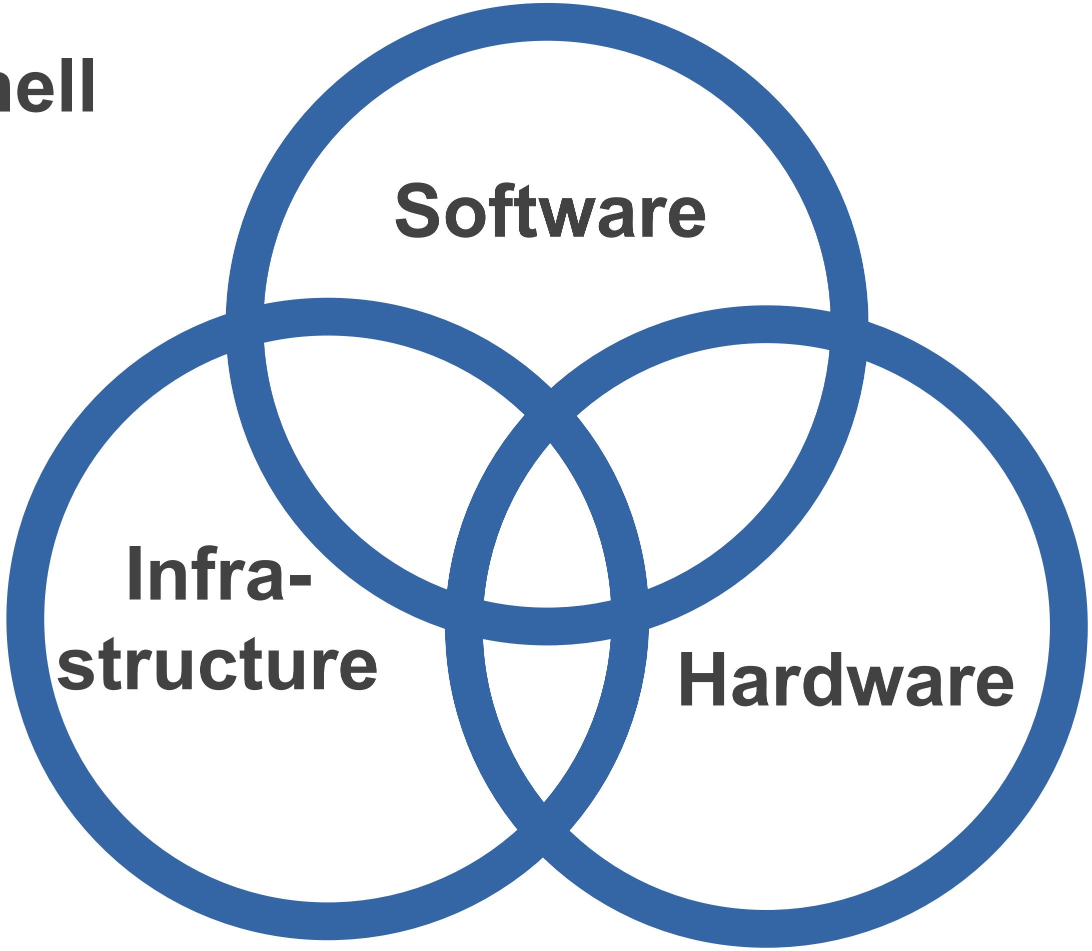
Sustainability aspects

Caterina Doglioni (U of Manchester) and Kristin Lohwasser (U of Sheffield)
17th September 2025, Workshop on Artificial Intelligence (AI)/Machine Learning (ML) and Emerging Technologies in particle physics

Computing in a nutshell

”Classical” sustainability (e.g. *FAIR* where R=Reusable, requires training…)



In this contribution we will focus on **environmental sustainability**

Why care about sustainability in computing?

- Climate change is real: Extreme weather events have become more frequent

- Paris agreement is legally binding → Will require changes to accomplish NetZero

nion Sport Culture Lifestyle

The Guardian

UK

ate crisis Middle East Ukraine Football Newsletters Business Environment UK politics Science Tech Global development Obituaries

Europe's summer of extreme weather caused €43bn of short-term losses, analysis finds

Greatest damage from heat, drought and flooding done in Cyprus, Greece, Malta and Bulgaria

Advertisement

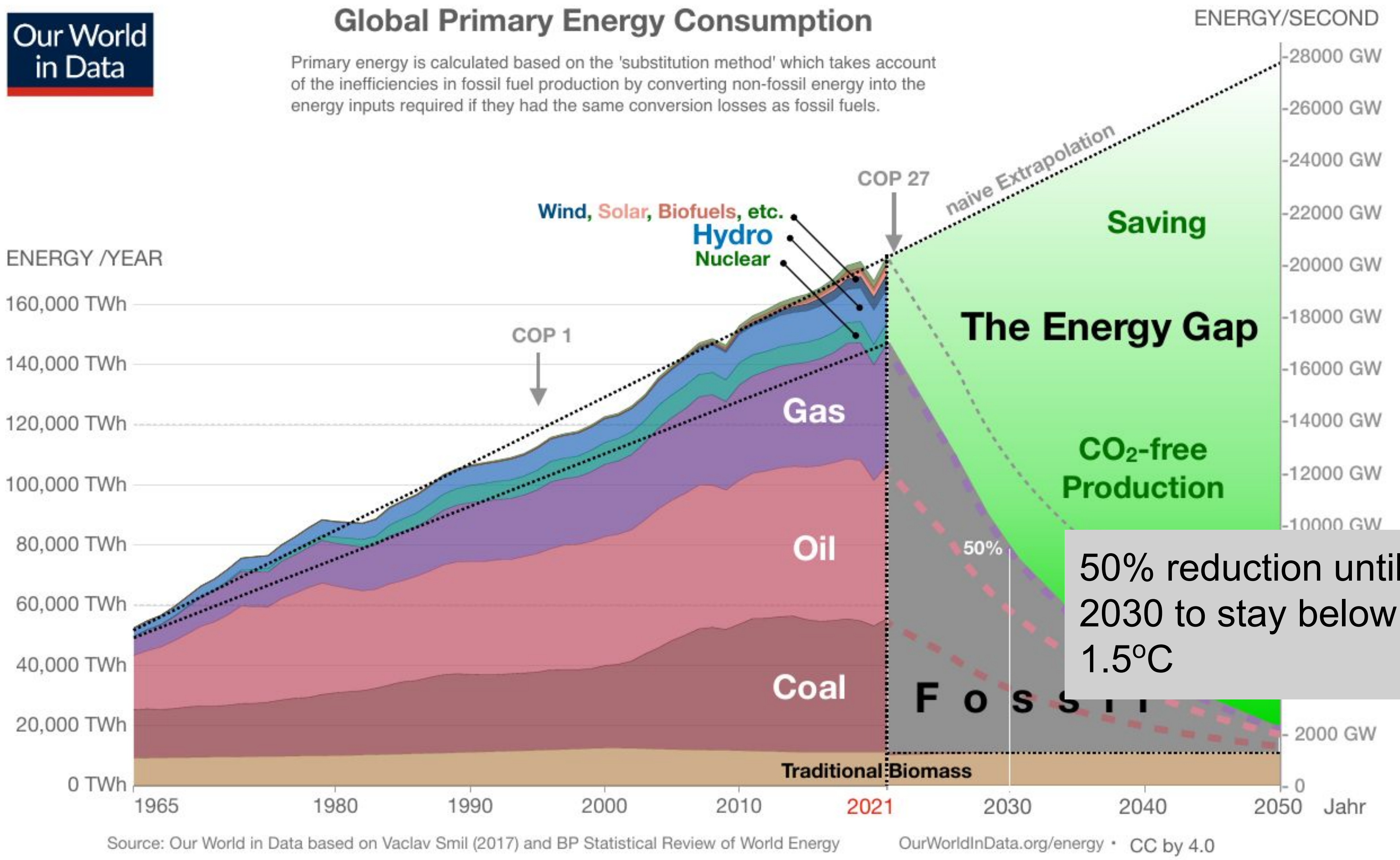
https://www.theguardian.com/world/2025/sep/15/europes-summer-of-extreme-weather-caused-43bn-of-short-term-losses-analysis-finds

