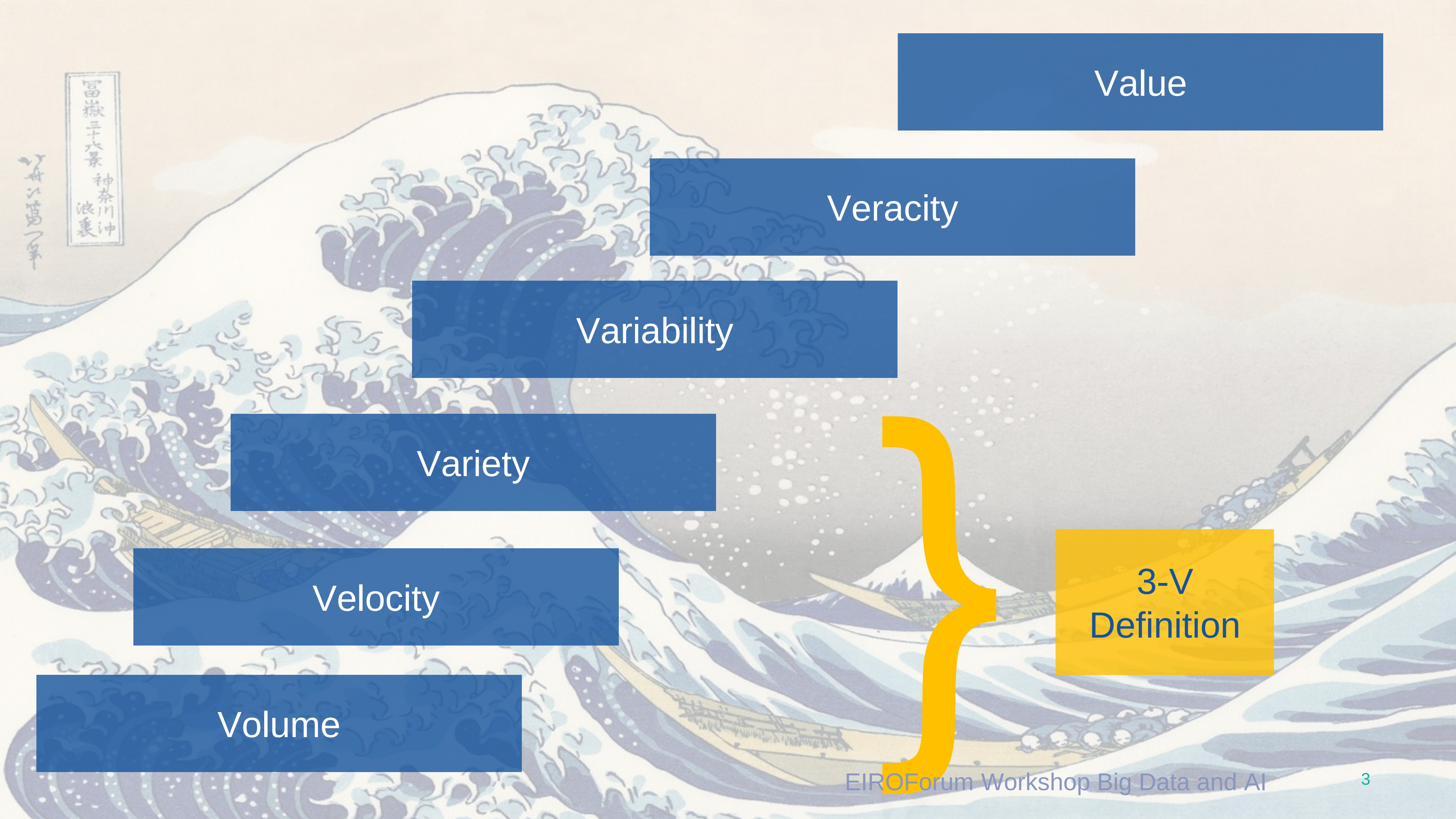
CERN

Big Data and AI for Data Analysis

26/10/2020

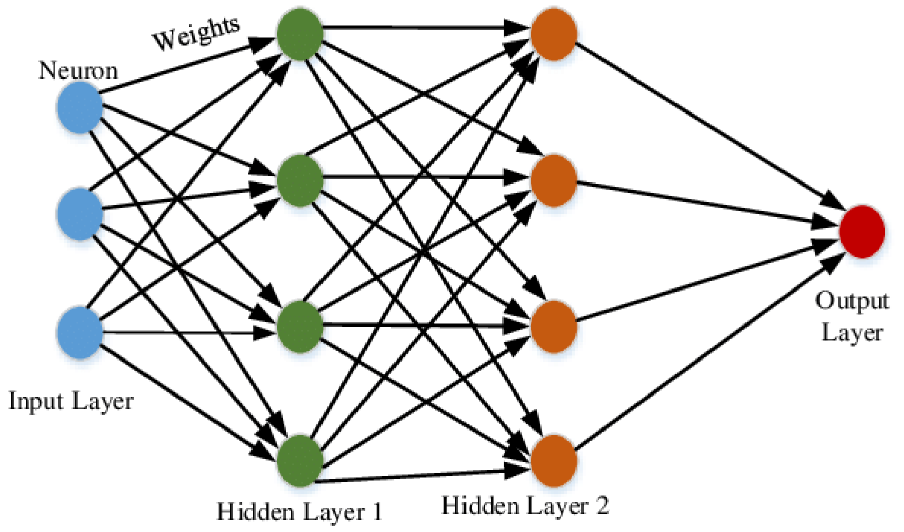
Alberto Di Meglio - CERN openlab Head, CERN IT Department
Sofia Vallecorsa - CERN openlab AI and Quantum Research, CERN IT Department
João Fernandes - EOSC ARCHIVER project coordinator, CERN IT Department



冨嶽三十六景
神奈川沖浪裏
Value
Veracity
Variability
Variety
Velocity
Volume
3-V Definition

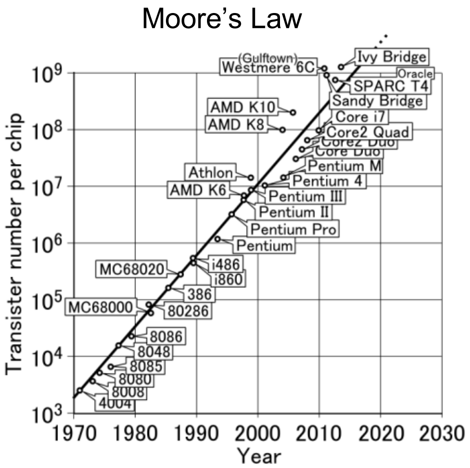
Artificial Intelligence (ML/DL)

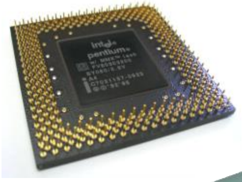
The coincidental combination of the availability of increasing amounts of (noisy!) data, more efficient algorithms, and faster hardware makes Machine/Deep Learning approaches more and more appealing



But… how do we implement it at large scale?

