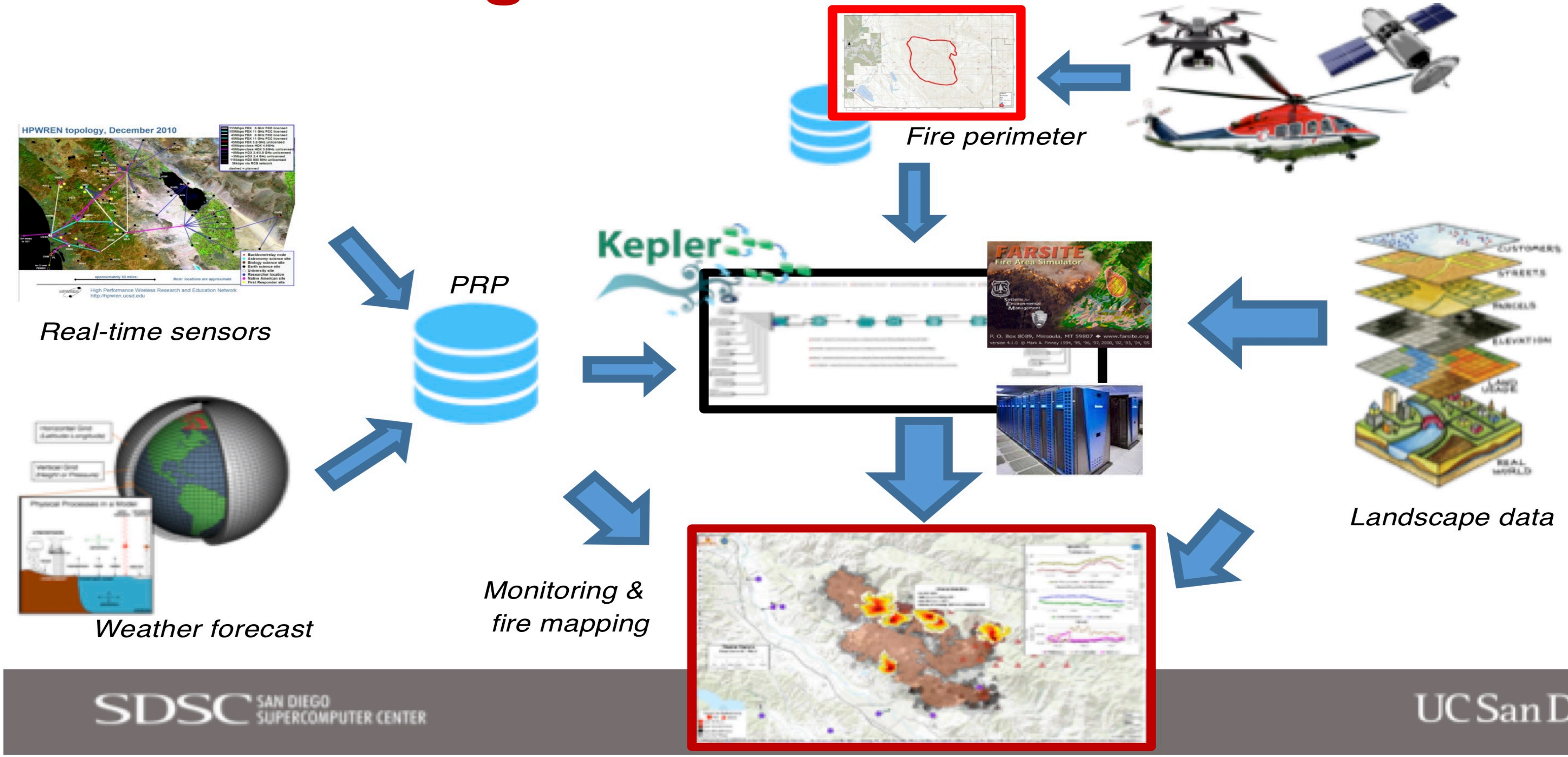
kepler-project.org



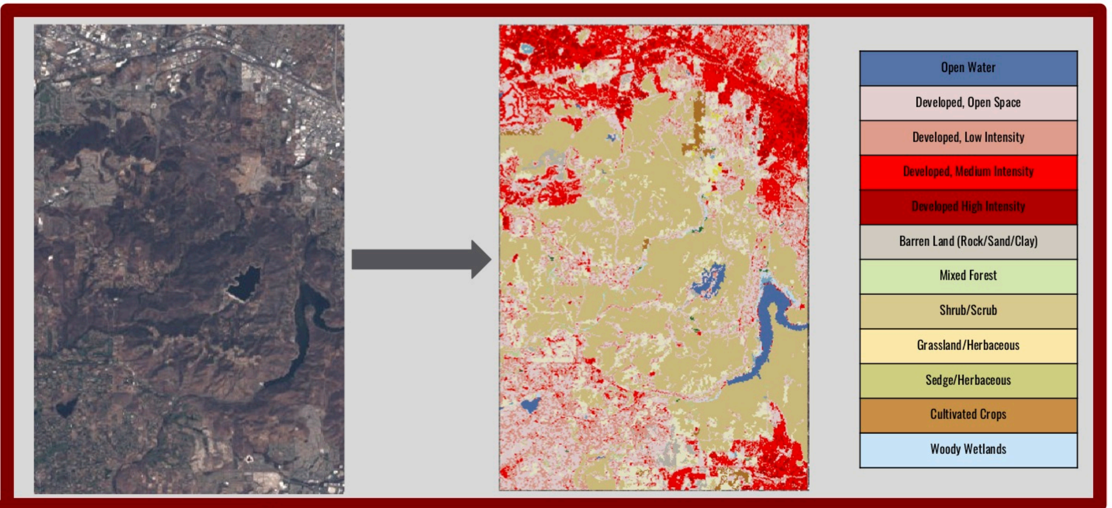
Goal: Methodology and tool development to build automated and operational workflow-driven solution architectures on big data and HPC platforms.

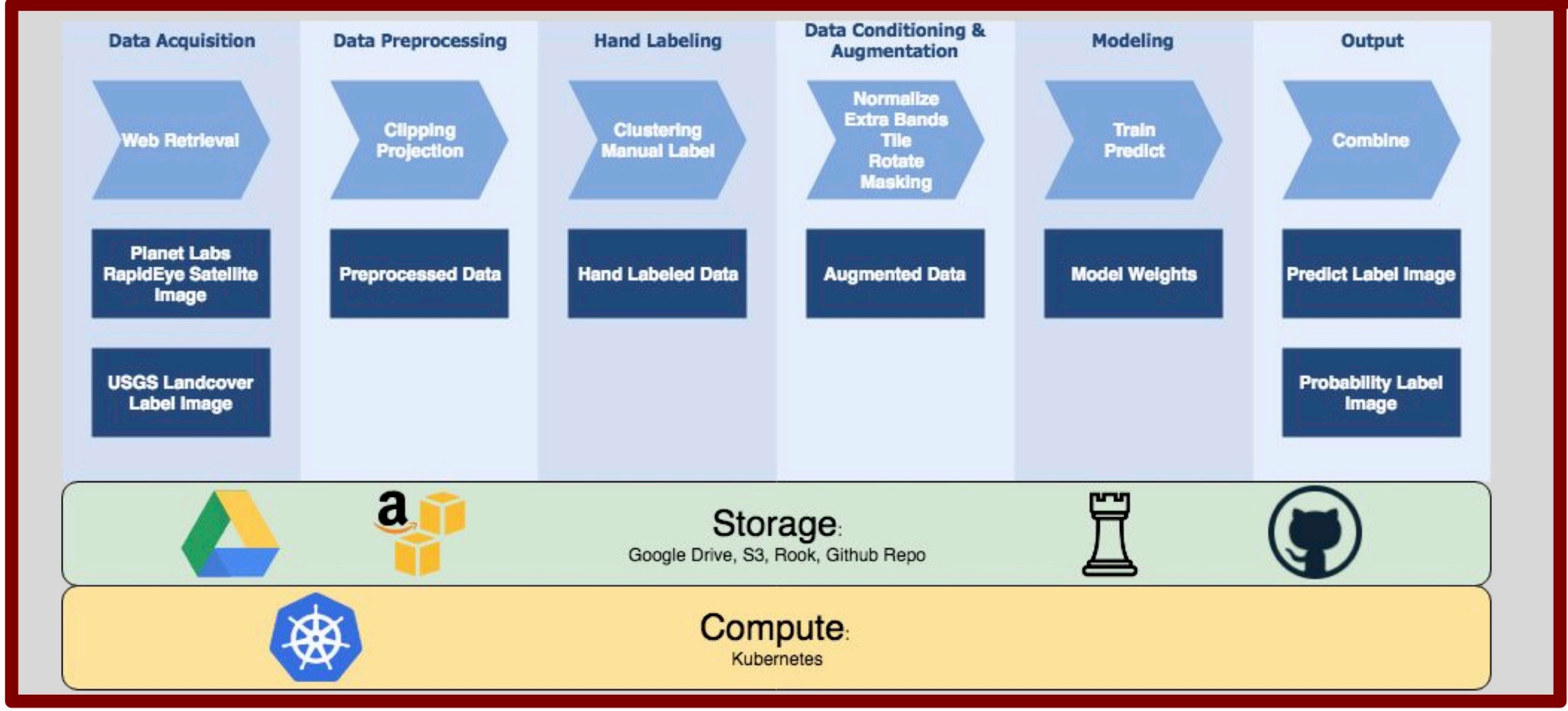
Scalable Automated Molecular Dynamics and Drug Discovery
nbcr.ucsd.edu