- The immediate hit to the economy from a single brutal summer of heat, drought and flooding amounted to 0.26% of the EU's economic output in 2024 (1/10 of European spending on science)

- Costs are expected to rise to €126bn by 2029.

→ Need to take steps to reduce our footprint in line with Paris expectations

The energy gap



50% reduction until 2030 to stay below 1.5°C

The energy gap

Our World in Data

Global Primary Energy Consumption

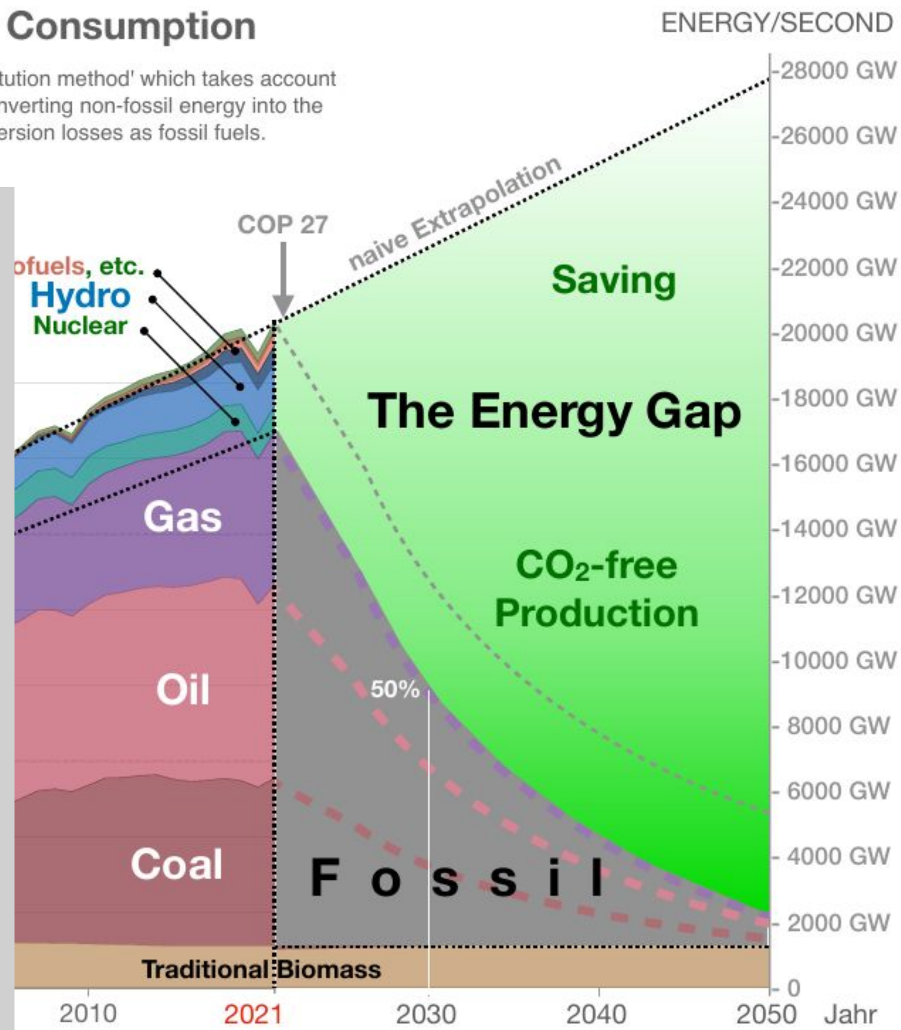
Primary energy is calculated based on the 'substitution method' which takes account of the inefficiencies in fossil fuel production by converting non-fossil energy into the energy inputs required if they had the same conversion losses as fossil fuels.

Options:

1) Expand CO2-free energies
→ factor ~12 in 7 years required;

2) Increase energy efficiency
→ factor ~2 in 7 years
e.g. Electrification of engines (factor 3-5 vs. combustion engine)
e.g. LEDs for lighting (factor 10 vs. light bulb)

3) Save energy
→ factor ~2 in 7 years
e.g.Less travel: online conferences, holidays nearby
e.g. Fewer consumer items, more repair options
e.g. Energy priority for essential things



Source: Our World in Data based on Vaclav Smil (2017) and BP Statistical Review of World Energy OurWorldInData.org/energy • CC by 4.0

What does this mean for computing?