New Computing Platforms





1995

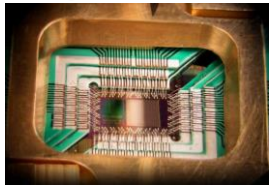
2000

2010

2014

2015

2019



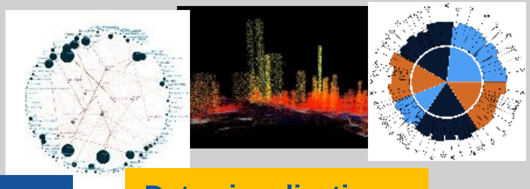
Radically new computing platforms are rapidly moving from pure computer science to realistic devices, e.g. **Neuromorphic Computing** and **Quantum Computing**

A **Quantum Computing Initiative** has been launched in 2020 as a long-term investigation activity

What does it Really Mean?

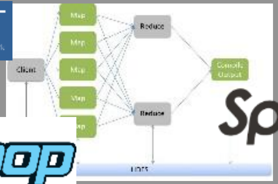
Information

Software

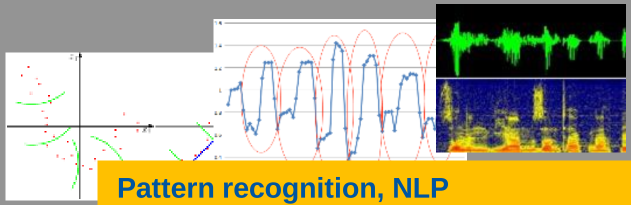


Data visualization, representation

Applications designed for “big data”



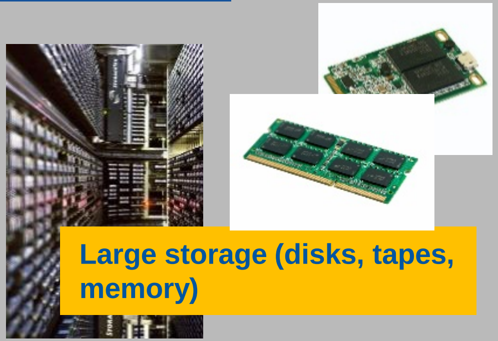
Data analysis and analytics platforms



Pattern recognition, NLP machine learning, predictive analytics

Value
Veracity
Variability
Variety
Volume
Velocity

Platform



Large storage (disks, tapes, memory)

Fast data acquisition systems

Flexible networks

Distributed Computing and Data Grids, Clouds, HPC, Crowd Computing

Infrastructure

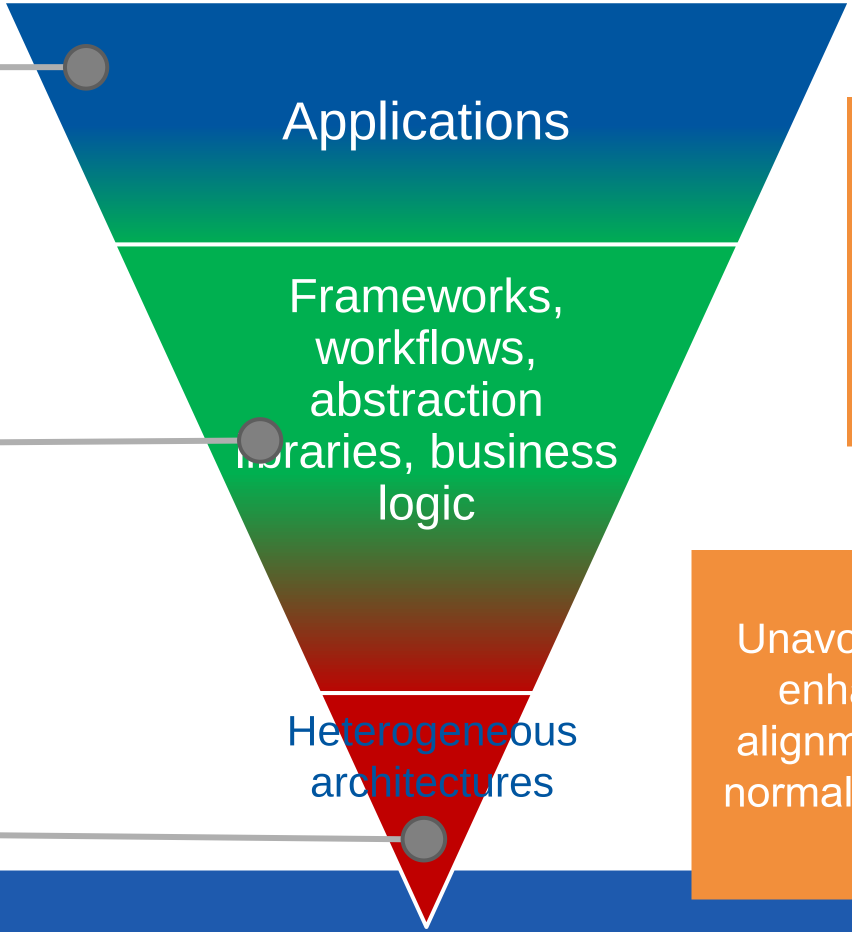
As Easy as ABCDE:

AI, Big data, Clouds, Direction, Education

User-domain apps and software

This is where the main effort should go, what the users need, high-level services abstracted from the underlying infrastructures and where future activities should focus

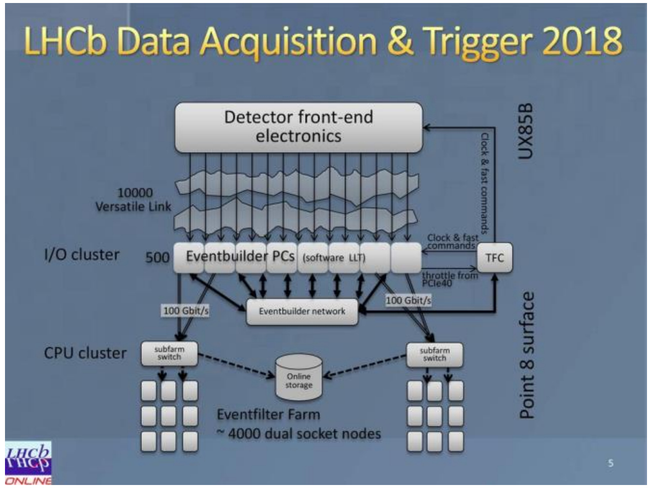
This is the general distributed infrastructure, sites, clouds, HPC Centres, in most cases they "just exist", but we need to learn how to use them cost-effectively



