ICT to support the transformation of Science in the Roaring Twenties

Cees de Laat

Systems and Networking Laboratory
University of Amsterdam
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**

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

“BIG” DATA

SDSC SAN DIEGO SUPERCOMPUTER CENTER

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

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

- 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
Fire Modeling Workflows in WIFIRE

Real-time sensors

PRP

Weather forecast

Monitoring & fire mapping

Fire perimeter

Landscape data

SDSC

UC San Diego
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

- Cosmic Microwave Background
- DNA Structure
- Periodic Table of the Elements
- Special Relativity

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

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

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 support network operations.
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 example:
    - biodiversity
Science DMZs for Science Applications

High-performance feature-rich science network ecosystem

Data Portal

Single Lab

Parallel Filesystem

HPC Facilities

Experiments

Experiment Data Archive

LHC Experiments

University Computing

DATA

DATA

DATA

DATA

Courtesy Eli Dart, ESnet

Petascale DTN Project
November 2017
L380 Data Set
Gigabits per second (min/avg/max), three transfers

ALCF DTN cluster
Globus endpoint: alcf#dtm_mira
Filesystem: /projects

43.0/50.0/56.3 Gbps
33.0/35.0/37.8 Gbps
44.1/46.8/48.4 Gbps

OLCF DTN cluster
Globus endpoint: olcf#dtm_atlas
Filesystem: atlas2

34.6/47.5/56.8 Gbps
29.9/33.1/35.5 Gbps
33.2/43.4/50.3 Gbps

NERSC DTN cluster
Globus endpoint: nersc#dtm
Filesystem: /project

41.0/42.2/43.9 Gbps
35.9/39.0/40.7 Gbps

OCSA DTN cluster
Globus endpoint: ncsa#BlueWaters
Filesystem: /scratch

21.2/22.6/24.5 Gbps
55.4/56.7/57.4 Gbps
26.7/34.7/39.9 Gbps

Data set: L380
Files: 19240
Directories: 211
Other files: 0
Total bytes: 6442781786482 (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 file
100 - 1K bytes: 53 files
1K - 10K bytes: 3170 files
10K - 100K bytes: 1560 files
100K - 1M bytes: 2817 files
1M - 10M bytes: 3901 files
10M - 100M bytes: 3880 files
100M - 1G bytes: 2295 files
1G - 10G bytes: 1647 files
10G - 100G bytes: 3 files

Courtesy Eli Dart, ESnet
eScience Middleware

+ Machine Learning + Reasoning + Scheduling + ...

Service Plane

Chromium  SAGE  MTP  OCCI  JSDL  CAGLX  GIR  OpenFlow  SDN / NSI  PerfSonar  ICMP  Cassandra  iRODS  Hadoop  Storm  WSRF  SensorML  INSPR
Networks of ScienceDMZ’s & SDX’s

Internet
Peer ISP’s

Supercomputing centers
(NCSA, ANL, LBNL)

ISP

SDX

NFV

Func-c1
Func-c2

ISP

SDX

NFV

Func-c4

ISP

SDX

DTN

Ownership/trust relation

client 1
client 2
client 3
client 4

... client n

Petabyte email service 😊

Contains a

DMZ

DTN

Petabyte email service 😊

J

DMZ

DMZ

DMZ

DMZ

Contains a
Superfacility Model for Productive, Reproducible Science

Interconnected facilities where data is acquired, stored, analyzed and served

User Community

Expertise

Computing and Data Facilities

Experimental Facilities

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?
Big Data Sharing use cases placed in airline context

- **Global Scale**
  - Aircraft Component Health Monitoring (Big) Data
    - NWO CIMPLO project
    - 3.5 FTE

- **National Scale**
  - Cargo Logistics Data
    - (C1) DaL4LoD
    - (C2) Secure scalable policy-enforced distributed data Processing (using blockchain)
    - NWO iShare project