Options:

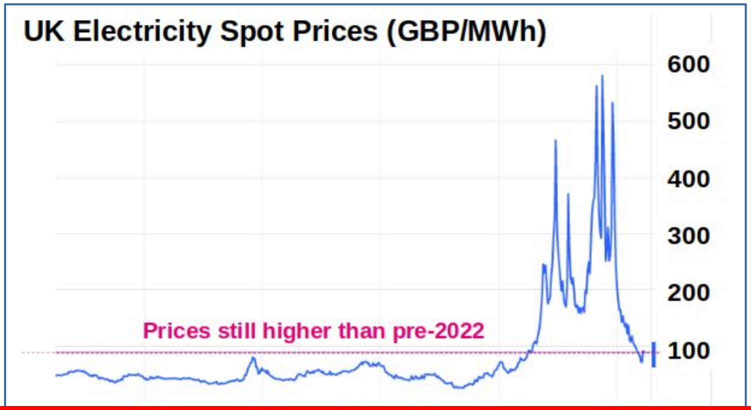
1) Expand CO2-free energies (factor 12)
Renewable power for computing: processors and cooling;
Consider district heating and site selection;
Job scheduling according to energy availability (might not always be possible); ...

2) Increase energy efficiency (factor 2)
Optimised processors (clocks, GPUs),
architecture, cooling system,
software, in the future: quantum/neuromorphic computing?, ...

3) Save energy (factor 2)
Prioritise research questions
Optimise debugging, statistics and precision;
Modular and reusable software;
Modular and repairable hardware, reduce purchases;

Can’t we just use green energy and not do anything?

- **Electricity prices are volatile**
- **EU projections from 2016 predict about 25% rise of prices (consumer)**
  → Cut 25% of the physics?



The short answer is NO

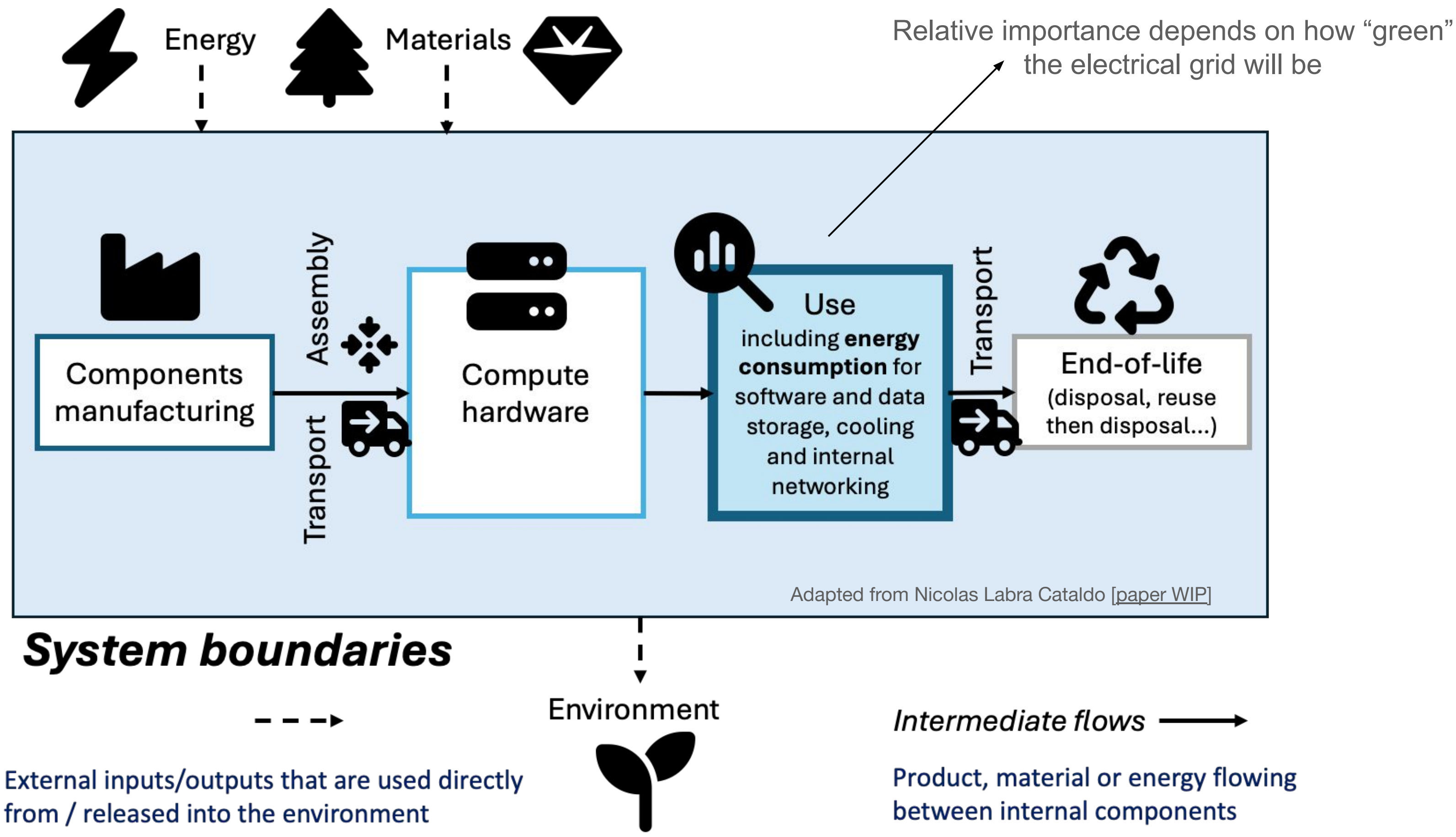
(slide also in backup)

Unit costs (10-20% decrease/yr), 5 years of lifetime
→ **Costs could rise between 0.5 – 5.5 (best vs. worst case scenario)**

- Electricity costs (based on average) consider inflation, power efficiency (30% decrease → no improvement), high prices+high inflation versus both dropping
  → **Costs could rise ranging by 1.6 – 3 – 7 (based on mid capacity)**

Chris Brew (RAL)

From energy awareness to <u>resource</u> awareness



https://www.iso.org/standard/37456.html, but also https://arxiv.org/abs/2506.14365 from our community

Icons: https://fontawesome.com

Sustainability in HECAP+ and Computing

- **Sustainability has become a big topic at the past few CHEP conferences with activities triggered by the 2022 energy crisis**

- eg. focus of large project funded in German community (https://erumdatahub.de/)
  https://indico.desy.de/event/37480/
  https://indico.desy.de/event/47133/
  (project for digital transformation)

- UKRI funding as well: Net Zero scoping Project, NetDRIVE

- It's not a new topic: https://indico.esrf.fr/event/2/
  Sixth Workshop on Energy for Sustainable Science at Research Infrastructures

- Recent reflection document on sustainability in High Energy Physics, Cosmology and Astroparticle Physics + Hadron and Nuclear Physics (Computing only a small part of this document)


Environmental sustainability in basic research

A perspective from HECAP+

Abstract

The climate crisis and the degradation of the world's ecosystems require humanity to take immediate action. The international scientific community has a responsibility to limit the negative environmental impacts of basic research. The **HECAP+ communities** (**High Energy Physics, Cosmology, Astroparticle Physics, and Hadron and Nuclear Physics**) make use of common and similar experimental infrastructure, such as accelerators and observatories, and rely similarly on the processing of big data. Our communities therefore face similar challenges to improving the sustainability of our research. This document aims to reflect on the environmental impacts of our work practices and research infrastructure, to highlight best practice, to make recommendations for positive changes, and to identify the opportunities and challenges that such changes present for wider aspects of social responsibility.

