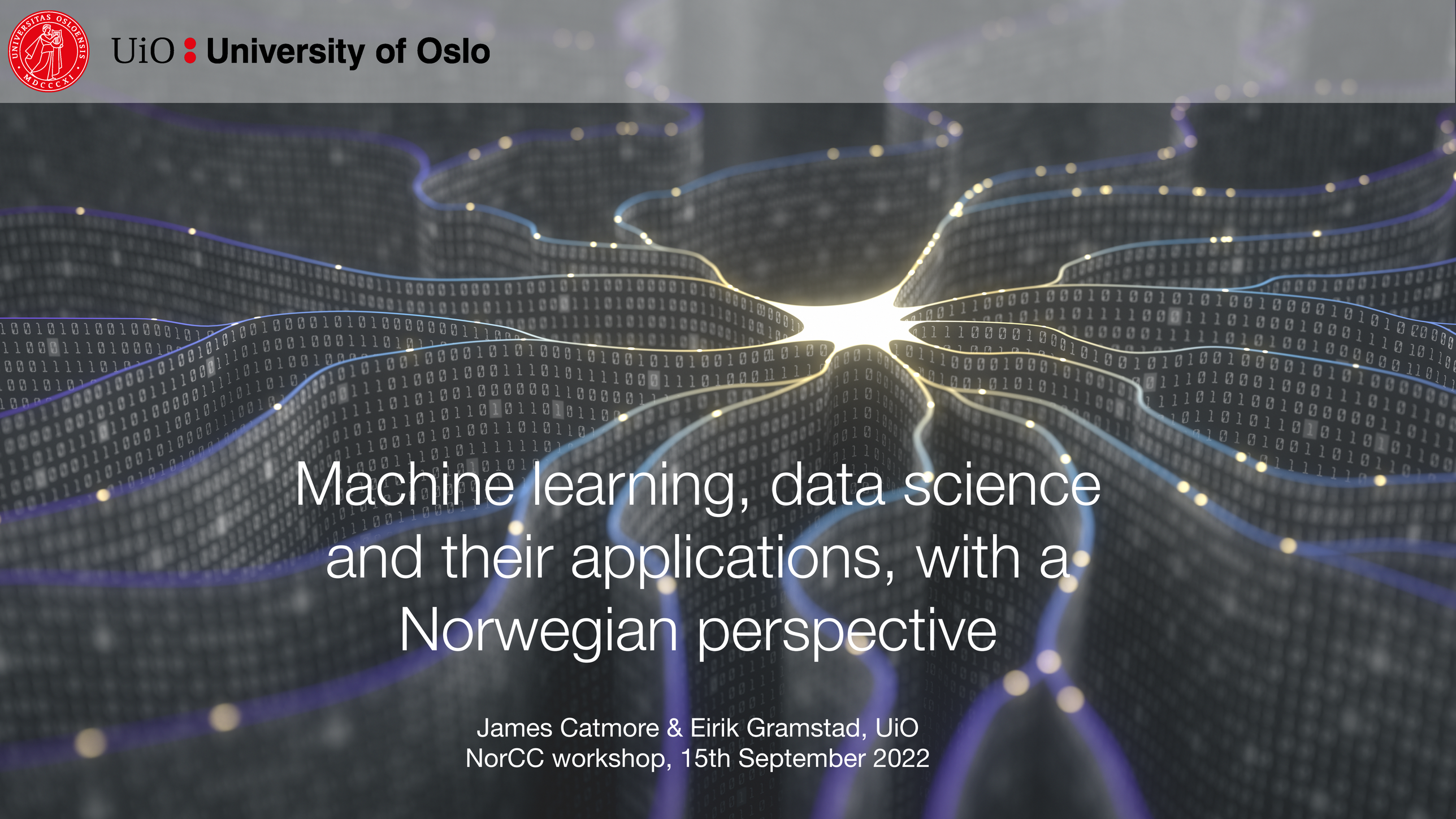