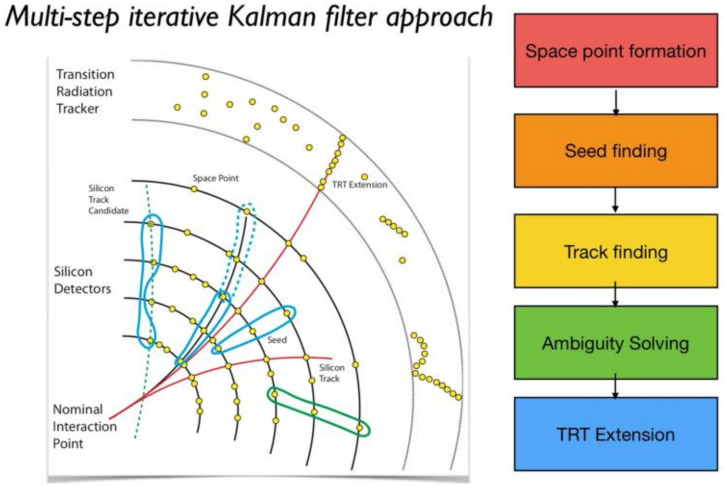
Education and Training
Targeted to both users and service providers
Needed to increase and evolve skills, prepare users for new services, and keep our people "market-worthy"

Governance & Compliance
Unavoidable need to address privacy & trust enhancing technologies that can ensure alignment with regulation as part of the "new normal" on technological requirements across all layers