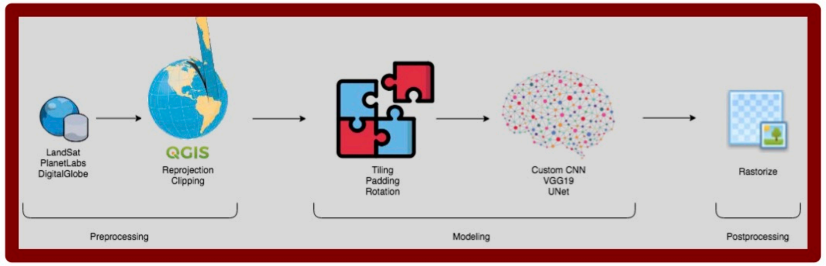
Fire Modeling Workflows in WIFIRE

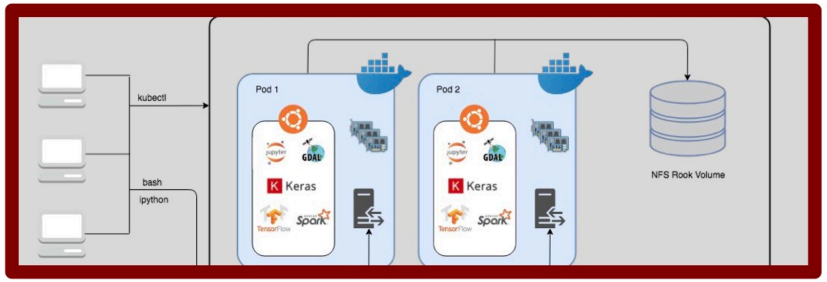


One Piece of the Puzzle: Vegetation Classification using Satellite Imagery





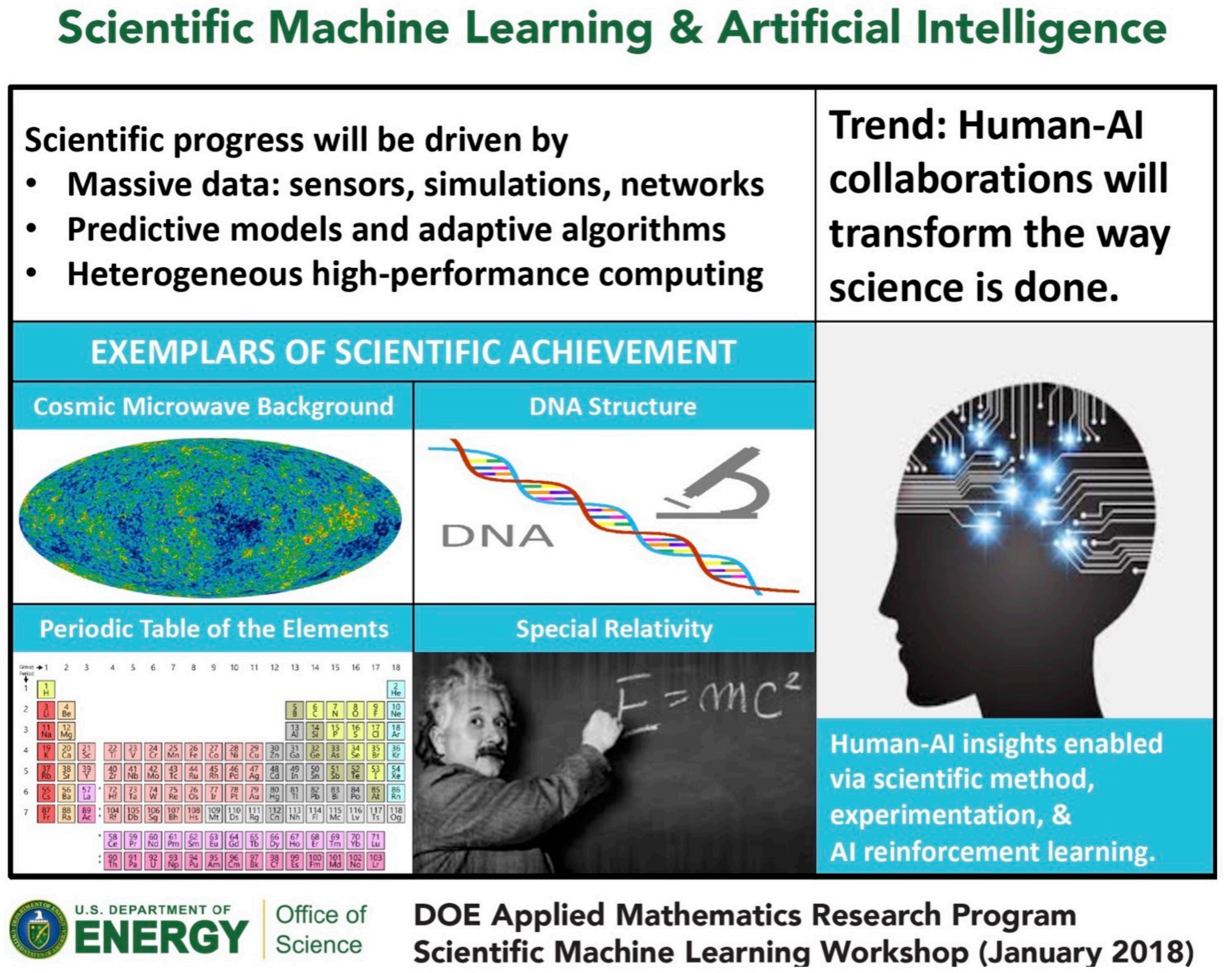




SDSC SAN DIEGO SUPERCOMPUTER CENTER

İlkay ALTINTAŞ, Ph.D.
ialtintas@ucsd.edu

UC San Diego



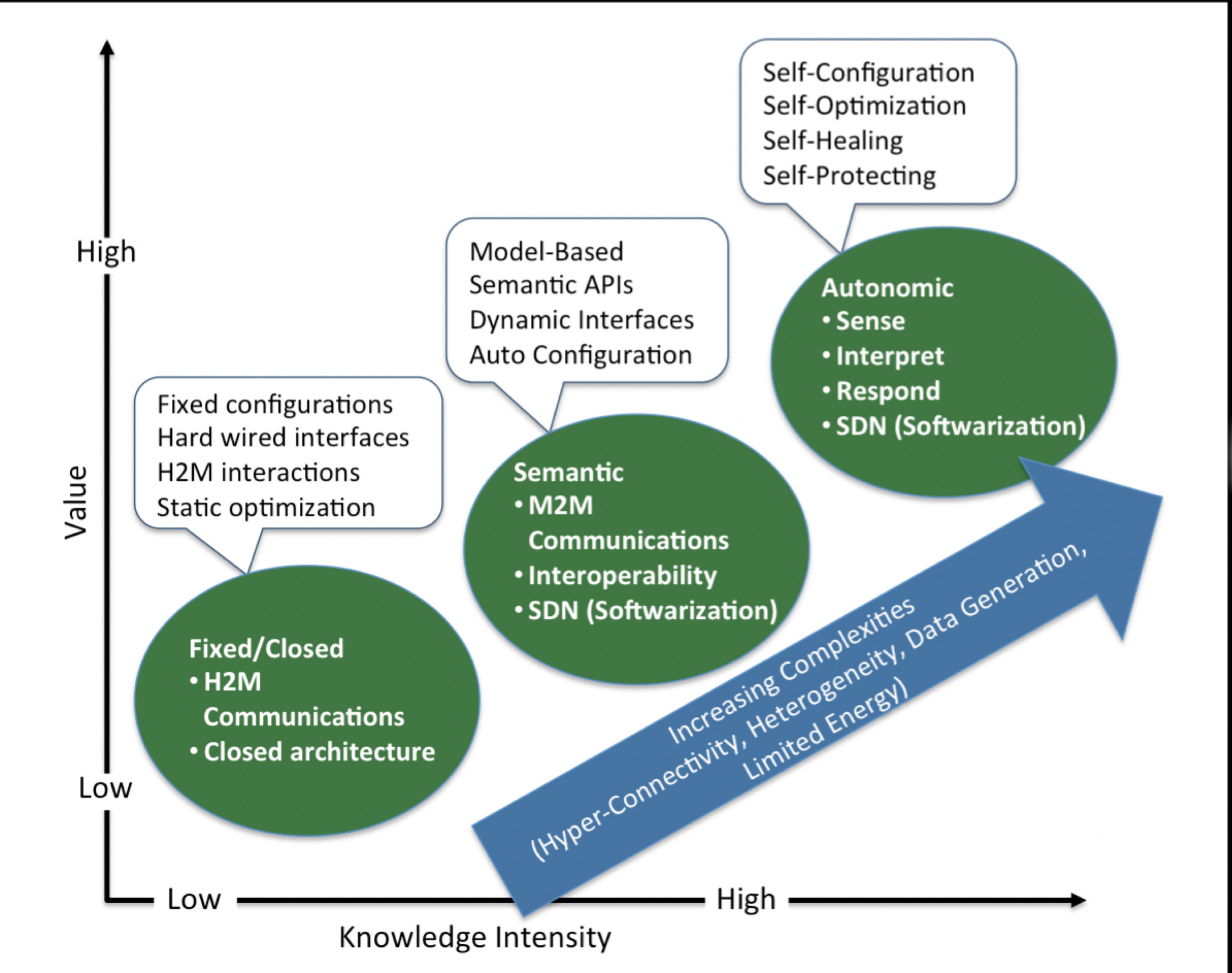
Workshop report:

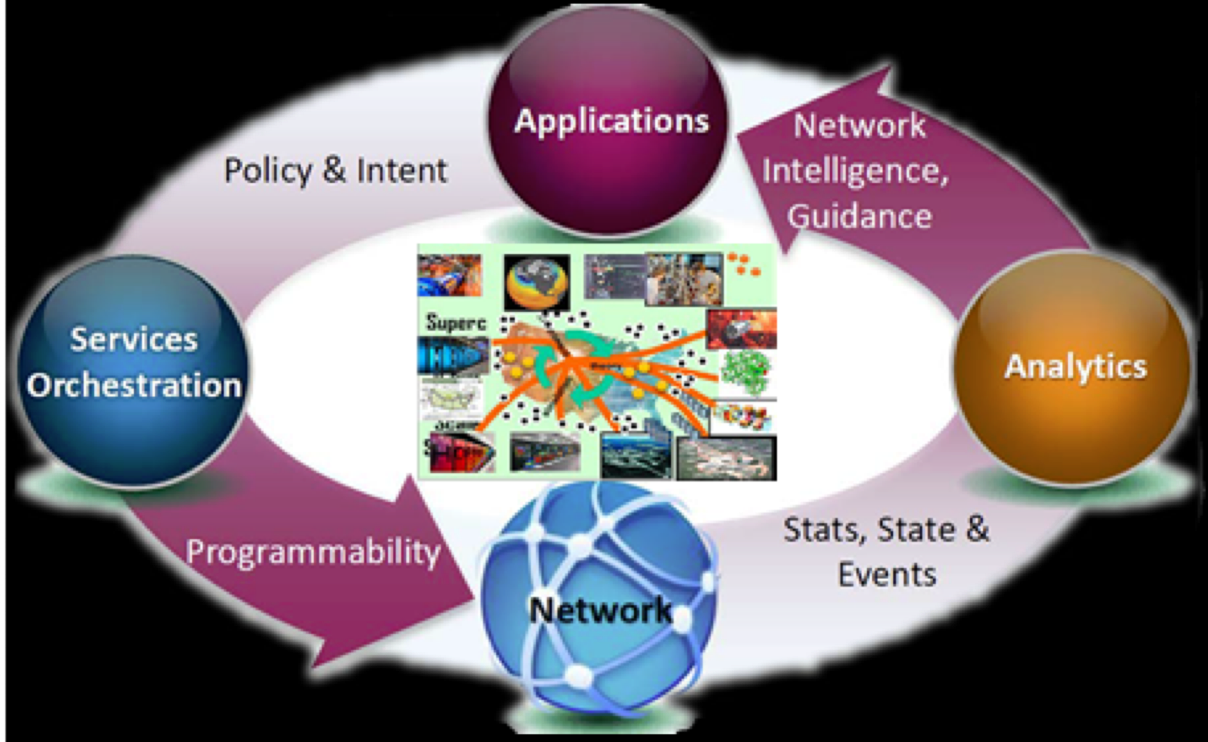
https://www.osti.gov/biblio/1478744

DoE workshop on smart networks

Bring AI in control plane to harness complexity

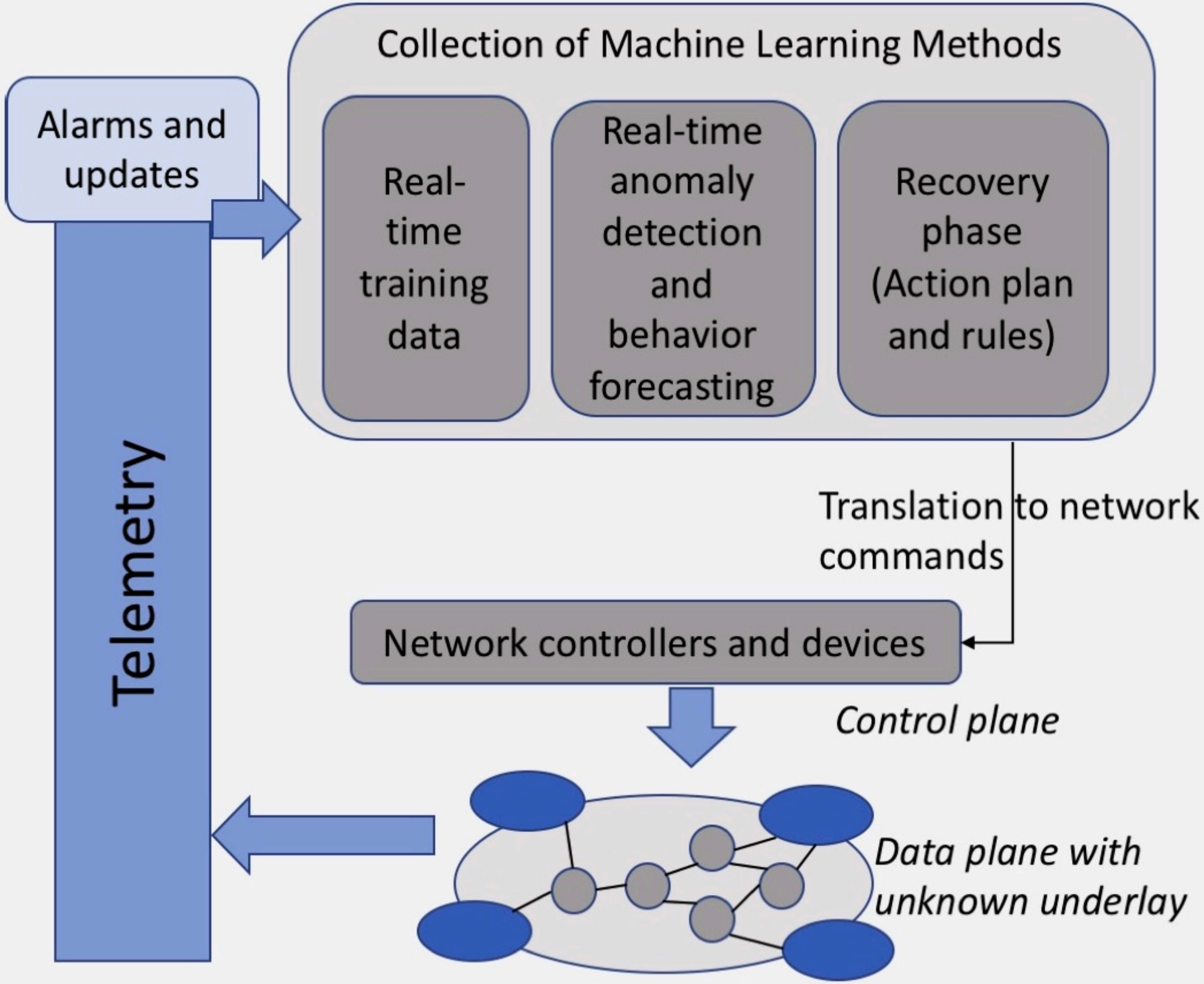
https://www.orau.gov/smarthp2016/





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-



Rethinking NSF's Computational Ecosystem for 21st Century Science and Engineering

Workshop Website: https://uiowa.edu/nsfcyberinfrastructure

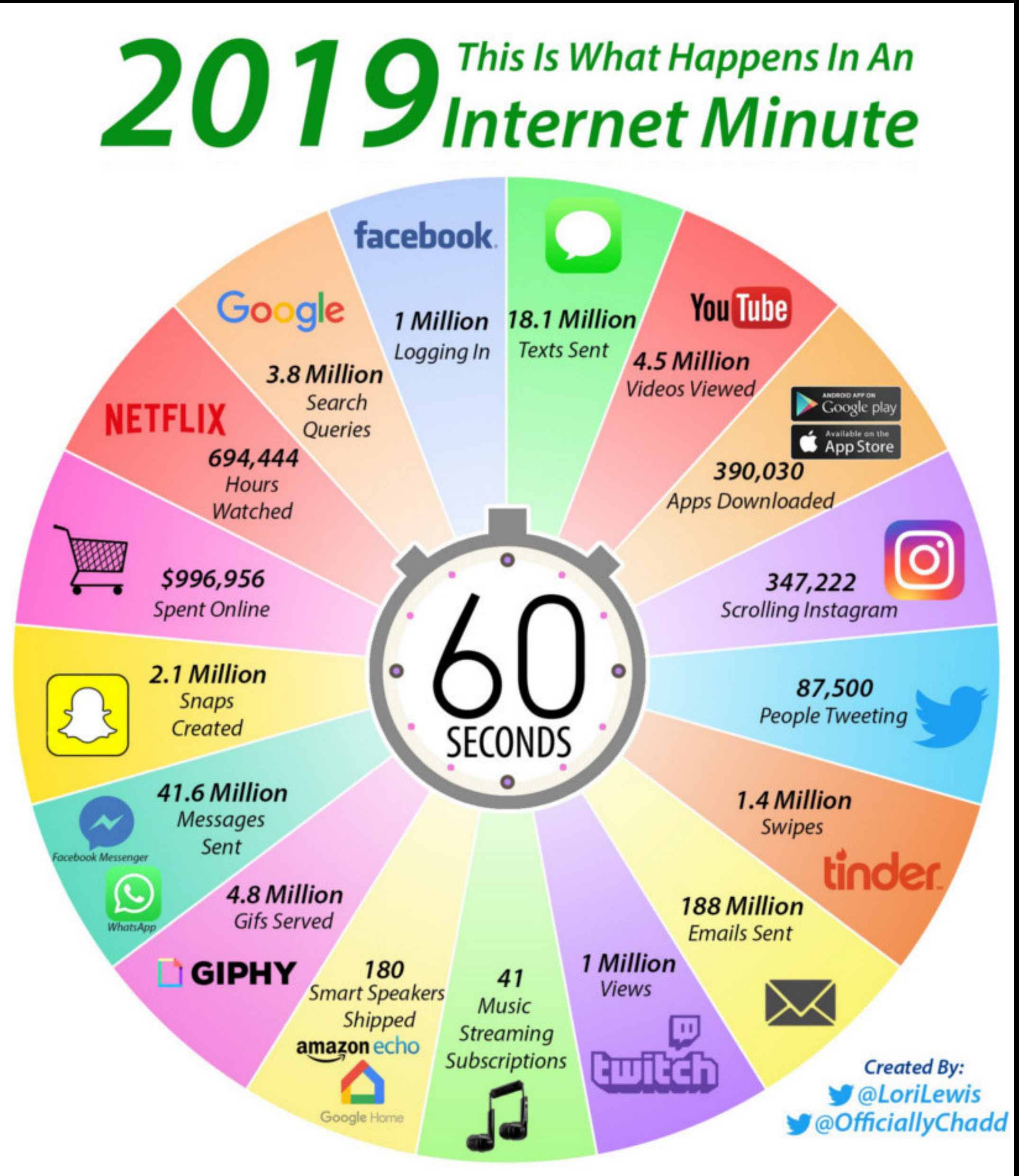
Workshop Report: https://www.uiowa.edu/nsfcyberinfrastructure/report.pdf

Initial debates about resource management and delivery options focused on expert personnel as a critical component of successful cyberinfrastructure delivery. Several examples such as Campus Champions (CC) or XSEDE's ECSS were described as critical to scientific advance but insufficient in numbers to meet demand. Regionally tasked staff might help to alleviate this shortfall. Benefits could include greater use of cloud or national resources if there was a local expert to help researchers with initial utilization. Along these lines, it was mentioned that the NSF CC* programs changed campus culture, spurring local networking expertise. A similar program to promote workforce development to incentivize local computational and data scientists could, for instance, result in the integration of otherwise isolated clusters on campuses with national resources. These key personnel, ranging from ECSS experts and developers to CCs, are often in careers that need professionalization.

Change in computing

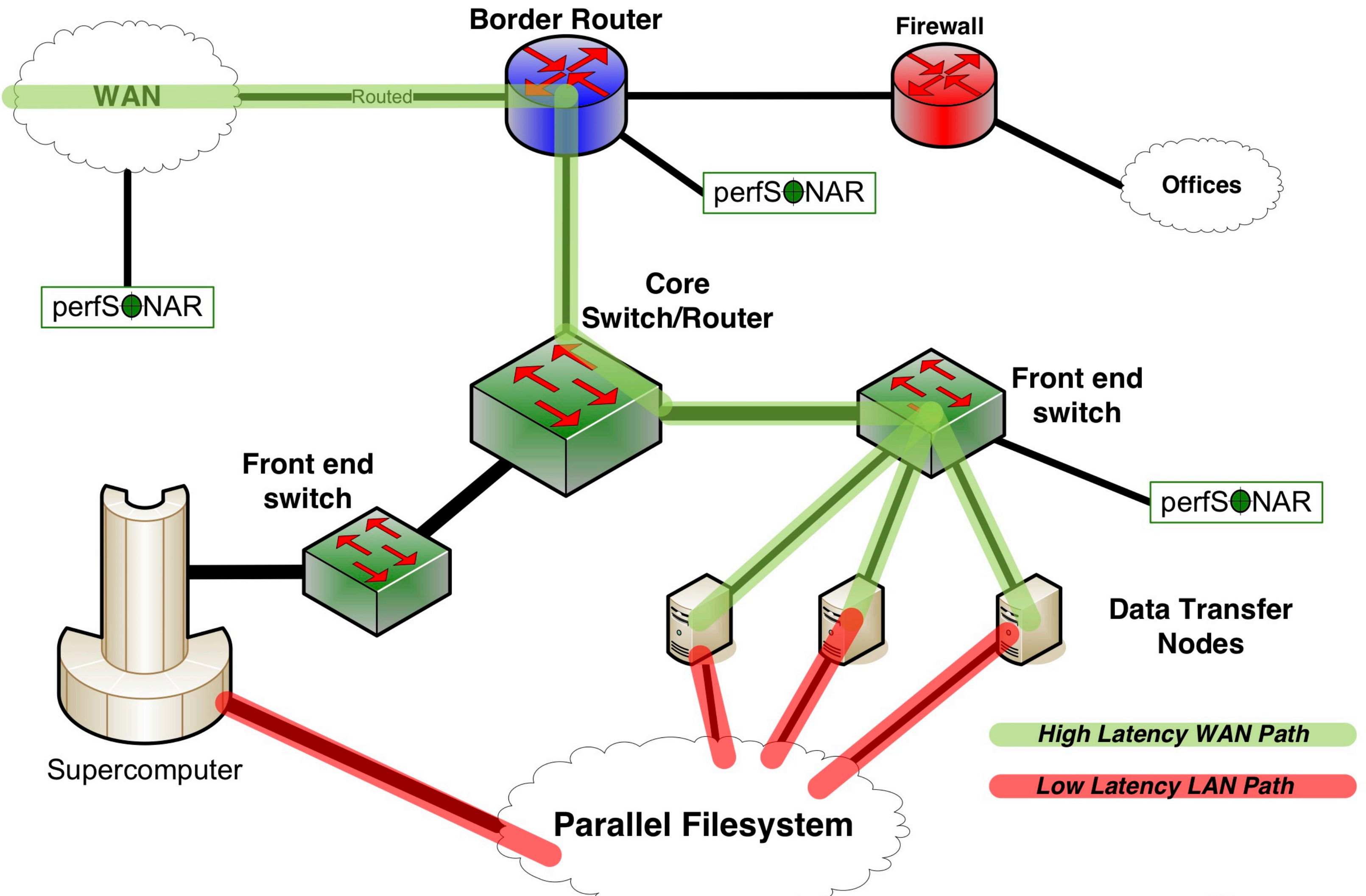
- Early days a few big Supercomputers
  - Mostly science domain
- Via grid to commercial cloud
  - AWS, Azure, Google Cloud, IBM, Salesforce
  - The big five: Apple, Alphabet, Microsoft, Facebook and Amazon
  - Computing has transformed into an utility
- Data => Information is the key

Now, how do we get and use data?