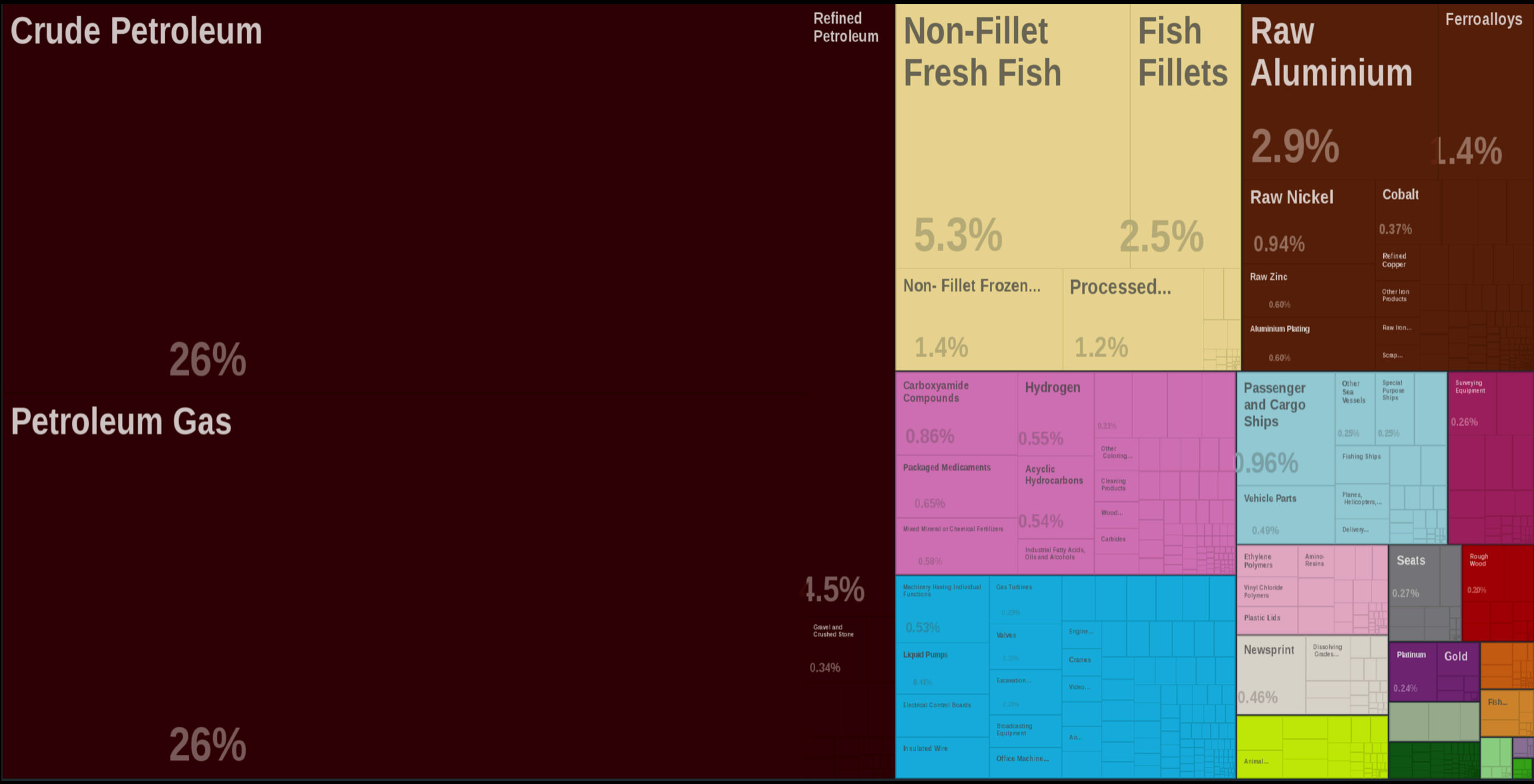