- **City / regional Scale**
  - Cybersecurity Big Data
    - NWO COMMIT/SARNET project
    - 4.5 FTE
Data value creation monopolies

Create an equal playing field

Sound Market principles

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!)
AmDex
Layer 2 exchange service
Ethernet frames
Routers - Internet – ISP’s – Cloud
IP packet service
Data objects & methods
Data & Algorithms service

AmsIX
Application domain
Application domain
Application domain
Application domain
Application domain

FAIR / USE
IP / BGP
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
The Big Data Challenge

Doing Science

- Wisdom
- Knowledge to act
- Information
- Data a.o. from ESFRI’s

ICT to enable Science

- e-IRG
- Workflows Schedulers to act
- OWL
- XML, RDF, rSpec, SNMP, Java based, etc.

GRID/CLOUD

Doing Science — ICT to enable Science

ICT to enable Science — Doing Science
The Big Data Challenge

Doing Science

Data
a.o. from ESFRI's

Knowledge

Wisdom

Information

MAGIC DATA CARPET

curation - description - trust - security - policy – integrity

ICT to enable Science

Workflows
Schedulers

e-IRG

OWL

XML, RDF, rSpec, SNMP, Java based, etc.

Scientists live here!

Science App Store?
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
  - Dedicated 10 Gb/s optical links between all sites
- DAS-4: Clouds, diversity, green IT (2010)
  - Hardware virtualization, accelerators, energy measurements
- DAS-5: Harnessing diversity, data-explosion (June 2015)
  - Wide variety of accelerators, larger memories and disks

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

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.
Past & future ICT research infrastructures

- TEN34 / TEN155
- Geant testbed & JRA’s
- FIRE
- Grid5000 (FR)  
  - Was connected by LightPath around 2010!
- DAS1-5 (NL)
- 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
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.
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)

IEEE NetSoft paper [5]
- Efficiency of SDN mitigations [8]

SC15 demo/poster <delaat.net/sc16>
- Austin (TX)
- Sec-VirtNet paper [3]
- Computational Trust Model (SCTM) [7]

Laboratory: ExoGeni & PRP
- Fieldlab with KLM & CIENA
OSA-Optical Forum Conference paper [1]

SC16 INDIS workshop paper [2]
- TNC short paper [4]

SC16 demo/poster <delaat.net/sc16>
- Salt Lake City (UT)
- IEEE Sec-VirtNet 2016 paper [3]
- Removing Undesirable Flows [6]

SC16 demo/poster <delaat.net/sc16>
- Salt Lake City (UT)
- CoreFlow: Enriching Bro security events [9]

7. Paper: Ameneh Deljoo, Tom van Engers, Leon Gommans, Cees de Laat, “Social Computational Trust Model (SCTM) as framework to facilitate selection of partners”, Proceedings of 2018 IEEE/EMC Engineering the Network for Data-Intensive Science (INDIS), Dallas, TX, USA, 2018
EPI Project goals

“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?
Position in the Instituut

ICAi
- Amsterdam Machine Learning Lab
- Computer Vision
- Information and Language Proc. syst.
- Intelligent Sensory Information Systems

Artificial intelligence

IAS
- Computational Science and Engineering
- Complex Systems

Computational Science

Amdex
- Multi-Scale Networking Systems
- Parallel Computing Systems
- Complex Cyber Infrastructure
- Security by Design

Systems and Networking

ABC
- Strategic theme
- Existing units
- New investments full chairs

Informatics Institute
CONCLUSIONS

• Observations:
  – parallels energy world and internet developments
    • move to micromarkets
    • IOT alike security treats
  – trend: ML & AI replaces Visualization
  – 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

• Overall 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
  – Super-facility
  – 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.
AI forking off

Artificial Intelligence

NOW
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://delaat.net/qn
- https://delaat.net/
  - https://delaat.net/sarnet
  - https://delaat.net/dl4ld
  - https://delaat.net/epi

This trip is supported by SARNET, DL4LD and EPI projects.