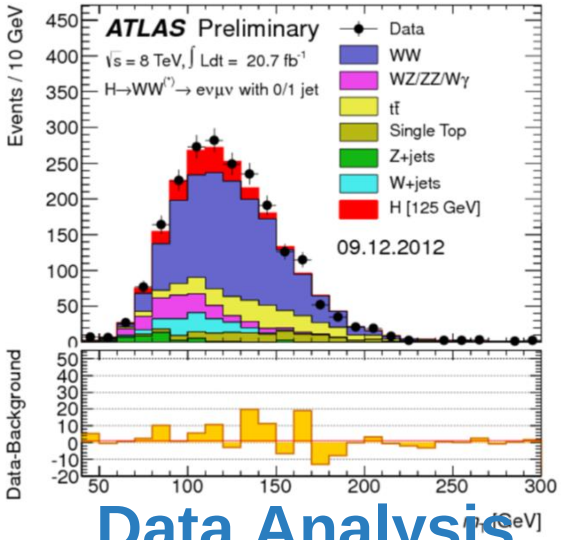
Typical LHC Experiments Workloads



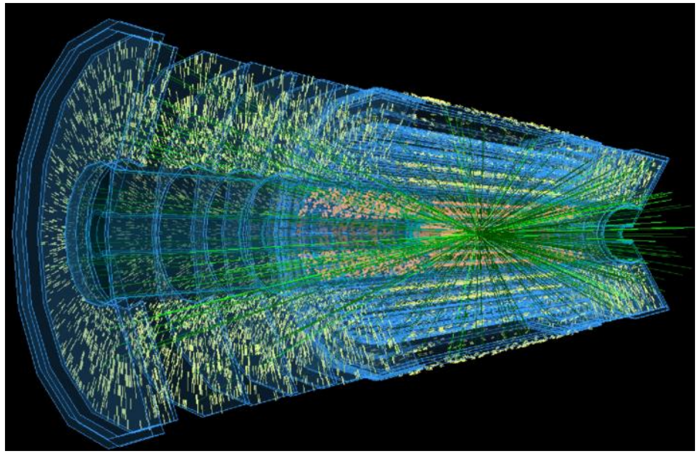
Data Acquisition



Track Reconstruction



Data Analysis

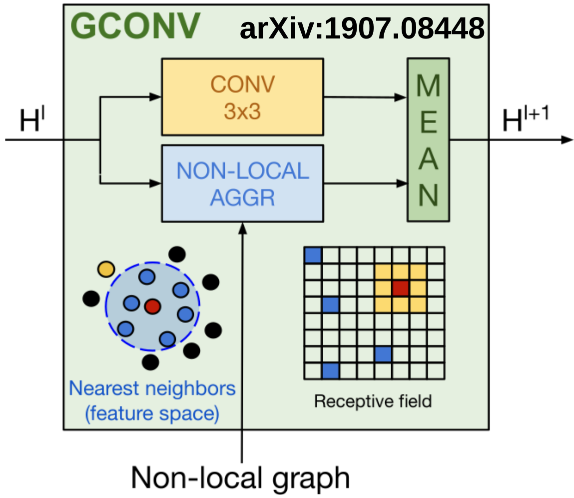


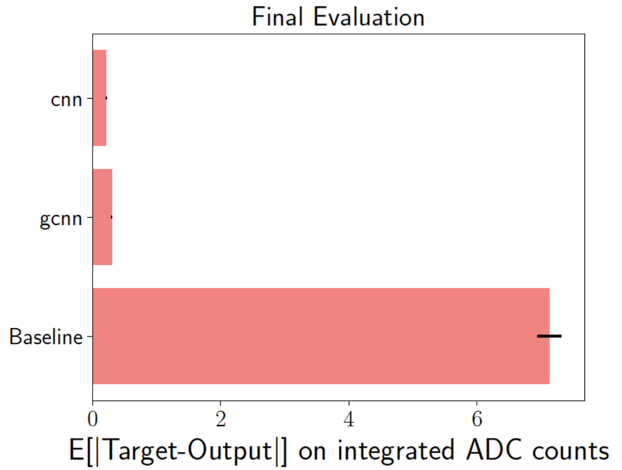
Simulation

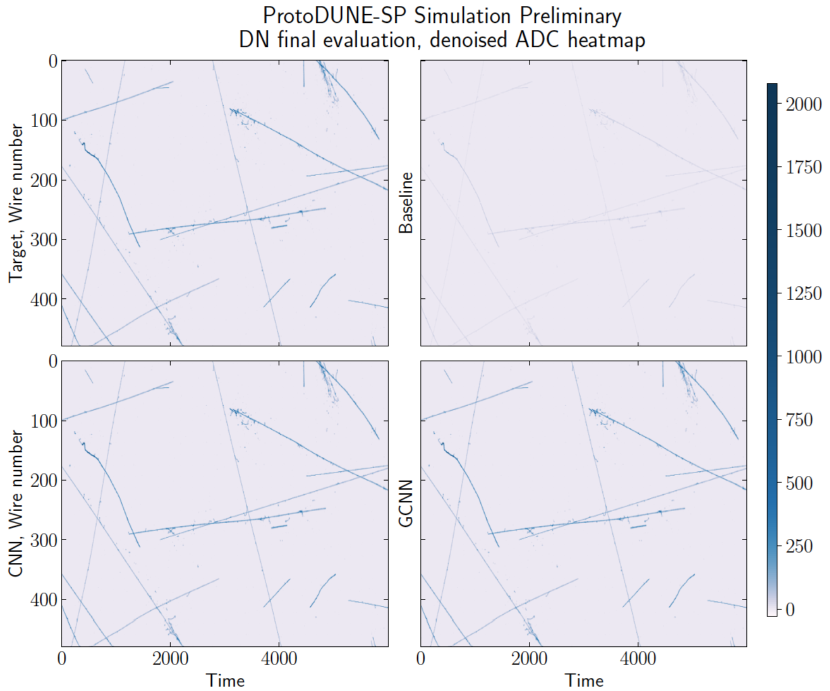
Data Acquisition

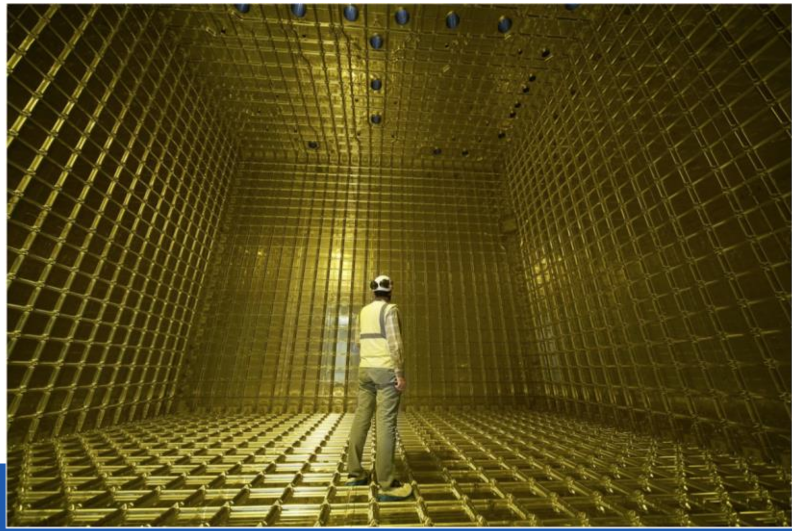
Deep graph-convolution NN to select and denoise raw detector data (Dune)

custom architecture







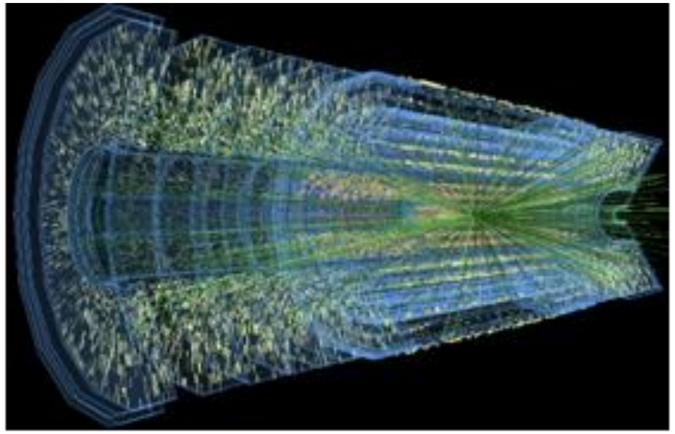


Track Reconstruction

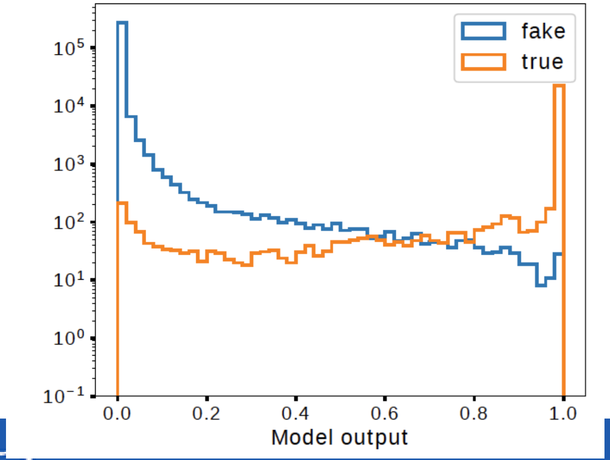
Exa.TrkX project introduces **Graph Neural Networks** for particle trajectory reconstruction

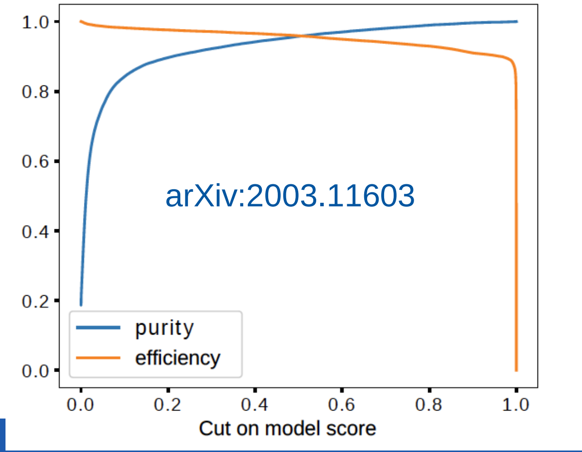
- Data as a **graph of connected hits**
- Connect plausibly-related hits using geometric constraints