UiO : University of Oslo
Machine learning, data science and their applications, with a Norwegian perspective
James Catmore & Eirik Gramstad, UiO
NorCC workshop, 15th September 2022

Overview

- The Norwegian economy
- How can the wider world help HEP and how can HEP help the wider world?
  - Data science and statistical analysis
  - Machine learning
  - Computing and software
- Transferable skills - life after HEP
- Opportunities for collaboration beyond HEP



The Norwegian economy

- Norway’s economy is dominated by the petroleum industry (>50% of exports)
- All of this has to go by 2050 as part of the transition to net zero greenhouse gas emissions
  - Norway’s main gas customers have similar targets
- This is a huge challenge and a huge opportunity
  - Massive transition to renewable electricity to replace petroleum exports
  - Diversification of the economy away from energy exports
  - Electrification of other leading industries (shipping, metals/chemicals, food)
- Data science and machine learning will have a major role in this transition
  - Norway is a highly digitised country with a well educated population → well placed to address the challenge of the coming years
- Most services are provided digitally in Norway so the public and private sectors are awash with data to be exploited
- People with HEP training can make a huge contribution in these areas!

Data science

HEP data is unusual

- A large part of HEP data analysis is what is now called “data science”
  - Data collection, triggering, reconstruction, feature extraction, cutting, projecting, visualisation, statistical analysis
- Although our reconstructed data is highly structured, each “feature” (variable) may have any number of instances in each “example” (event)
  - There may be six jets in one event and one jet in the next
- These “ragged” structures are highly unusual in the wider world
- Those with a HEP background have skills which are widely sought in industry

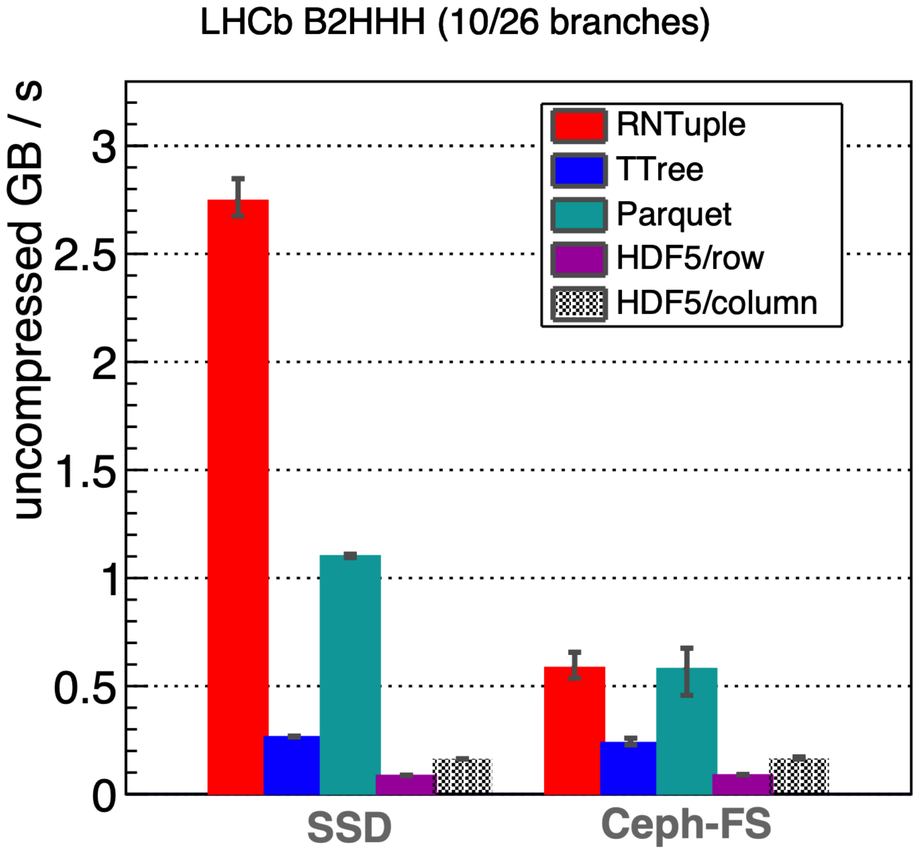
Var 1	Var 2	Var 3
#	#	#
#	#	#
#	#	#
#	#	#