Version 1.0, 5 June 2023

Please read this document in electronic format where possible and refrain from printing it unless absolutely necessary. Thank you.

https://sustainable-hecap-plus.github.io/

Hardware

- **Manufacturing 50% - 80% of a devices CO2e footprint (server vs. laptop)**
  → Infrastructure to keep, reuse, recycle, repair! Extend use lifecycle
  → staff extensive, on the level of single institution
  → on larger scale (clusters): potentially complicated to organise (especially when moving old hardware to a different cluster)

- **Use of certified products (e.g. TCO certification, though that probably already covers most of hardware)**

- **'Energy proportionality' is important:**
  energy consumption should be proportional to computing performance over the full range of applications → idle hardware consumes power as well

- → tests needed to find optimum usage, depending on architecture

- Potential in reducing clock frequency ~about same amount of HEP work at 94-98%

Rodney Walker:
https://indico.desy.de/event/37480/contributions/140510/attachments/82246/108365/Meinerzhagern_compOps(2).pdf

Infrastructure

- **Well managed, centralized systems key to address challenges**

  → Optimized PUE (=Power Usage effectiveness → Total Power/Energy used by IT)

  → Current **best centres:** 1.05-1.2 mainly due to heat recovery from cooling system for heating (HECAP+ examples: GSI green cude 1.07, CERN data centre: 1.5 (1.1 planned), Swiss National supercomputing (1.2 at 25% full load)

  → world average ~1.55, WLCG assumed ~1.5

  → Note: You can get below 1 (e.g. district heating, heat usage), water may be trade-off

- **Centralization here helps, in particular to run hardware optimized for specific (HEP) applications**

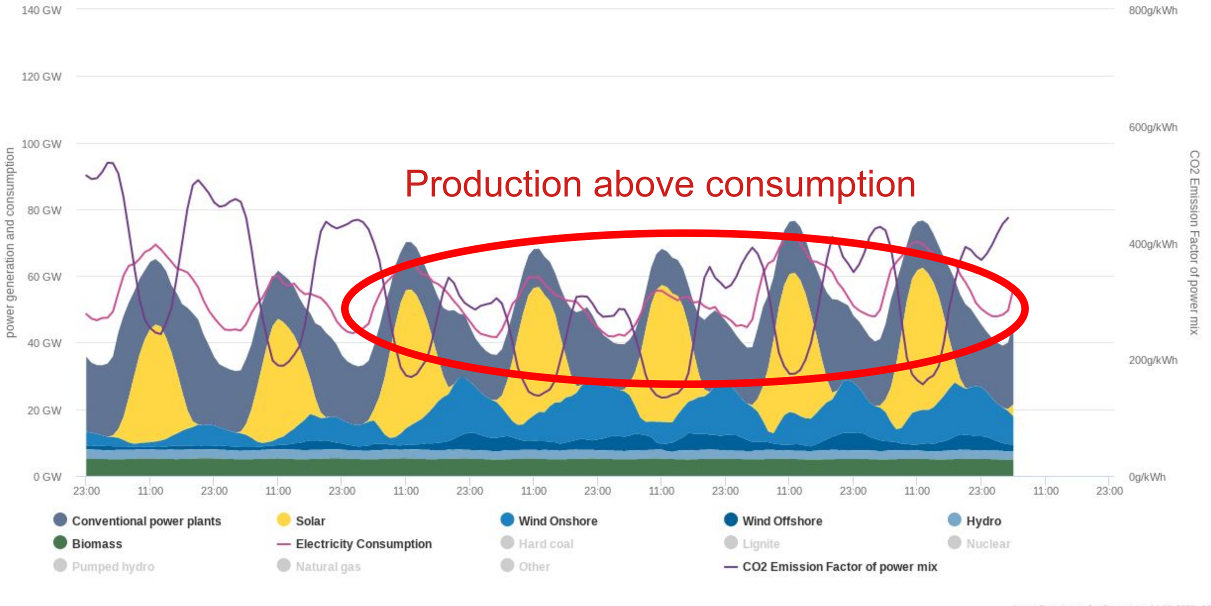
  → (HTC versus HCP which can make local resources difficult to use)

Infrastructure

- **Usage of carbon-free energy paramount**

  → “Own” production (requires investment into solar + potentially storage)

  → Regulation of load according to prices (“Follow the money” – R.W.), prices can be negative, but requires special tariff that can be used → well maintained data centres reacting to production and other grid loads, can help balance grid



Lancium Computing centre
https://indico.desy.de/event/37480/contributions/138296/attachments/82407/108618/2023-05-30%20Concrete%20Action.pdf

Software

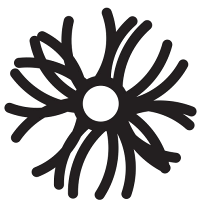
- **HECAP+ Code relies on libraries and public codes, general frameworks and software infrastructure provided by experts in the experiments.**
- Strict requirements posed by the computing environment.
- Impact directly measureable e.g. cosmological analyses → using Likelihood Inference Neural Network Accelerator (LINNA) for efficiency could save **$300,000 in energy costs and around 2,200 tCO2** in first-year for Rubin Observatory's Legacy Survey of Space and Time (LSST) analyses (https://dx.doi.org/10.1088/1475-7516/2023/01/016)
- **Dedicated efforts can have a huge impact!**
- Need sustained effort, continued recognition and dedicated and well-trained person power
→ need to use leverage with experiments, mechanisms to allow ***more*** people to make a career of these efforts **within** the field

What does it mean for “AI software” in HEP?