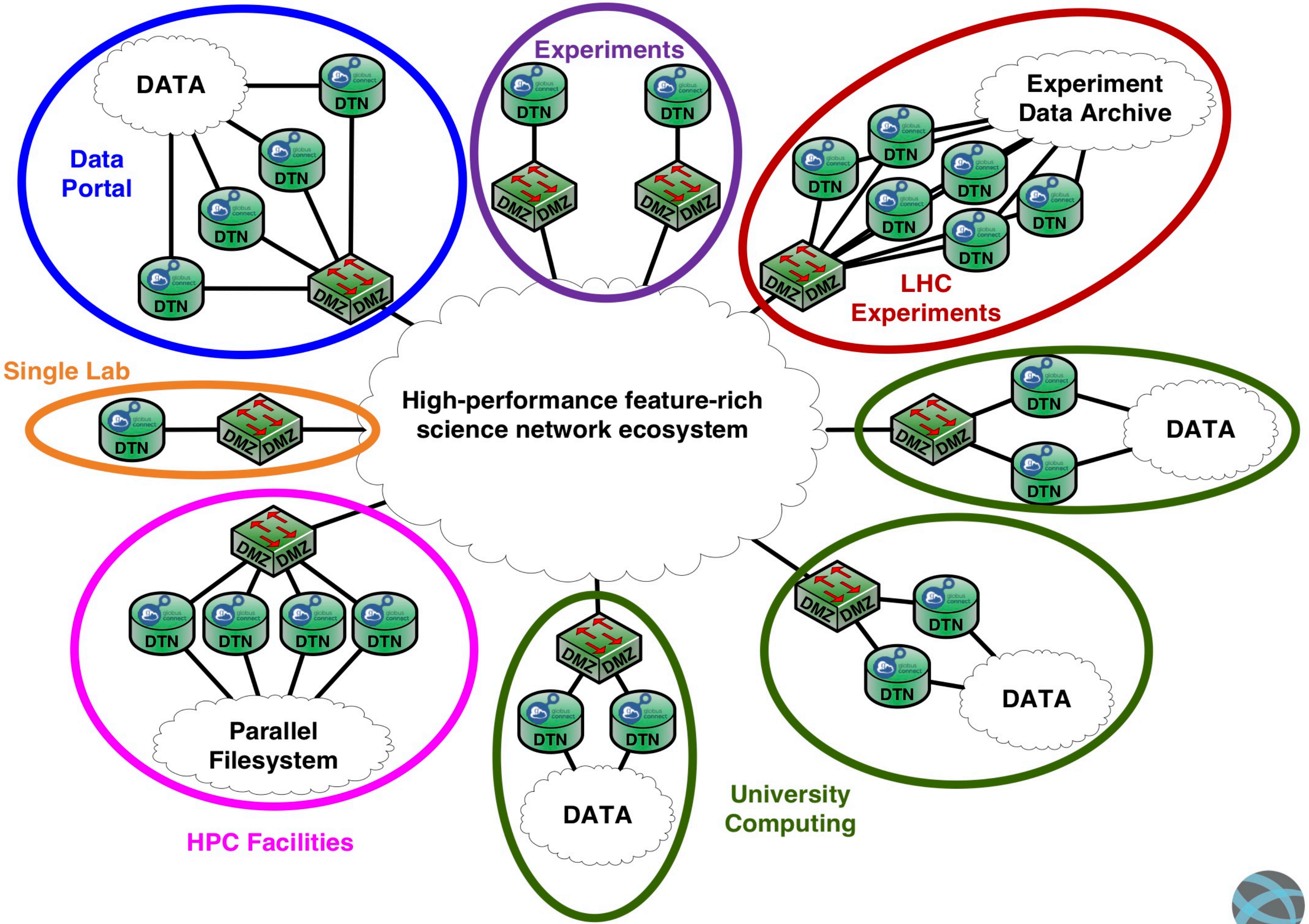
- Move towards streaming
  - Netflix
  - youtube
- Same in science world
  - SKA / LOFAR
  - Light Source
  - Environmental (Marine, Meteorology, …)
- Data is not always huge
  - Sometimes it is very complex
  - Some examples

Science DMZ – HPC Center DTN Cluster



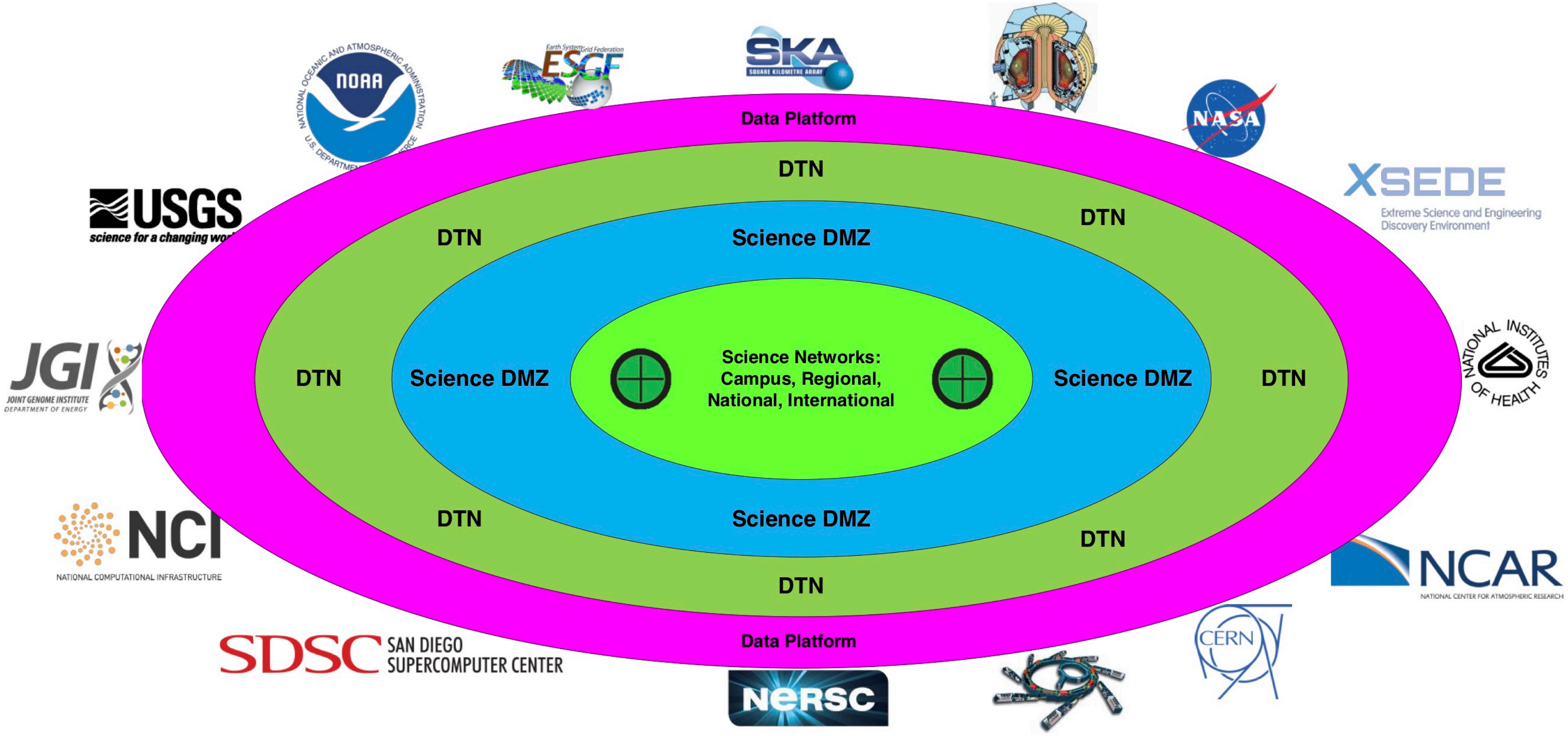
Courtesy Eli Dart, ESnet

Science DMZs for Science Applications



Courtesy Eli Dart, ESnet

Data Ecosystem – Concentric View



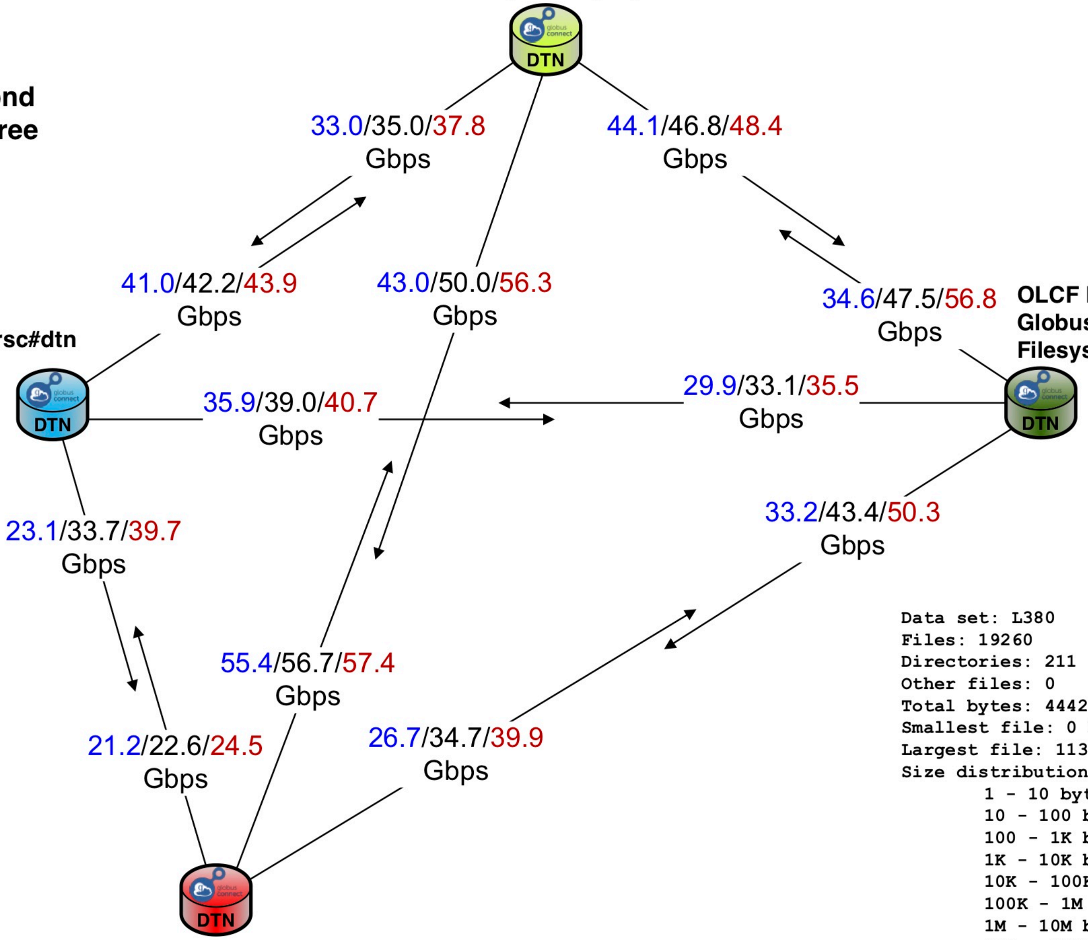
Courtesy Eli Dart, ESnet

DTN Cluster Performance – HPC Facilities (2017)