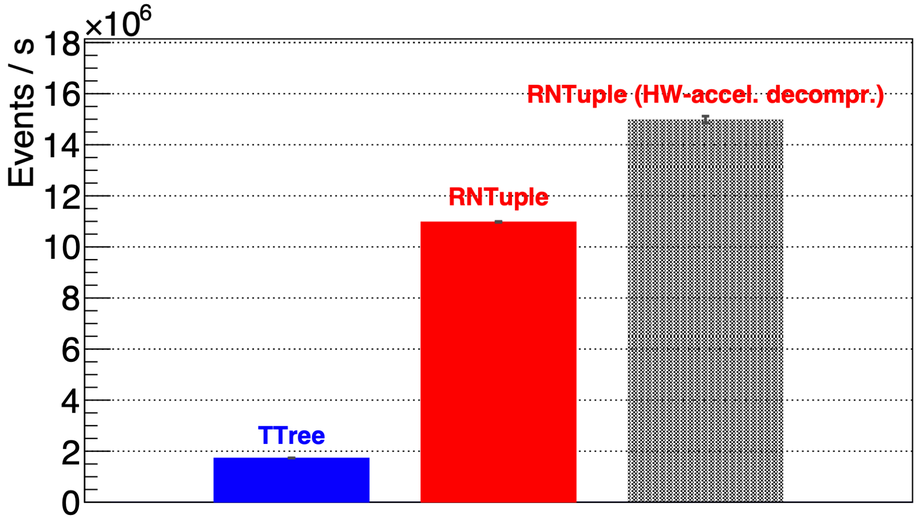
Var 1	Var 2	Var 3
#, #	#, #	#
#, #, #, #	#	#, #, #, #
#	#,#,#	#,#,#,#,#,#
#,#,#,#,#,#	#	#

Data science

HEP data science and statistical tools are world leading

- The ROOT framework and HEP statistical packages are world leading
  - ROOT and its data structures have exceptional I/O performance and memory management
  - It is designed to handle ragged data
  - ROOT7 will introduce significant further developments including RDataFrame, RNTuple and GPU support
- Data science tools in the wider world (Pandas, Numpy etc) are less well adapted to ragged data and are often less performant than ROOT
  - But they are generally better documented and more readily interfaced with machine learning tools (see next slide)…
  - Also more popular with students…
  - This has led to attempts to better the two “worlds” with tools such as AwkwardArray as well as massive efforts from the ROOT team to fully integrate with Python