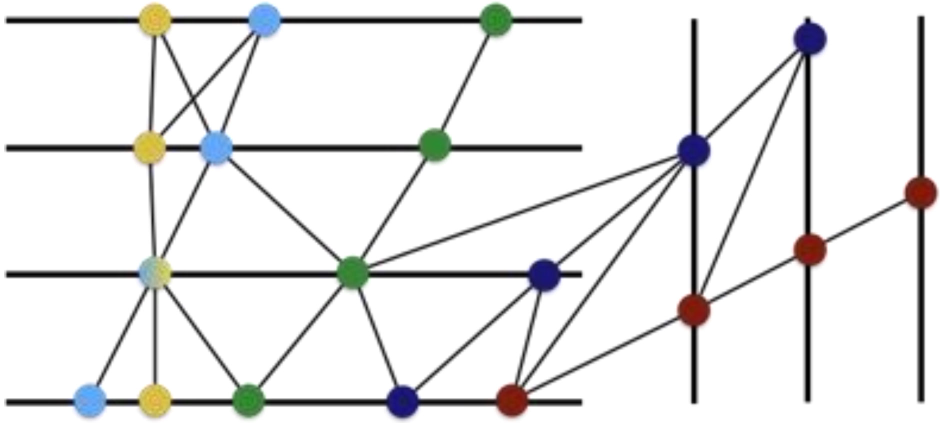
Use a cascade of **Input**, **Edge** and **Node** GNN Networks



https://exatrkx.github.io/







Event Classification

Simple, portable and scalable ML platform running on Kubernetes

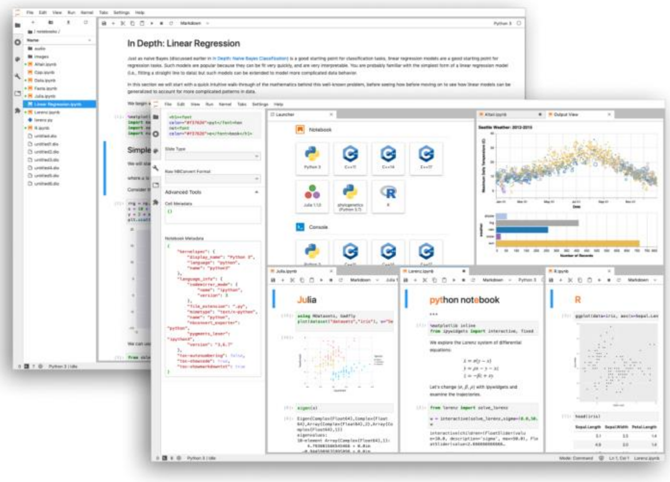
- Auto Scaling, GPU and Accelerators, Cloud Bursting

Support entire ML lifecycle

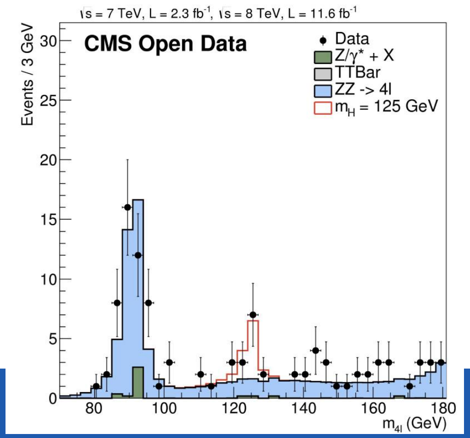
- Development, Training, Inference/Model Serving
- Notebooks, ML Pipelines, Hyper-parameter Optimization
- Tensorflow, PyTorch, scikit-learn, MXNet, MPI, …

Use Cases: 3DGAN/Fast Simulation, CNNs DUNE, CMS HLT

H→4l re-discovery from CMS Open Data **in 4 minutes using 25k core on Google cloud:** cloud native approach to ML-based analysis for HEP



L.Heinrich, R. Rocha, Keynote at KubeCon EU 2019



Event Classification

H→4l re-discovery from CMS Open Data **in 4 minutes using 25k core on Google cloud:** cloud native approach to ML-based analysis for HEP

L.Heinrich, R. Rocha, Keynote at KubeCon EU 2019

Google Cloud Storage

70 TB Dataset

Cluster on GKE

Max **25000 Cores**

Single Region, 3 Zones

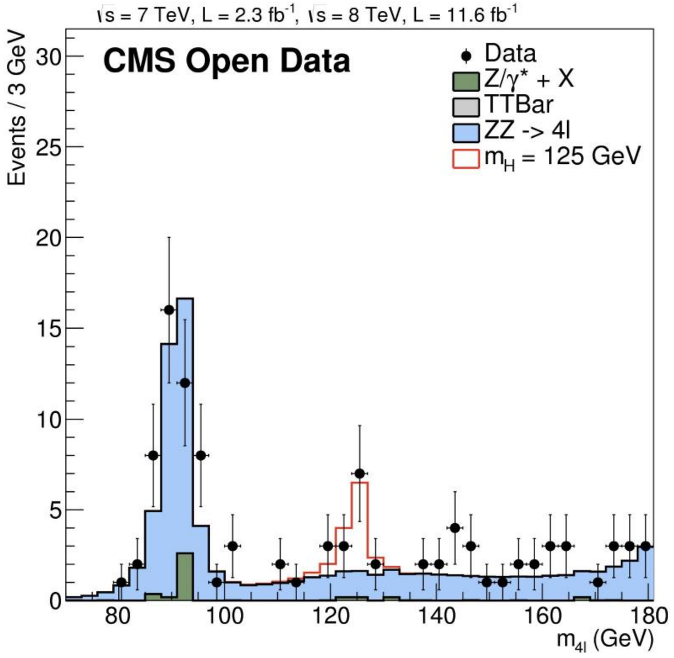
25000 **Kubernetes Jobs**

Job Results

jupyter

Interactive Visualization

Aggregation



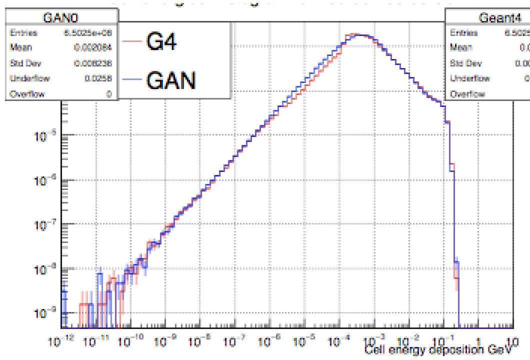
Simulation

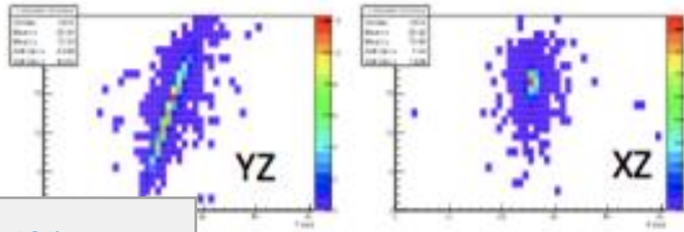
Simulation is a **major workload** in terms of computing resources.