Petascale DTN Project

November 2017
L380 Data Set

Gigabits per second (min/avg/max), three transfers

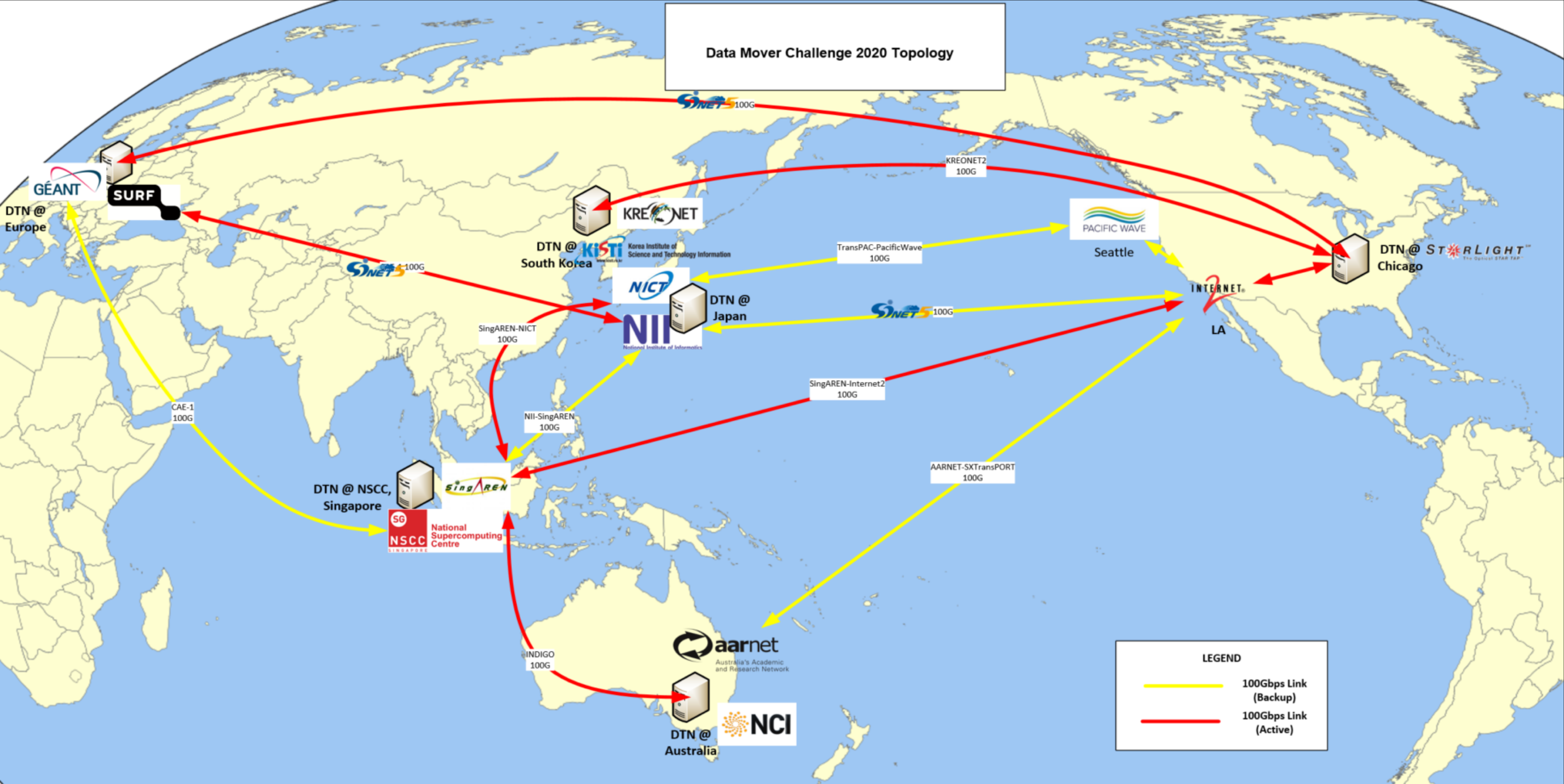



```
Data set: L380
Files: 19260
Directories: 211
Other files: 0
Total bytes: 4442781786482 (4.4T bytes)
Smallest file: 0 bytes (0 bytes)
Largest file: 11313896248 bytes (11G bytes)
Size distribution:
        1 - 10 bytes: 7 files
        10 - 100 bytes: 1 files
        100 - 1K bytes: 59 files
        1K - 10K bytes: 3170 files
        10K - 100K bytes: 1560 files
        100K - 1M bytes: 2817 files
        1M - 10M bytes: 3901 files
        10M - 100M bytes: 3800 files
        100M - 1G bytes: 2295 files
        1G - 10G bytes: 1647 files
        10G - 100G bytes: 3 files
```

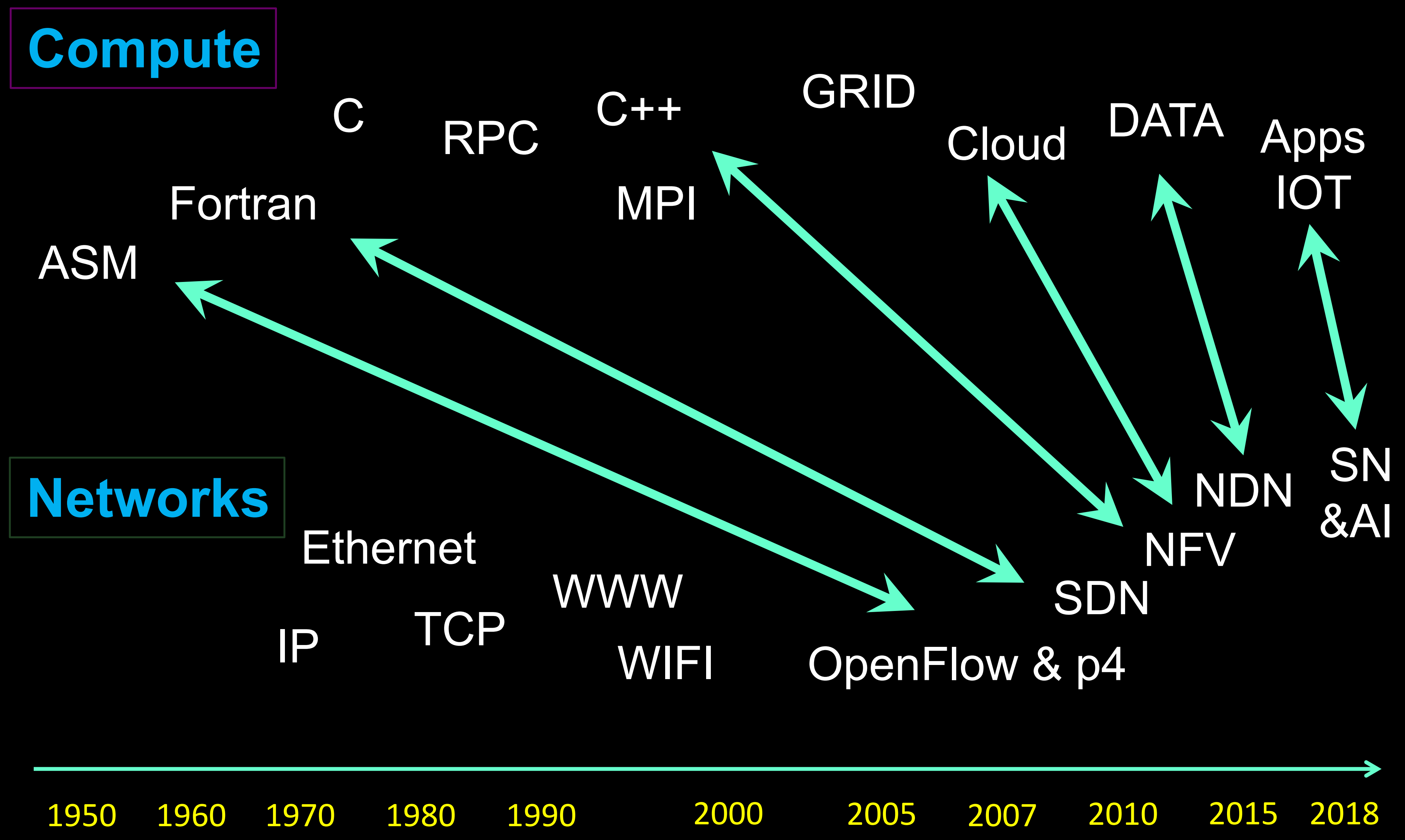
ESnet

Courtesy Eli Dart, ESnet

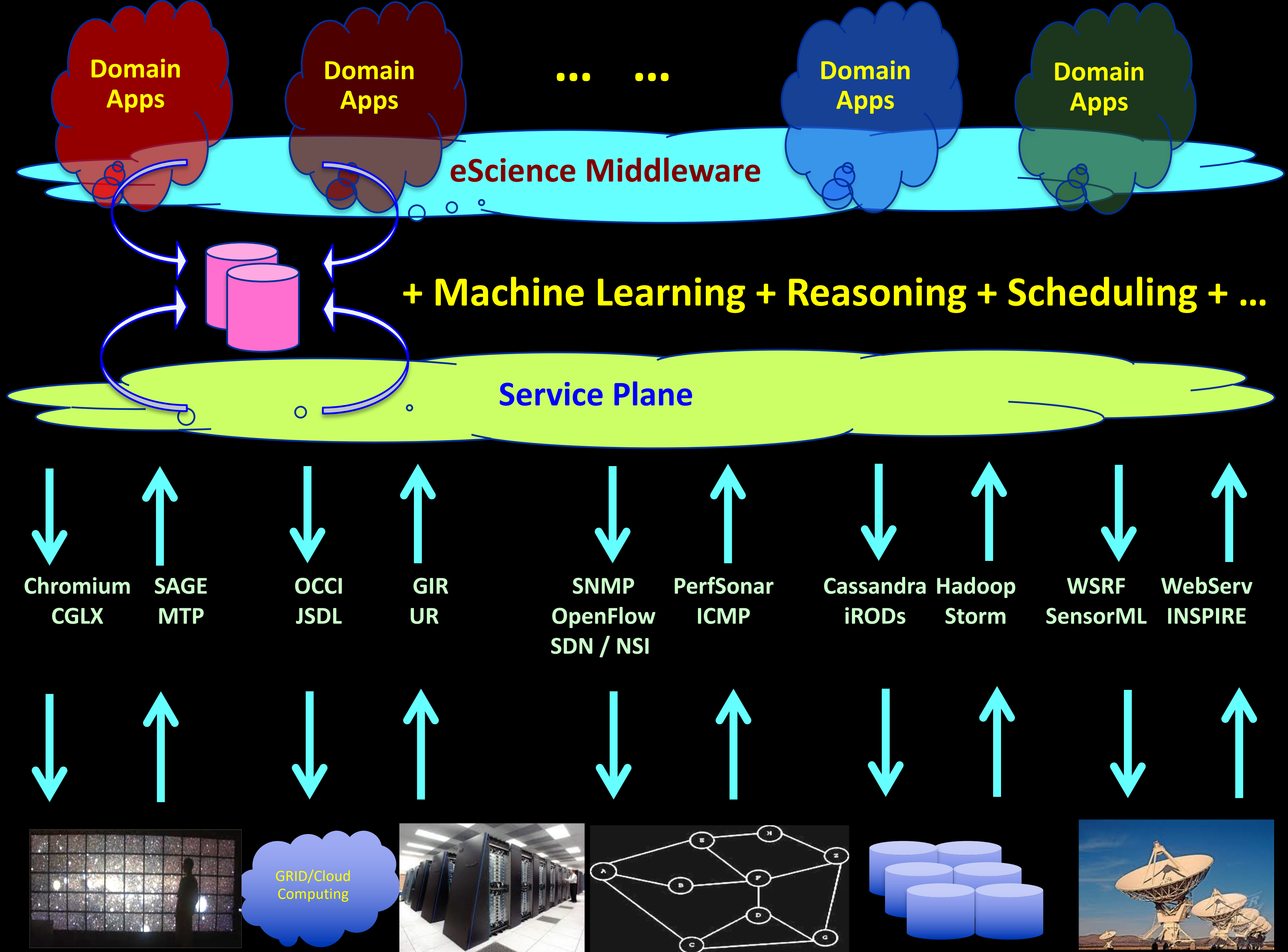
https://www.sc-asia.org/data-mover-challenge/