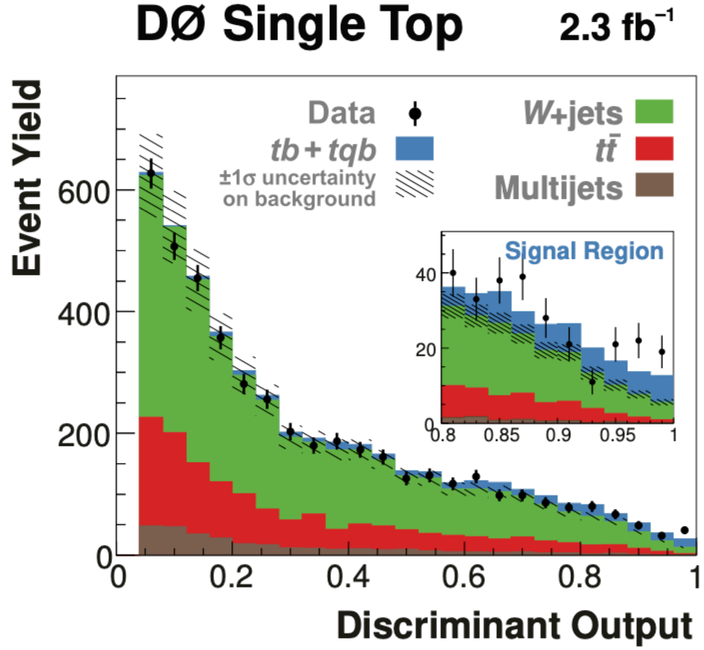


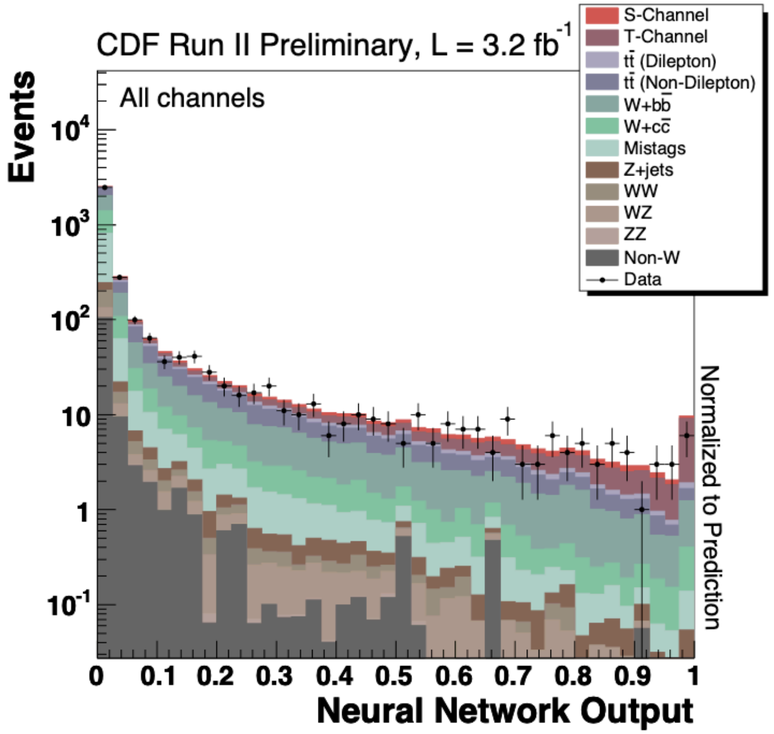


Machine learning

What HEP has contributed

- HEP was an early adopter of machine learning (we called it “multi variate analysis”)
  - In use since the LEP/TeVatron era; an early fundamental contribution was the discovery of single top production
  - Now used widely and increasingly for many applications in HEP
- HEP has inspired several developments now used in the wider world
  - Regressing cyclic observables and transformed observables (from modelling phi and eta)
  - Max-likelihood fits to classifier outputs (and related statistical interpretation)
  - The HiggsML challenge in 2014
  - The popular Keras framework boasts on their front page that “Keras is used at the LHC”