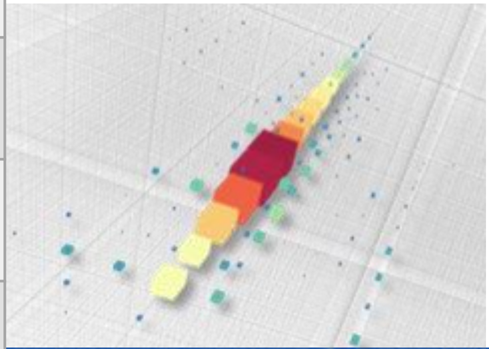
- With High Luminosity LHC we expect a x100 increase in simulation need

Investigate **Deep Generative Models** to replace Monte Carlo

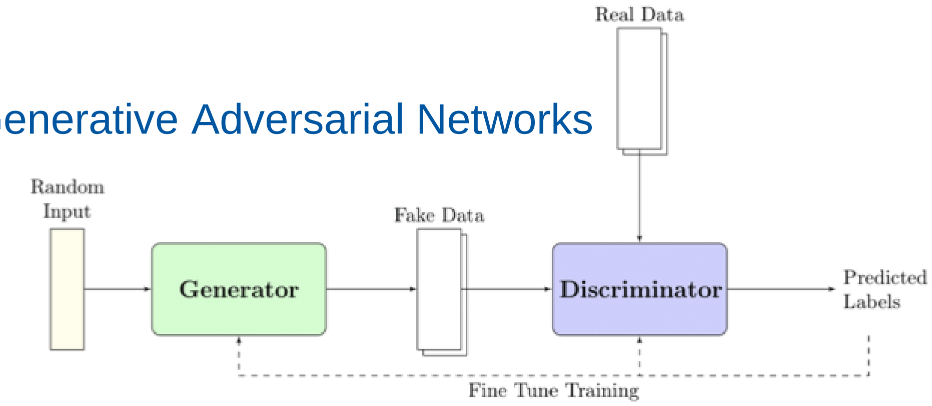
G.Khattak, ICMLA2019







Generative Adversarial Networks



Time to create an electron shower		
Method	Machine	Time/particle (msec)
MC Simulation	Intel Xeon Platinum 8180	17000
3DGAN (batch size 128)	**Intel Xeon Platinum 8160 (TF 1.8)**	**1**

Quantum Machine Learning

Quantum linear algebra **is generally faster** than classical counterpart

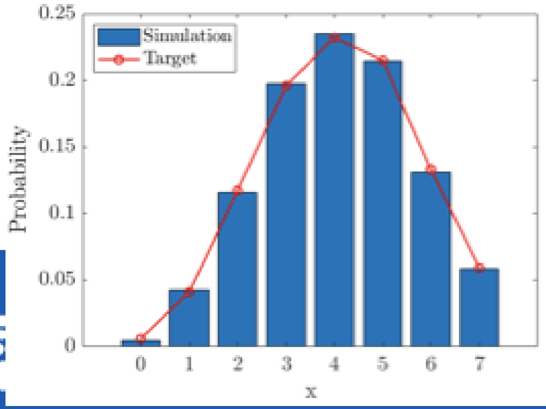
Some standard ML techniques estimate the **ground state of Hamiltonians**

ML algorithms have some **tolerance to errors**

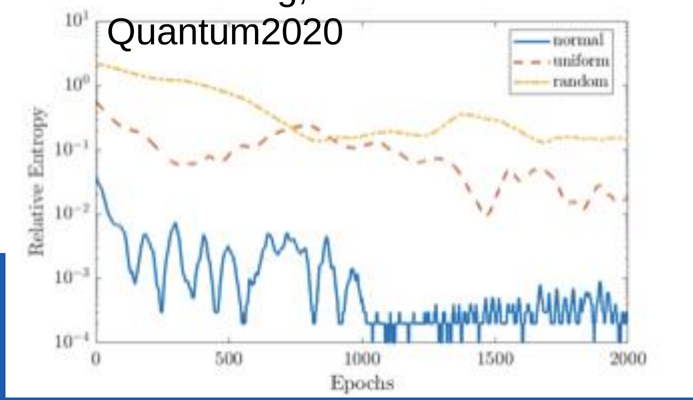
Specific **quantum techniques** can be exploited to bring further improvement

Partnership with ESA

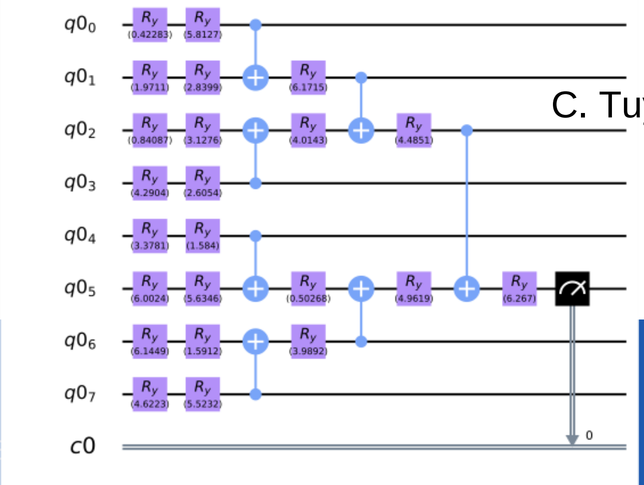
- Investigate quantum Generative Models Applications to applications to Earth Observation



S.Y. Chang, IOP Quantum2020



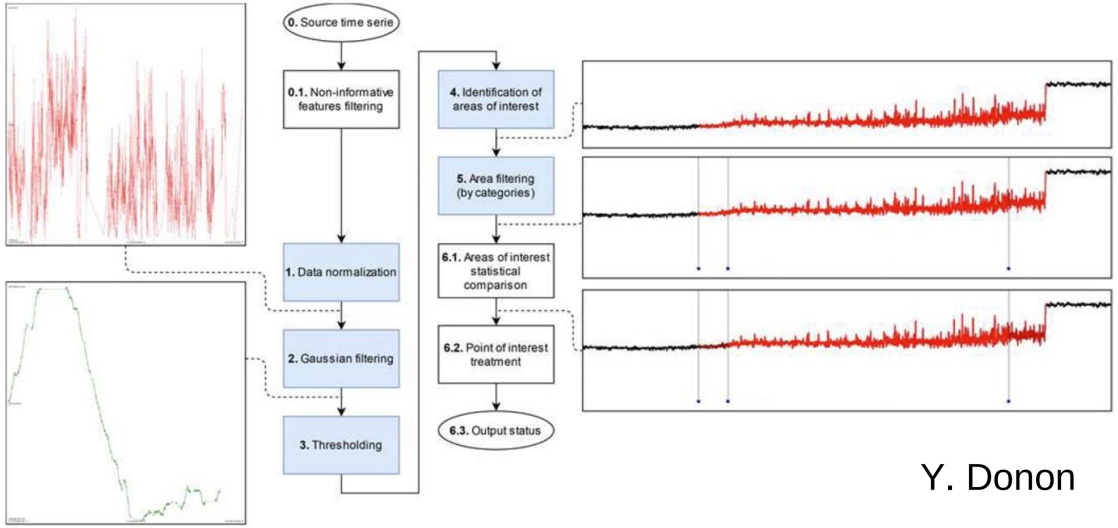
C. Tuysk, CDT020



Anomaly Detection, Preventive Maintenance, Control Systems

Many use cases and ongoing projects

- Anomaly identification in noisy data using SNiF (Series with Noise Featuring), tests with LINAC4 at CERN, collaborations on medical and industrial LINACs
- CNN-based image analysis for anomaly detection in industrial components
- Edge-computing, IoT, block chains for distributed control systems
- Beam stability and quality control for particle injectors