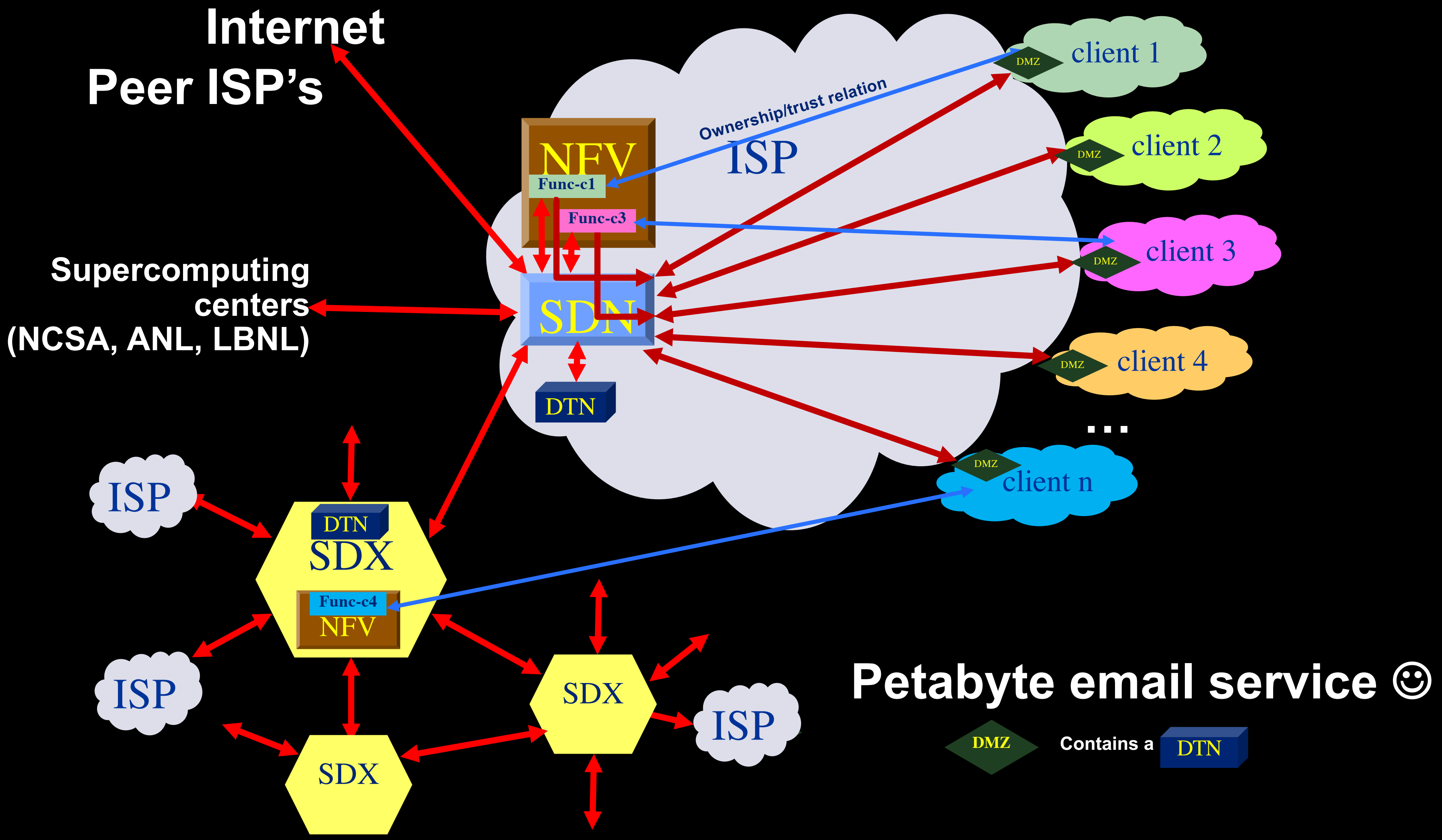
TimeLine



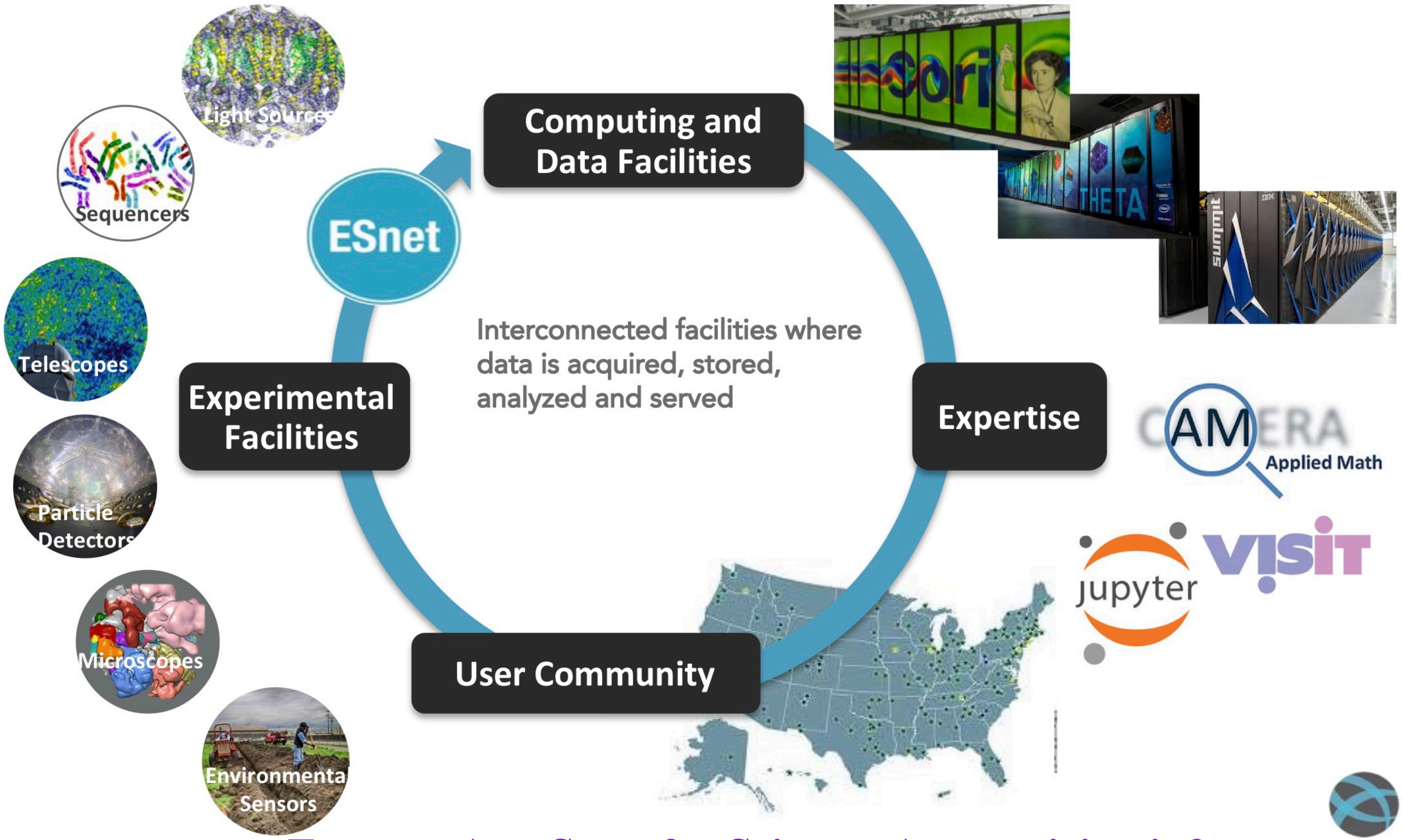
Compute
C
RPC
C++
GRID
Cloud
DATA
Apps
IOT
Fortran
MPI
ASM
Networks
Ethernet
WWW
IP
TCP
WIFI
OpenFlow & p4
SDN
NFV
NDN
SN
&AI
1950
1960
1970
1980
1990
2000
2005
2007
2010
2015
2018


Domain Apps
Domain Apps
... ...
Domain Apps
Domain Apps
eScience Middleware
+ Machine Learning + Reasoning + Scheduling + ...
Service Plane
Chromium CGLX
SAGE MTP
OCCI JSDL
GIR UR
SNMP OpenFlow SDN / NSI
PerfSonar ICMP
Cassandra iRODs
Hadoop Storm
WSRF SensorML
WebServ INSPIRE
GRID/Cloud Computing

Networks of ScienceDMZ’s & SDX’s



Superfacility Model for Productive, Reproducible Science



ESnet

Extreme App Store for Science; Amazonizing infrastructure

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 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?

Harvard Business Review

Harvard Business Review

Subscribe | Sign In | Register

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 3 COMMENT TEXT SIZE PRINT $8.95 BUY COPIES