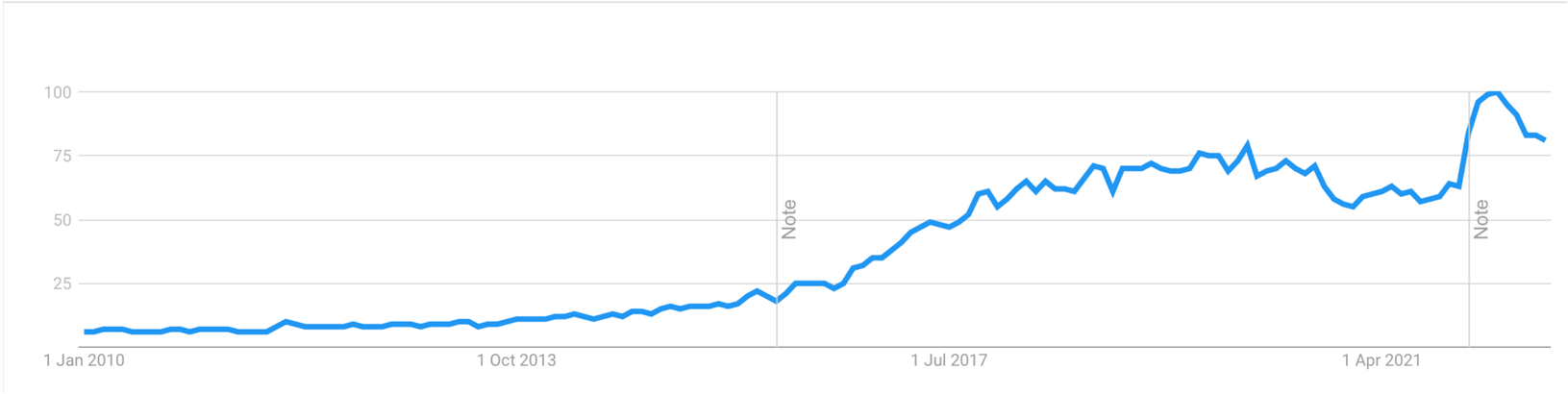


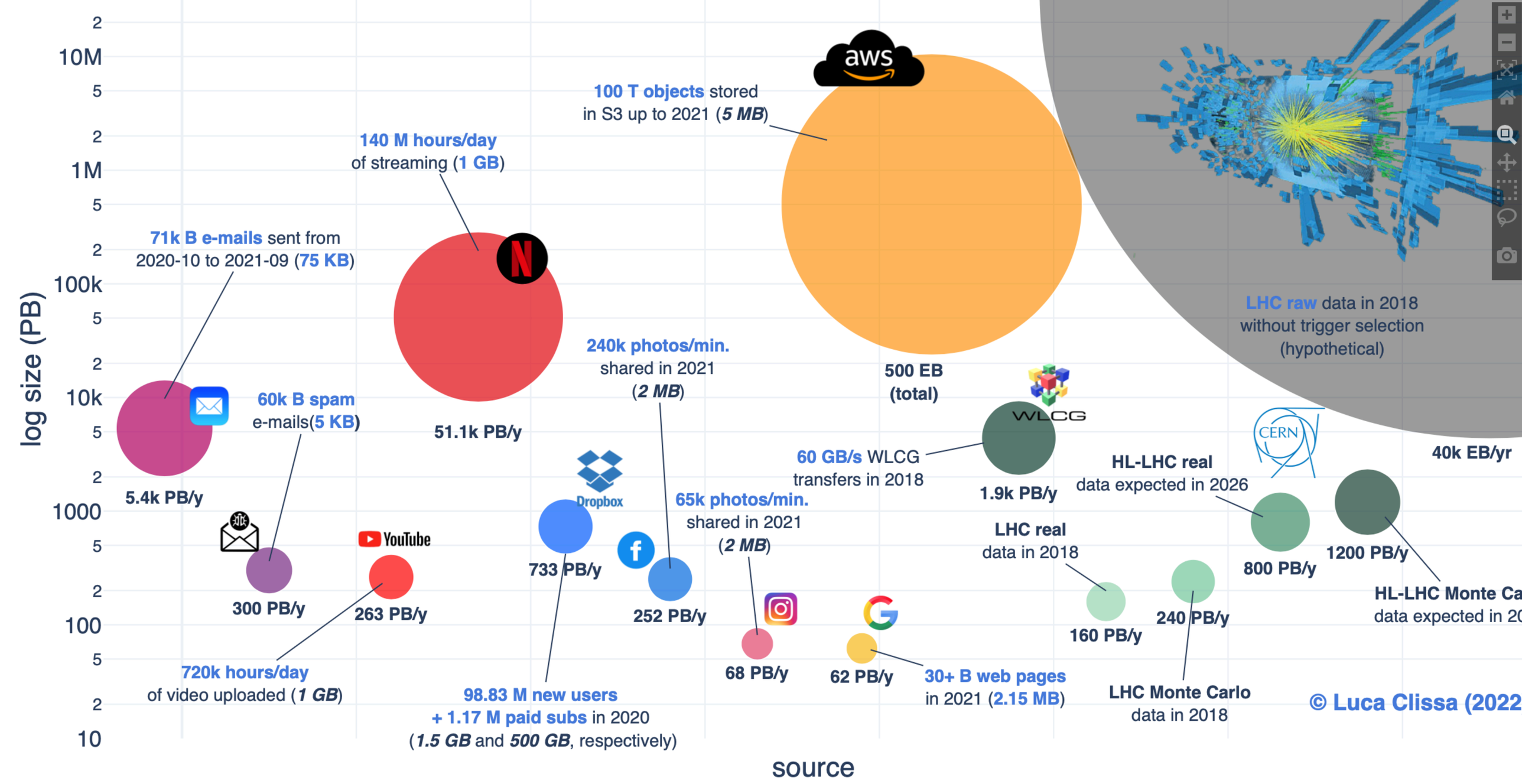
Machine learning

What HEP can learn

- The growth of computing power, the rise of huge technology companies and the availability of training data (“big data”) has led to an explosion in the use of ML, so we are no longer leaders in this area
  - We are able to benefit from new open-source tools developed by the wider community (TensorFlow/Keras, SKL, etc)
- Significant opportunity for collaboration with non-HEP experts
  - Introduction of new techniques: computer vision; anomaly detection; generative methods; …

Searches for “machine learning” on Google since 2010





https://datapane.com/u/lclissa/reports/dkjK28A/big-data-2021/
https://towardsdatascience.com/how-big-are-big-data-in-2021-6dc09aff5ced