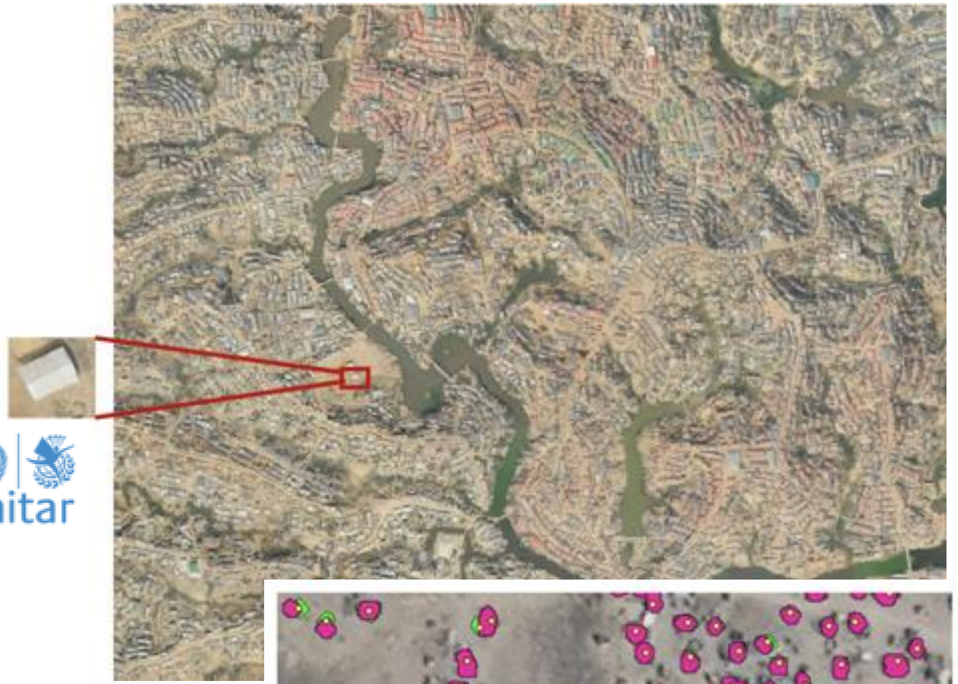


Y. Donon

AI for Earth Observation

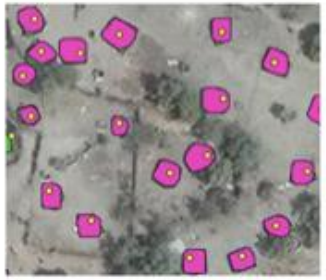
Automatic scan of high-resolution satellite images for **disaster relief**

High precision is required

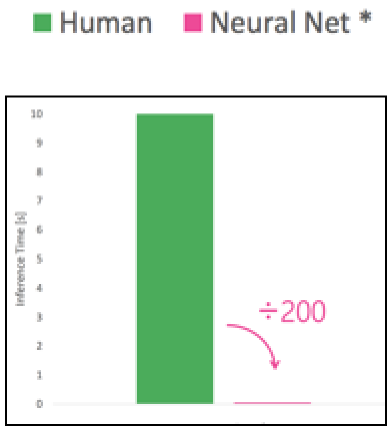


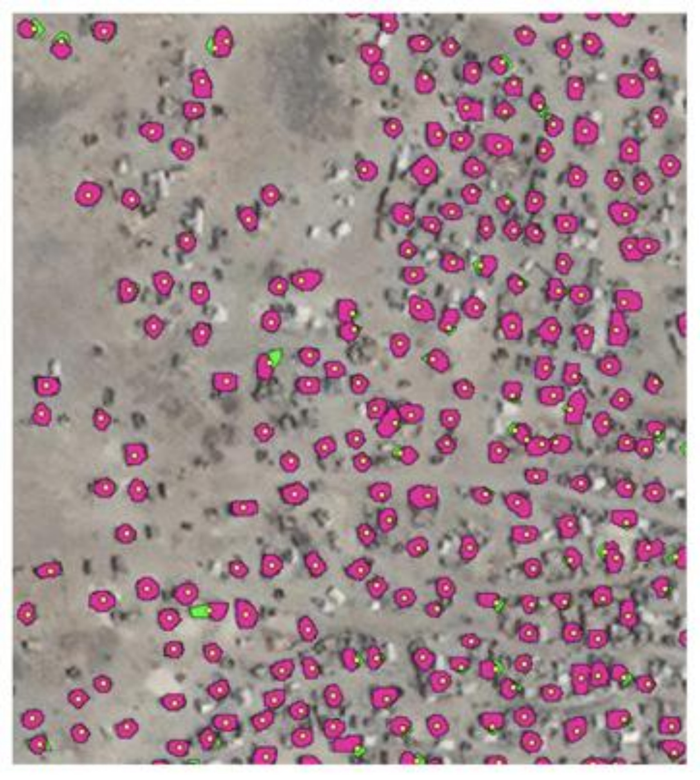
Detectron Framework (FacebookAI)