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

➔

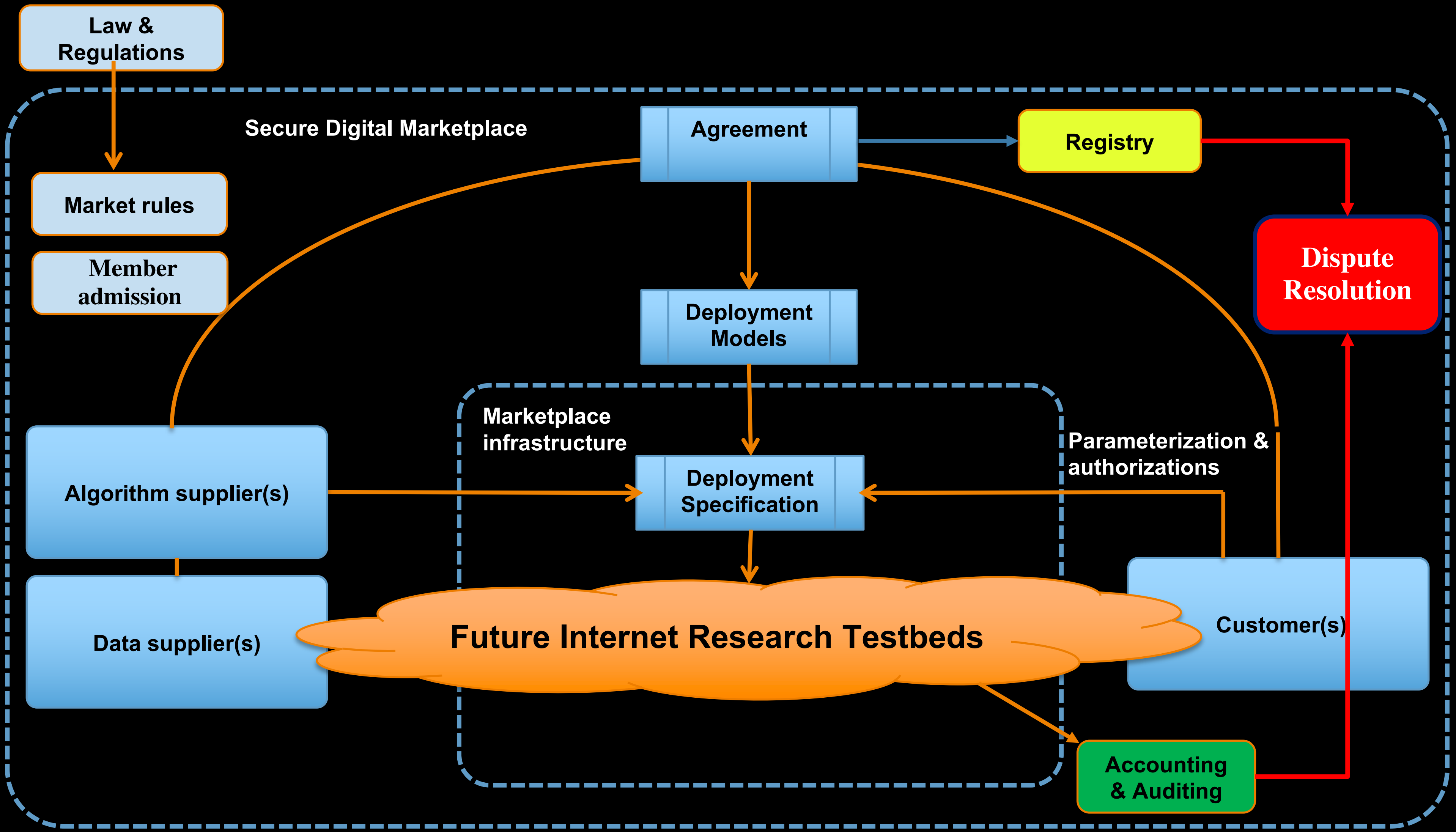
Sound Market principles

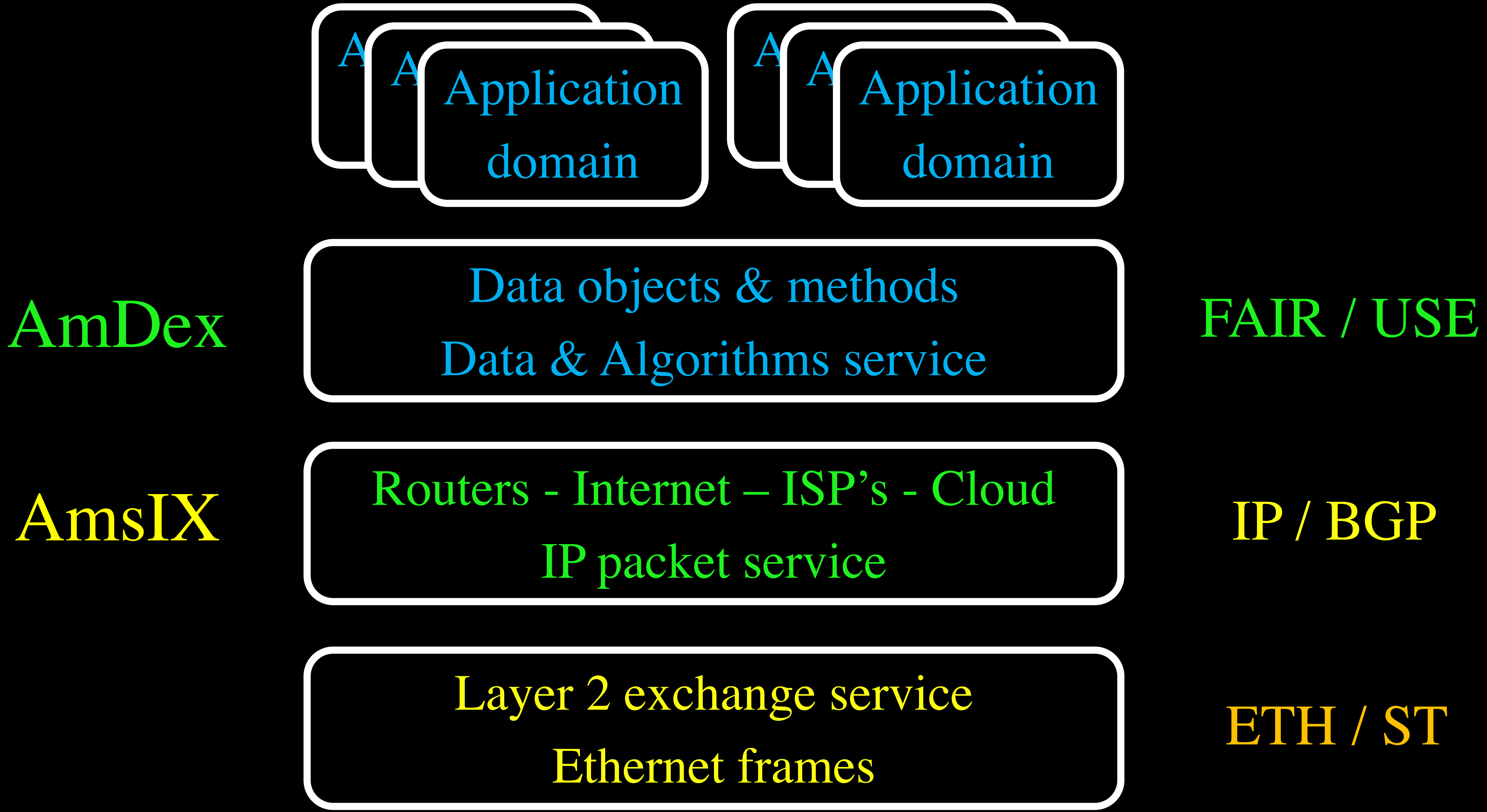
https://hbr.org/2017/09/managing-our-hub-economy

Approach

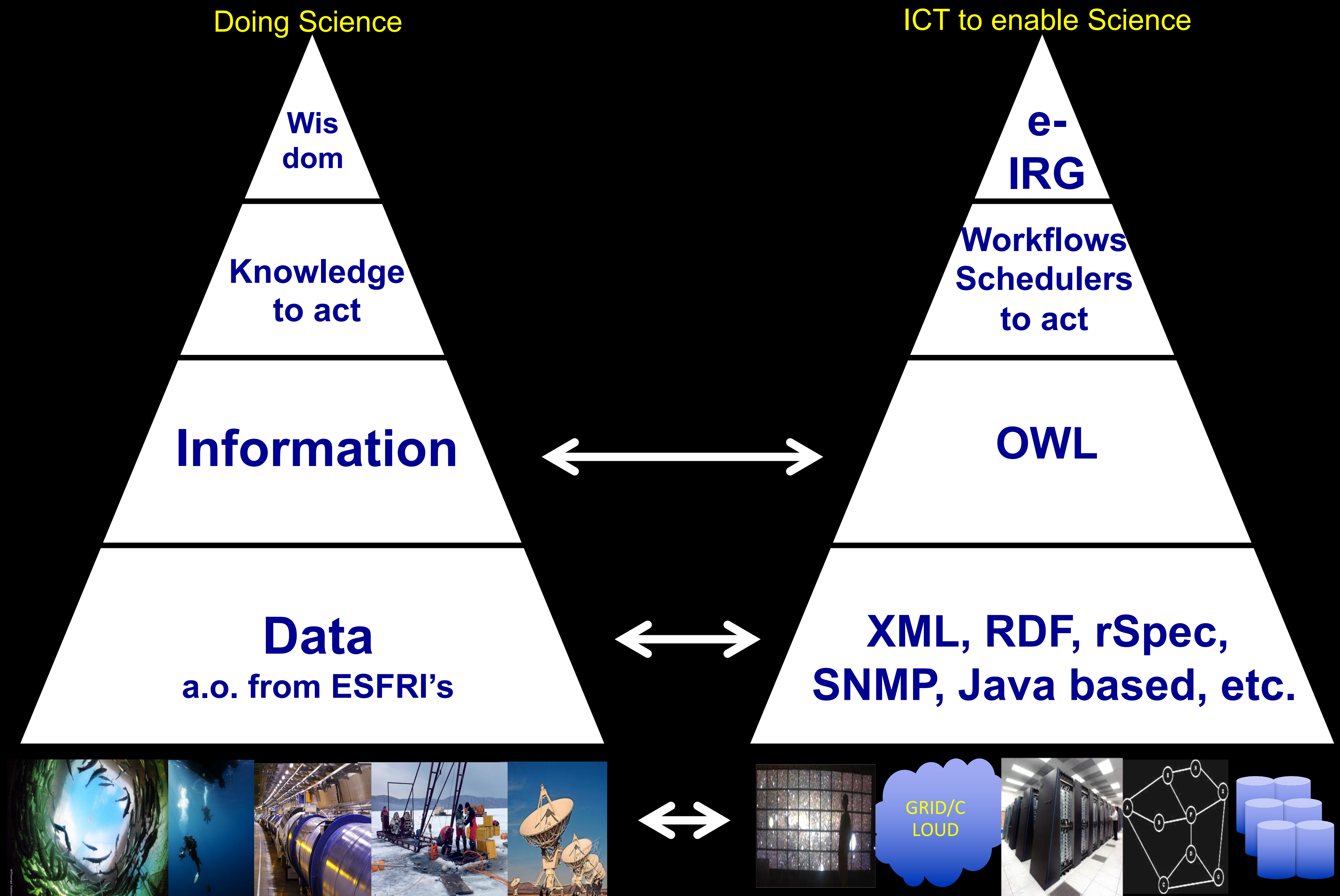
- Strategic:
  - Translate legislation into machine readable policy
  - Define data use policy
  - Trust evaluation models & metrics
- Tactical:
  - Map app given rules & policy & data and resources
  - Bring computing and data to (un)trusted third party
  - Resilience
- Operational:
  - TPM & Encryption 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!)

Secure Digital Market Place Research

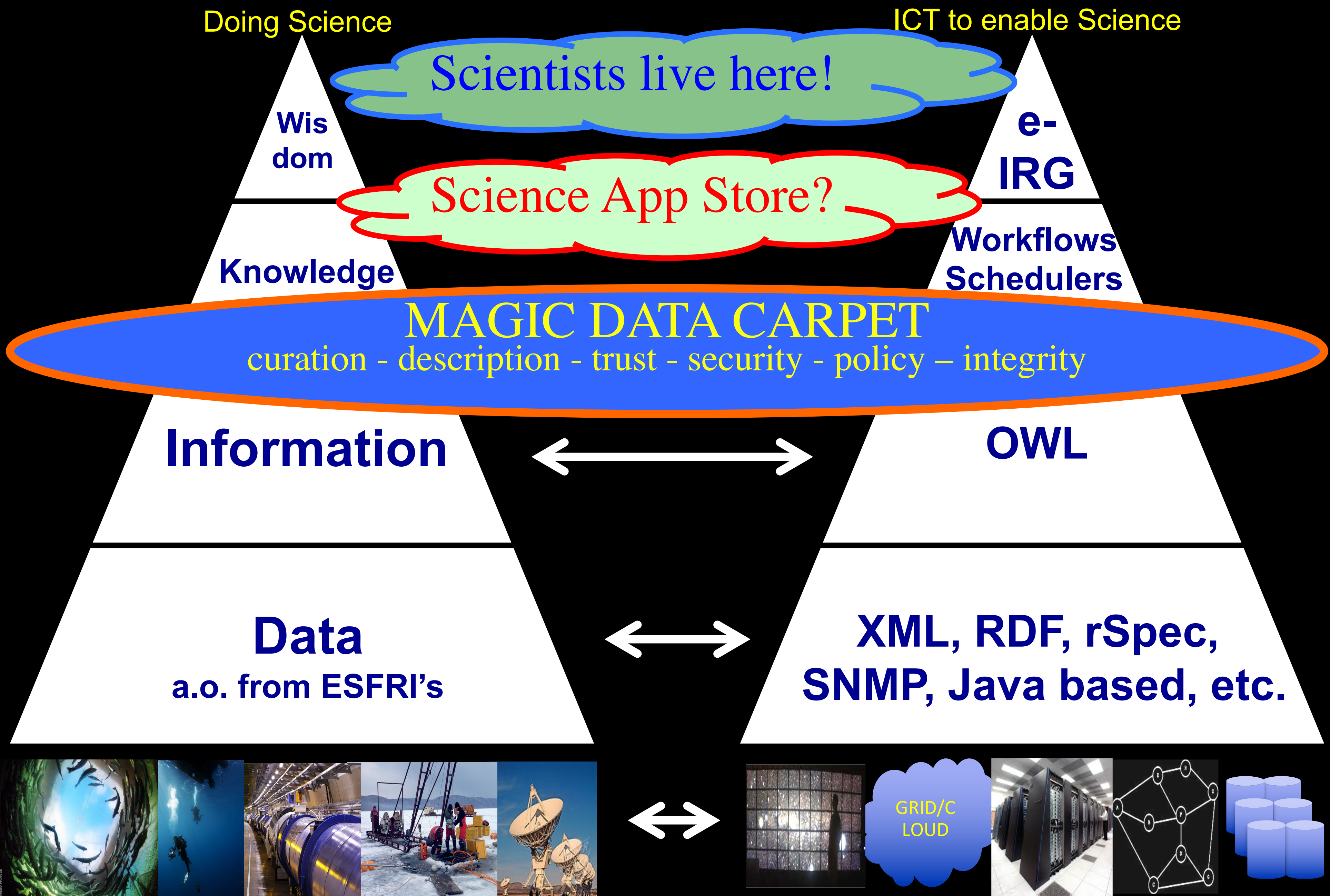



Application domain
Application domain
AmDex
Data objects & methods
Data & Algorithms service
FAIR / USE
AmsIX
Routers - Internet – ISP's - Cloud
IP packet service
IP / BGP
Layer 2 exchange service
Ethernet frames
ETH / ST

The Big Data Challenge



The Big Data Challenge

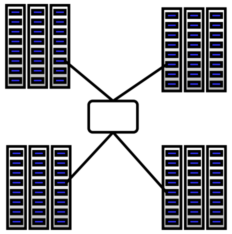


Past & future ICT research infrastructures

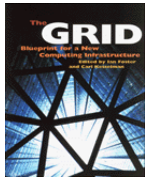
- TEN34 / TEN155
- Geant testbed & JRA’s
- FIRE
- Grid5000 (FR)
- DAS1-5 (NL)

Some years around 2010 connected by LightPath

DAS generations: visions



- DAS-1: Wide-area computing (1997)
  - Homogeneous hardware and software
- DAS-2: Grid computing (2002)
  - Globus middleware



- DAS-3: Optical Grids (2006)