Computing and software

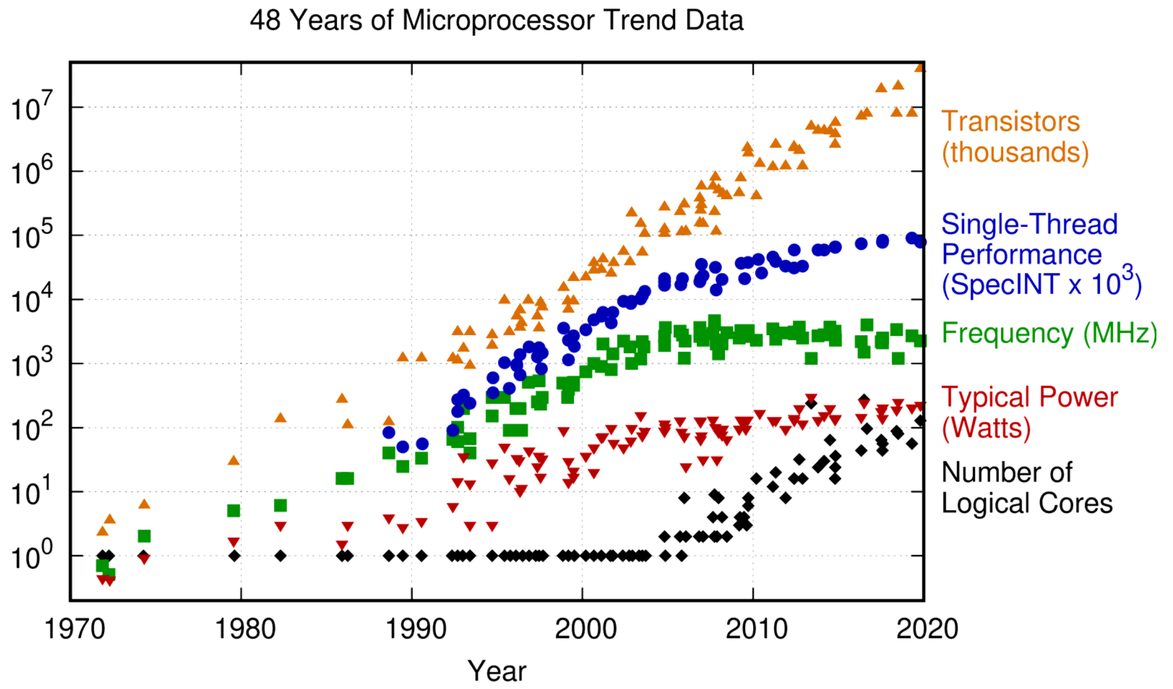
Grid computing

- In terms of data storage and processing volumes, HEP isn’t negligible
- The grid computing model used by HEP is robust and flexible and built to extend collaboration to other research fields
- The distributed data management system used by ATLAS and CMS (Rucio) is world leading and can have applications beyond HEP
- Personnel from NorCC have leading roles in these areas
- Upcoming challenges → common with other areas?
  - HL-LHC data volumes
  - Price of electricity
  - Integration of heterogeneous architectures (HPC, GPU?)
  - Use of commercial “cloud” services (collaboration at the experiment level with several companies including Google and Amazon Cloud)

Computing and software

Software development and optimisation

- Current HEP software stacks were developed in the early 2000s when hardware performance was continually increasing and the price falling
- Not the case any more - memory, CPU and disk will be critical for run 4
  - As will electricity consumption
- Making software more efficient is a major task and one in which few HEP people are expert
- Challenge is compounded by imperative to develop GPU-compatible software
- There is significant collaboration at the experiment and CERN level with Intel, Nvidia etc
  - Beta testing of new hardware/software, tutorials etc



Life after HEP

- Many ex-HEP people have gone on to contribute to the Norwegian economy/society in different roles related to data science and machine learning
- Recent example include
  - Law enforcement and defence
  - Personal finance
  - Shipping & engineering
  - Agribusiness
  - IT support
  - Non-HEP academic research foundations
- Recently we have also had several reverse cases, with people *returning* to HEP after a spell outside

Transferable skills

“Hard” skills

- Based on the comments from some of the people who have moved onto job outside, there are both “hard” and “soft” skills that transfer well from HEP to the wider world
- “Hard” skills
  - Management of large data volumes
  - Knowledge of machine learning techniques
  - Rigorous statistical analysis of large data volumes, especially related to the emphasis on evaluation of uncertainties
    - “*The emphasis that HEP puts on understanding and modeling of uncertainties, and to be conscious of possible biases in the data and models, gives us an advantage for building robust statistical models, e.g. in machine learning, and to understand the significance and limitations of their results.*”
  - Python data analysis software
  - Working with a variety of computing environments/languages/scripts and large code bases

Transferable skills

“Soft” skills

- Based on the comments from some of the people who have moved onto job outside, there are both “hard” and “soft” skills that transfer well from HEP to the wider world
- “Soft” skills
  - Analytical thinking, identifying critical elements of a problem and structuring the approach to solving it
  - Clearly presenting complex ideas and work to colleagues
  - Collaborating effectively in large groups, sometimes remotely
  - Having a background in fundamental research earns respect with colleagues and customers

Opportunities for external collaboration

and an obstacle…

- UiO dScience – Centre for Computational and Data Science
  - Centre collecting all data science and AI activities across the faculty → we need to make better connections with this
- NORA - Norwegian Artificial Intelligence Research Consortium
  - For PhD candidates across Norway working on AI linked topics, NORA arrange two important activities in November
    - Nordic AI Meet, 14-15 November (https://nordicaimeet.com/)
    - Research School Annual Conference on 16 November :https://www.nora.ai/research-school/symposium.html
- HVL: master and PhD students are technically in computer science or computer engineering programs → strong interaction with informatics/CS
- Possible collaboration with certain French institutes on ML, where there is significant expertise
- CICERO: climate change centre
- Industry - SINTEF, Simula, companies such as Inmeta, energy, shipping etc?
- Obstacle: sharing ATLAS/ALICE experimental data (& even simulation) is impossible without a formal STA