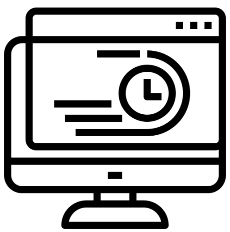
There are **incentives** to make software/AI more environmentally sustainable

Generally, faster software/AI == fewer resources needed to run it

I want results ASAP



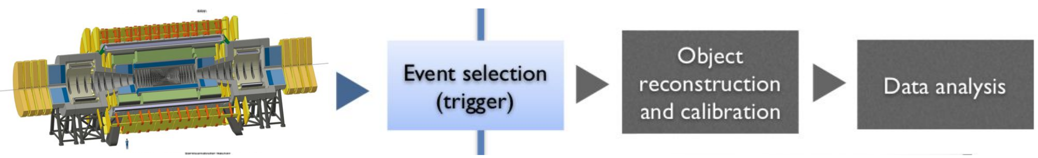
Attribution link



Attribution link

Attribution link

I’m running on a time/latency-constrained environment



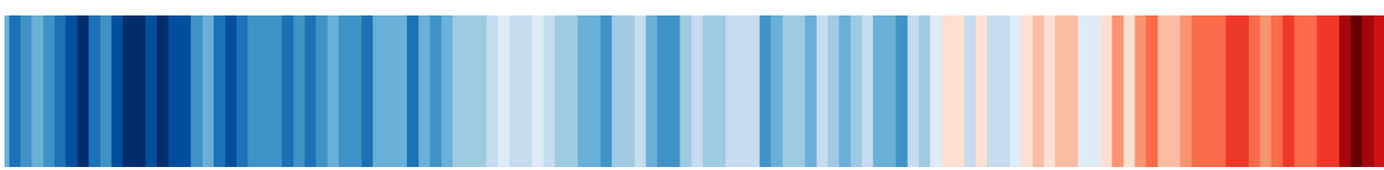
CERN-2022-02

Also, we all want our AI papers to be presented at NeurIPS

8. **Experiments Compute Resource:** For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

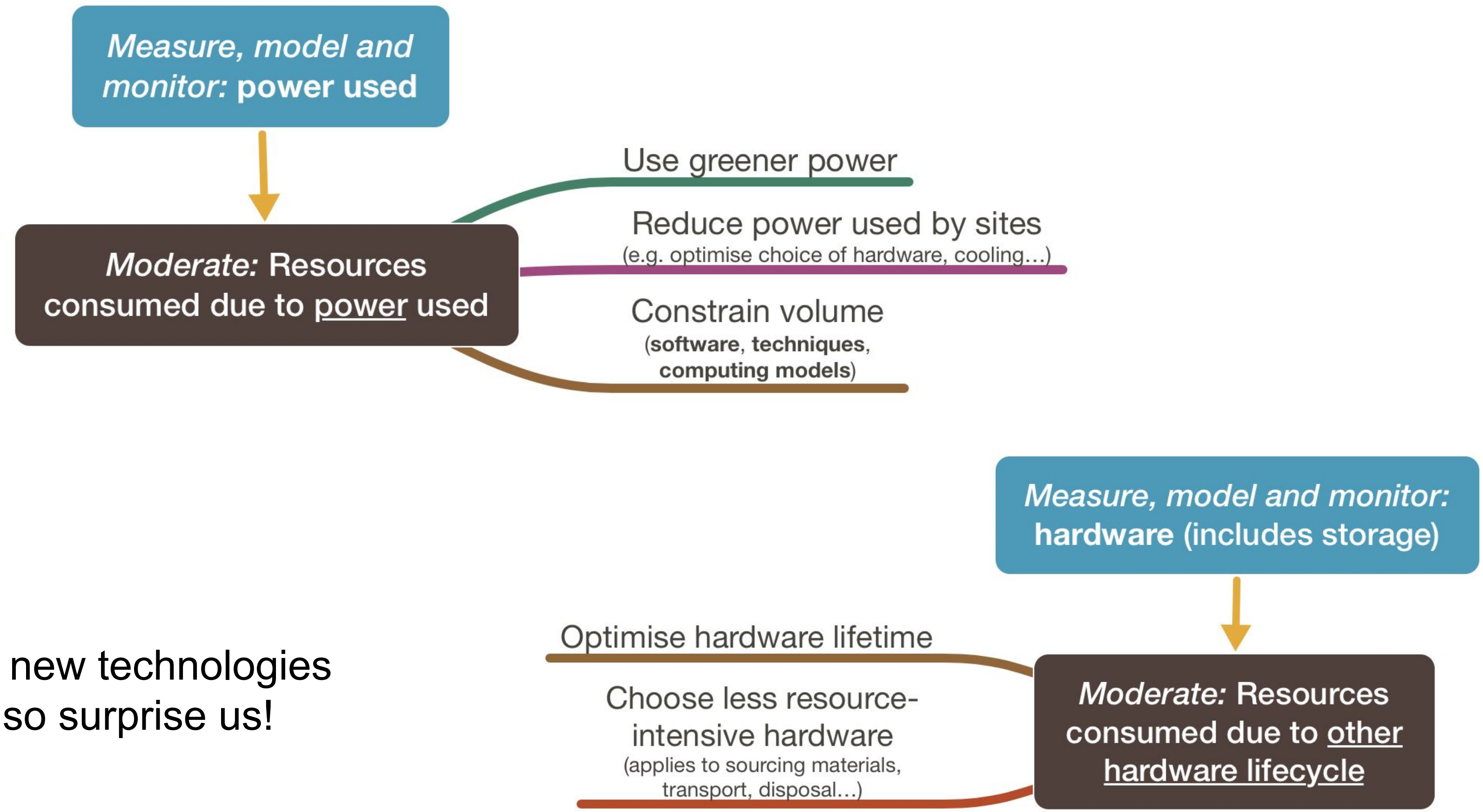
- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute. The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper)

https://neurips.cc/public/guides/PaperChecklist



How to get there, in principle? [by Dave Britton]

Original diagram by David Britton, WLCG environmental sustainability workshop



Note: new technologies can also surprise us!

The University of Manchester

Risks (from the UKRI Net Zero scoping report)

2.3.4 Tensions between competing objectives

There will be tensions between objectives that could be seen as opposing or competing. In many cases it is not that one set of objectives can or should be chosen in exclusion of other objectives. Rather, for example, we need to find a balance. This is not a modern problem, and is illustrated in Figure 2.3.4.A where it is necessary to steer (via regular reviews) between the extreme of unfettered innovation (Scylla) or of excessive restraint (Charybdis).

This section explores the key tensions discovered during the Scoping Project.