Retrain & encode point data cleverly



Unosat Adapted model

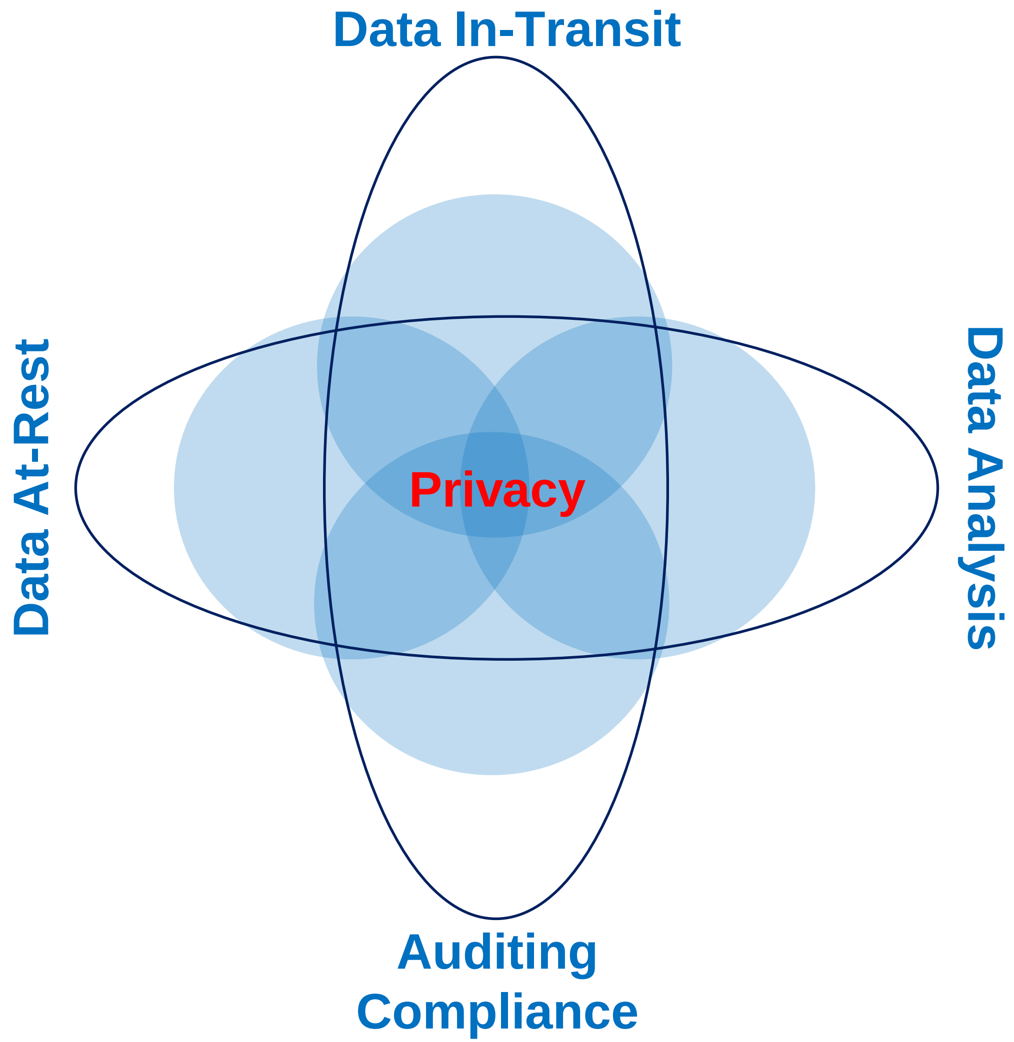




Transfer learning from Region-based CNN

Average precision is 82%, 200x speedup

Privacy-Preserving Analysis for Medical Research



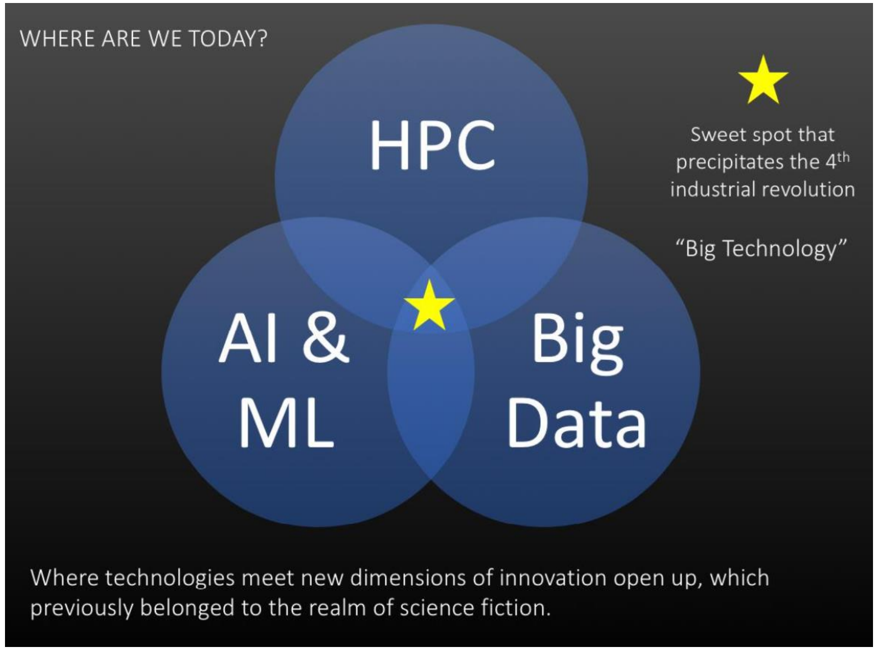
The possible risks of large-scale use of data must be managed

Use cases at CERN openlab based on methods developed for HEP

- Parkinson's detection from wearable devices using different statistical/ML/DL methods (event classification)
- Explicable AI for phenotype/genotype ("deconvolutional" Neural Networks)
- Image classification and segmentation for neurological diseases research using homomorphic encryption (noise-resilient networks for DAQ, radiation-hardening)
- Block chains and other transaction ledgers (security, but also analysis reproducibility)
- Applications of classic and quantum key distribution methods

Changing Business Models

Transparent integration of commercial cloud services, with default data governance: in DMPs, considering procurement cycles & research grant periods



Promote choice: ecosystem for innovation

Stay mainstream by adopting internationally recognized standards

Create a sustainable level playing field:
same requirements for commercial and not-for-profit providers

European Infrastructures for Science: EOSC

Framework to make access to scientific data widely available for public/private sectors across domains

Key point on data preservation & reuse

- ensure a continuum after
  a mission/experiment finishes

Create the conditions to develop a sovereign European data shari environment (e.g. explore links with GAIA-X)

Collaboration and joint initiatives

Joint investigation across distributed heterogeneous platforms are critical for the future of research

- CERN and ESA announced in September a joint partnership on Quantum and AI research for Earth Observation and Physics
- CERN and ESRF are discussing about collaborations on ML/DL for accelerators operations and maintenance
- CERN, SKA, PRACE, GEANT signed an agreement to work on common access models for HPC
- CERN and EMBL have a long history of collaboration on Grid and Cloud infrastructures

Parting messages

- Tipping point for large-scale research
- ABC computing offers unprecedented opportunities
- Distributed platforms and services must become foundational
- Collaborations across Europe and internationally are critical
- All domains of science, technology, industry, society can benefit
- Good governance, trust, reliability, explainability must be built-in

Thanks!

alberto.di.meglio@cern.ch
@AlbertoDiMeglio