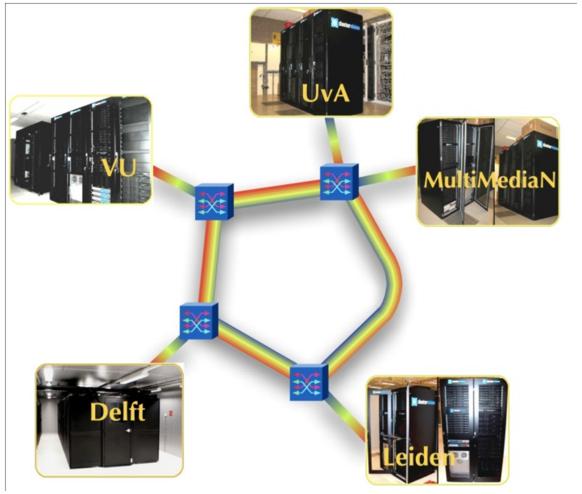
StarPlane

  - Dedicated 10 Gb/s optical links between all sites
- DAS-4: Clouds, diversity, green IT (2010)

DAS 4

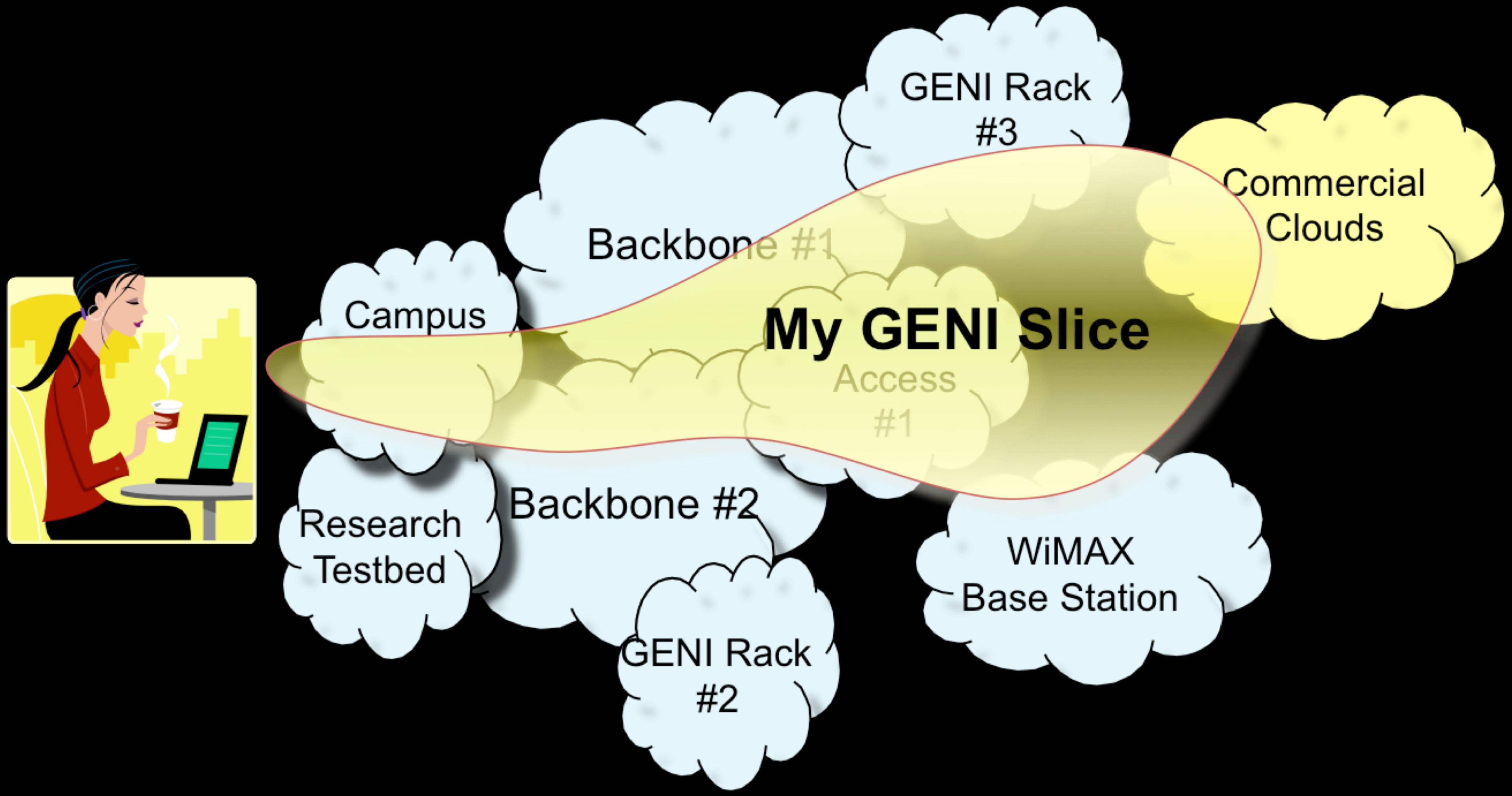
  - Hardware virtualization, accelerators, energy measurements
- DAS-5: Harnessing diversity, data-explosion (June 2015)
  - Wide variety of accelerators, larger memories and disks

das 5



Henri Bal, Dick Epema, Cees de Laat, Rob van Nieuwpoort, John Romein, Frank Seinstra, Cees Snoek, Harry Wijshoff, "The Distributed ASCI Supercomputer: a Long-term Computer Science Research Infrastructure.", IEEE Computer, Vol 49, Ed 5, pp 54-63, May 2016 (2015).

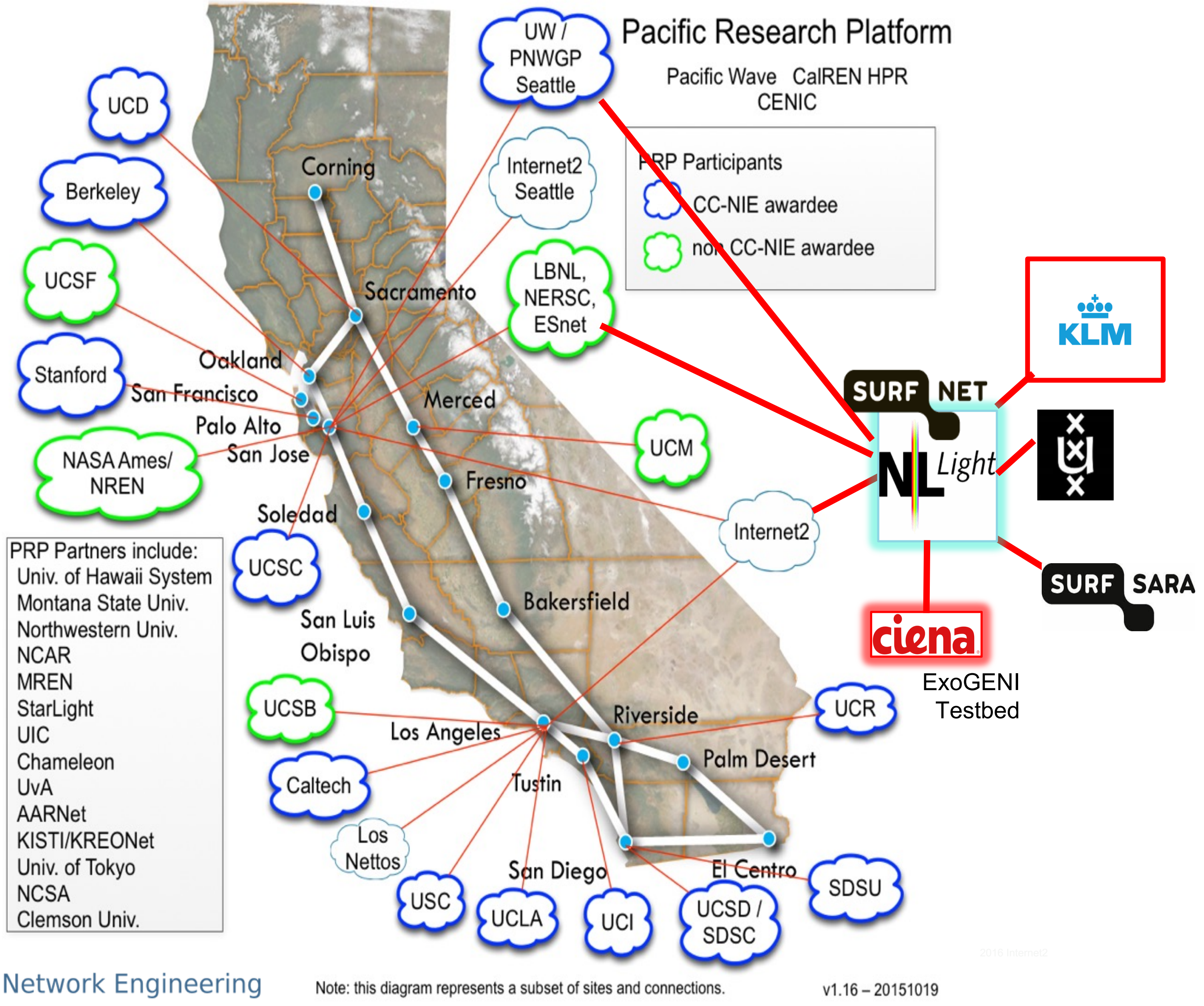
GENI: Virtualizing CI



Pacific Research Platform testbed involvement

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

PACIFIC RESEARCH PLATFORM



Past & future ICT research infrastructures

- TEN34 / TEN155
- Geant testbed & JRA's
- FIRE
- Grid5000 (FR)
- DAS1-5 (NL)

Was connected by LightPath around 2010!

- Need for breakable CS oriented testbed
- Must include: Programmable networks, Cloud, Exascale SC, DTN's, streaming, access to public services, IOT, Wireless
- Must include work on AI & ML, fundamental data security
- At Scale ➔ SILECS - https://www.silecs.net

CONCLUSIONS

- Observations:
  - parallels energy world and internet developments
    - move to micromarkets
    - iot alike security treats
  - trend: ML replaces Visualization
  - Illinois governor (1998) noting: canals - railroads - cars - fibers, and now we add trusted data exchange driving economy and markets
  - San Diego Super Center aligns with data science and portal for sustainability in RNE
  - LEGO model for CI & Data & Methods
  - Industry recognized need for new data related approaches
  - Data Value creates an economy for data sharing.

CONCLUSIONS

- Overal advice
  - It is about people & knowledge
  - Base on society relevant applications
  - Get faculty drivers from each campus
  - Governance model is essential
  - align with education (soft&hard money)
- Applications
  - Health
  - Instrumenting IOT
  - Energy transition/critical infrastructures IT
  - CyberSecurity

CONCLUSIONS

- Themes
  - global data & methods ecosystem supporting applications
  - Explainable AI to aid managing CI
  - Security
  - Superfacility
  - revisiting Internet standards with current technology in mind
  - Quantum compute and networking

Remarks, Quotes:

- Wouter Los: Considering the list of conclusions, it comes in my mind that any future data infrastructure should accommodate the preferred governance model. And this is related to the cultural dimension. What kind of data market do we foresee, what are checks and balances, and who decides (has power) on what? How is this framed in the context of (self regulating) micro markets, when billions of agents interact.
- Tom Defanti: ML is like training your dog without knowing how the dog works.
- Larry Smarr: Manage the exponential.
- Mike Norman: It is not about hardware, it is about the people!!!
- Inder & me: The kids of today are the decision makers tomorrow and have no feeling for classic CI.

Conclusion, Q&A

Need for Network to Data level experimental Infrastructure.

Europe’s own DTN infra, CC program, CI Ambitions

Data at scale.

P.S. I did not mention Quantum Compute & Networking; See:

- https://www.orau.gov/quantumnetworks2018/default.htm
- https://science.energy.gov/%7E/media/ascr/pdf/programdocuments/docs/2019/QNOS_Workshop_Final_Report.pdf
- https://delaat.net/qn

https://delaat.net/

This trip is supported by DL4LD and EPI projects.