Figure 2.3.4.A: Steering a passage between the rock of Scylla (inadequate steps and the to cut emissions leading to a crash) whirlpool of Charybdis (constraints on innovations leading to a tragic downward spiral). (John Doyle, National Portrait Gallery)

AI is consuming
a lot of resources

Let's put
(funding/moral/…)
caps on resource-
consuming
AI for HEP

European Research Council
MANCHESTER 1824
The University of Manchester

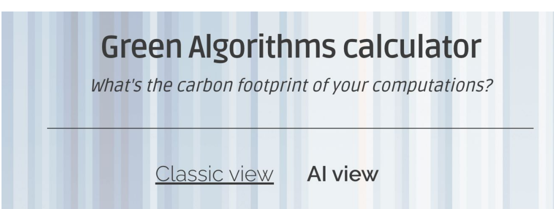
How to get there, in practice? Start <u>measuring</u>

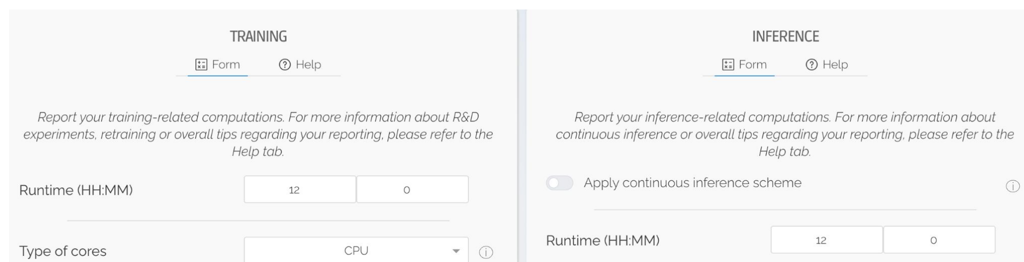
we're physicists after all

Tool Name	Type of Tool	Measured Components	Process or Machine-le...	Limitations	Carbon footprint or En..
GreenAlgorithms	Online Calculator	CPU GPU RAM	Process		Carbon Footprint
CodeCarbon	Embedded Package	CPU RAM GPU	Both	Nvidia GPU	Carbon Footprint
Eco2AI	Embedded Package	CPU RAM GPU	Machine	Nvidia GPU	Carbon Footprint
CarbonTracker	Embedded Package	CPU RAM GPU	Machine	Intel CPU Nvidia GPU Linux/mac OS	Carbon Footprint
MLCO2	Online Calculator	GPU	Machine		Carbon Footprint
CloudCarbonFootprint	Online Calculator	vCPU GPU RAM S	Machine	Cloud providers only	Carbon Footprint
Scaphandre	Server-side Tool	CPU RAM	Process	Intel/AMD CPU Linux/Windows OS	Energy
Kepler	Server-side Tool	CPU GPU RAM	Both	Linux OS Intel/AMD CPU	Energy
PowerJoular	Server-side Tool	CPU GPU RAM	Both	Intel/AMD CPU Nvidia GPU Linux OS	Energy
JoularJX	Embedded Package (Java)	CPU GPU RAM	Both	Nvidia GPU Intel/AMD CPU	Energy

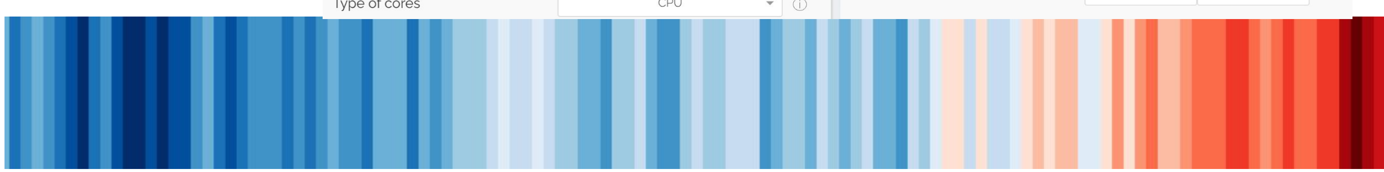
Rosie Schiffmann, UofM internship 2025, Zenodo link For links, see table

AI-specific calculator, separating training/inference





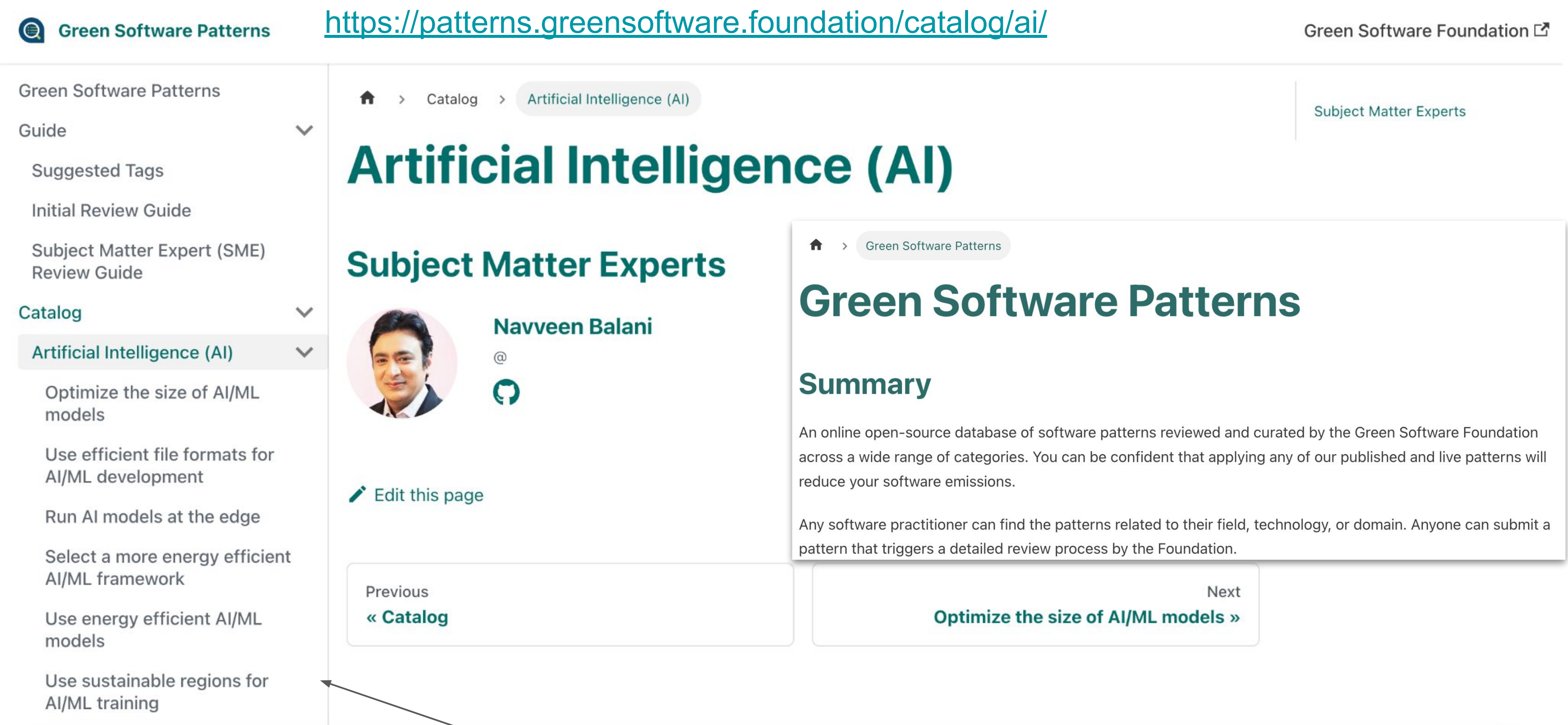
European Research Council
MANCHESTER 1824
The University of Manchester



How to get there, in practice? Start applying <u>patterns</u>

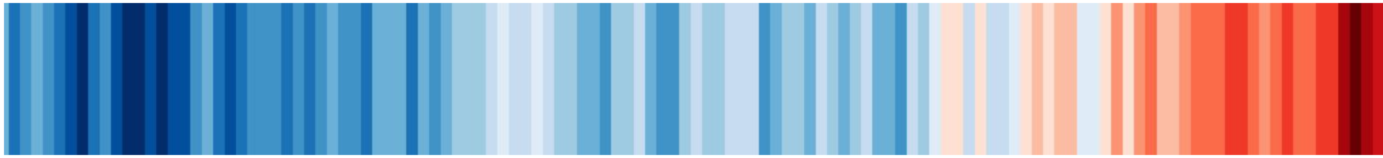
physicists like symmetries

https://patterns.greensoftware.foundation/catalog/ai/



Not all of them are usable/feasible for us, but it's a start

European Research Council
MANCHESTER 1824
The University of Manchester



Many research directions (from EuCAIF WG3 kick-off)

PATH TOWARD ENERGY EFFICIENT AI

STRATEGIC	HARDWARE	DEPLOYMENT	AI ARCHITECTURES
Optimise **use case definition**	Improve usage efficiency of available h/w	Optimise across the **AI lifecycle**	Improved/ compactified **data representation and computational graphs**
Optimise **integration** with existing software	**AI models are based on a few frameworks: optimising them impacts all use cases**	Optimise workloads **definition, scheduling, …**	**Foundation models**
Estimate classical tools **replacement savings**	Introduce **new h/w technologies** (dedicated accelerators, Quantum Computing…)	**Data centers choice:** centralisation allows better resource managament	New **approaches to training**
Actively contribute to **existing green intiatives beyond HEP**			**Neural Architecture Search (NAS)**

Sofia Vallecorsa (CERN)

Where to learn / discuss / collaborate further?

EuCAIF Working Group 3 - FAIR & Sustainable AI [indico]
[CERN egroups: *EUCAIF-wg-FAIRSUSTAINABLE*]

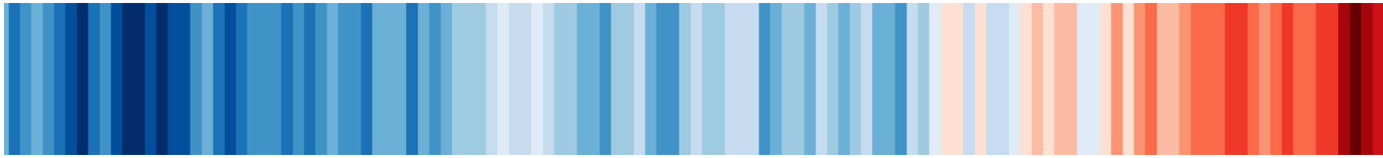
WLCG Environmental Sustainability Forum
First meeting: 1/10/2025 (save the date, agenda will appear here)
Coordinators: David Britton, Caterina Doglioni, Markus Schulz

UK's **Software Sustainability Institute** (search for "green software" for resources and activities), in synergy with EVERSE EU project [environmental sustainability page WIP] and the HEP Software Foundation, with E. Rodrigues (Liverpool) as chair of the Core Steering Group

Environmentally Sustainable Computer Science Forum
hosted by Loïc Lannelongue at Cambridge

The NetDRIVE project annual meeting in Durham next week (seems like it's still possible to join the hybrid part of the event)

ASCEND initiative - see afternoon slides, if interested in joining a "task force" for writing about environmental sustainability of software and AI in HEP in the next few weeks, contact us (Caterina and Kristin)



Some conclusions

- **We (as a community) have made big progress and substantial improvements**
  (considering the constraints potentially as much as e.g. google/amazon)

- **But is it enough to achieve 50% overall reduction of CO2e?**

- **3 handles:**
  Green energy → factor of 12
  Energy efficiency → factor of 2
  Energy saving → factor of 2

- **Will need a hard look and many, many sacrifices (not only in the computing sector)**

- **Will require a concerted effort and dedicated funding**
  → but as a community we are certainly better placed than other fields of science (which are/will also come under scrutiny)

Need framework with benchmarks and goals and
Ability to *shape* (institutional/funding) constraints to allow achieving goals

Climate benchmarks that need to be met
(→restricted physics exploitation scenarios, what can we sacrifice?)

Some questions

- **Infrastructure:**
  - → Further centralize?
    → Possibility for own power generation? (→saves money, but are those eligible costs?)
    → tools for power management, using negative tariffs, “follow the money”
    → Support for small local cluster → easy to handle tools to allow this for small sites?

- **Hardware:**
  - → Keep old hardware to run as backup / addition at high-energy production times
  - → How to manage old hardware (distribution to other sites possible?)
    → study interplay hardware/software
    → High-performance versus High-throughput (HPC versus HTC)

- **Software:**
  - → Needs dedicated efforts with possibility to retain people
    → ability to make this a career
    → ability for the end user to monitor
    → training, training, training

(said during the talk, but not in the original presentation slides shown):

Do we need contingency plans on restricting physics/research?

Reminder: Paris agreement is in principle legally binding
→ pressure on us / our savings might need to be increased
→ gives us negotiating power if we have a clear plan and strategy with demonstrable impacts and realistically achievable objectives in line with 1.5°C

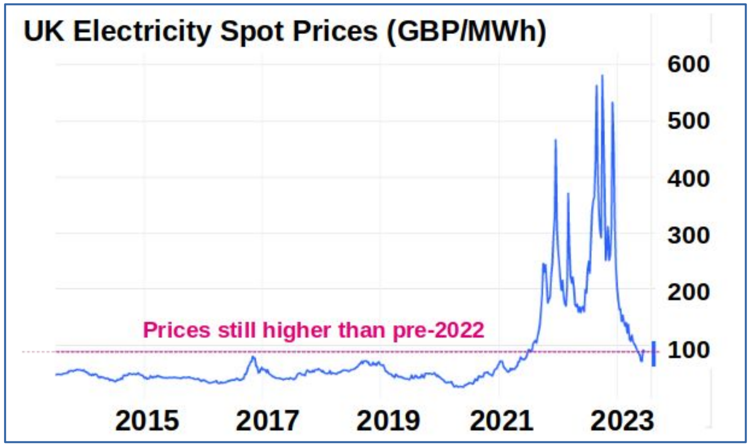
Backup

Can’t we just use green energy and not do anything?

- **Electricity prices are volatile**

- **EU projections from 2016 predict about 25% rise of prices (consumer)**
  → Cut 25% of the physics?

- And it’s not just electricity prices but also hardware



- **Costs of computing infrastructure evaluation 2032 (with 2021 as index)**

- Installed hardware based on computational requirements (15-20% increase/yr), Unit costs (10-20% decrease/yr), 5 years of lifetime
  → **Costs could rise between 0.5 – 5.5 (best vs. worst case scenario)**

- Electricity costs (based on average) consider inflation, power efficiency (30% decrease → no improvement), high prices+high inflation versus both dropping
  → **Costs could rise ranging by 1.6 – 3 – 7 (based on mid capacity)**

Chris Brew (RAL)

ASCEND process [link]

A Statement of Community Engagement and Needs for Digital Research Infrastructure (ASCEND)

- A chance to give our views on challenges/needs in terms of UK Digital Research Infrastructures on the following topics:

Statements of need submitted to this activity should be related to one or more of the key priorities of the programme:

- Interconnected DRI: Enabling transformative research and innovation through collaboration that delivers a strong economy and effective public services.
- Human DRI: Ensuring that we maximise the opportunities to digitalise research and innovation by supporting highly skilled DRI professionals.
- FAIR DRI: Enabling the UK's researchers and innovators to contribute to an open, accessible, and collaborative digital research landscape.
- Sustainable DRI: Building a resilient and environmentally responsible federated national resource that enhances the productivity of researchers and innovators while minimizing its environmental impact and ensuring the security of research data and infrastructure.

The whitepaper from EuCAIF WP4 and from last year's meeting include these topics

What we are looking for

The DRI Programme periodically reviews its investment portfolio and strategy and is considering what projects, facilities and systems will be included in the next phases of the Programme.

Funding

No direct funding is available through this activity. A successful outcome may inform a future funding opportunity.

STRIKE group [link]

UKRI Digital Research Infrastructure Pool: Supporting Technical Review, Insight, Knowledge and Evaluation

- A chance to advise the UKRI Digital Research Infrastructure (DRI) Programme team and the UKRI Digital Research Infrastructure Committee (DRIC) with insight that include HEP expertise on a regular or ad-hoc basis

STRIKE members may be invited to:

- assess and provide feedback on submissions to the UKRI A Statement of Community Engagement and Needs for Digital Research Infrastructure (ASCEND) initiative.
- contribute expertise to time-limited task and finish groups, led by the UKRI DRI Programme Team or DRIC members
- conduct technical due diligence and value-for-money assessments of DRI Programme projects after they have been recommended by the UKRI AGD and approved the UKRI Executive Committee
- provide recommendations to the DRI Programme SRO to assure each initiative is clearly defined, well-designed and feasible
- contribute to peer review panels convened by the DRI Programme Team or DRIC to assess the quality and technical robustness of funding proposals
- participate in reviews of ongoing DRI Programme initiatives to assess technical risks and overall readiness to proceed. These reviews are intended to support informed decision-making and will complement council-level project boards.
- provide rapid input on ad hoc technical issues or questions raised by the DRI Programme Team, DRIC or other UKRI governance bodies

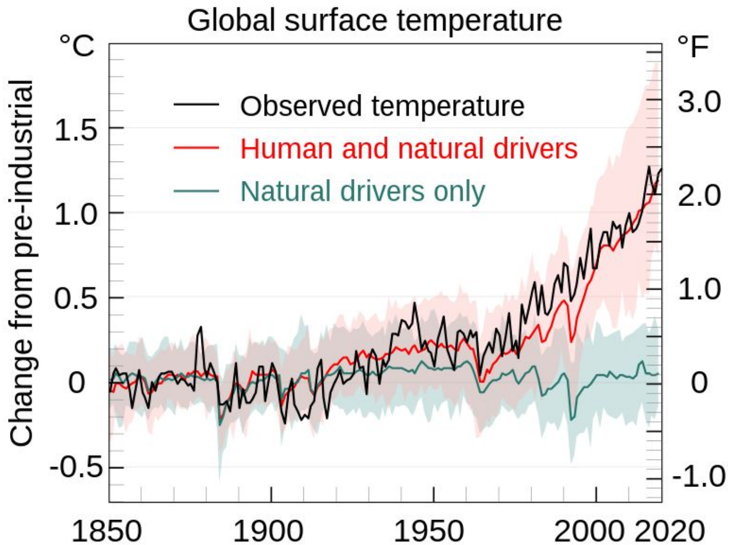
STRIKE is not a standing committee and does not hold regular full member meetings.

Weather or Climate? And is it “natural”?

- Whilst extreme weather events have a finite probability and therefore “just” can happen, this finite probability is strongly influenced by climate conditions
  - → studied in extreme event attribution / attribution science→ new field of study in meteorology and climate science using statistical methods and concepts not completely foreign to particle physicists.
  - → https://en.wikipedia.org/wiki/Extreme_event_attribution

- Using the framework of attribution science, the current level of climate change is fully attributed attributed to human activity

- Climate sets the probability (like a cross-section)
- Weather is a single event (like a collision) drawn from that cross-section



https://en.wikipedia.org/wiki/Attribution_of_recent_climate_change