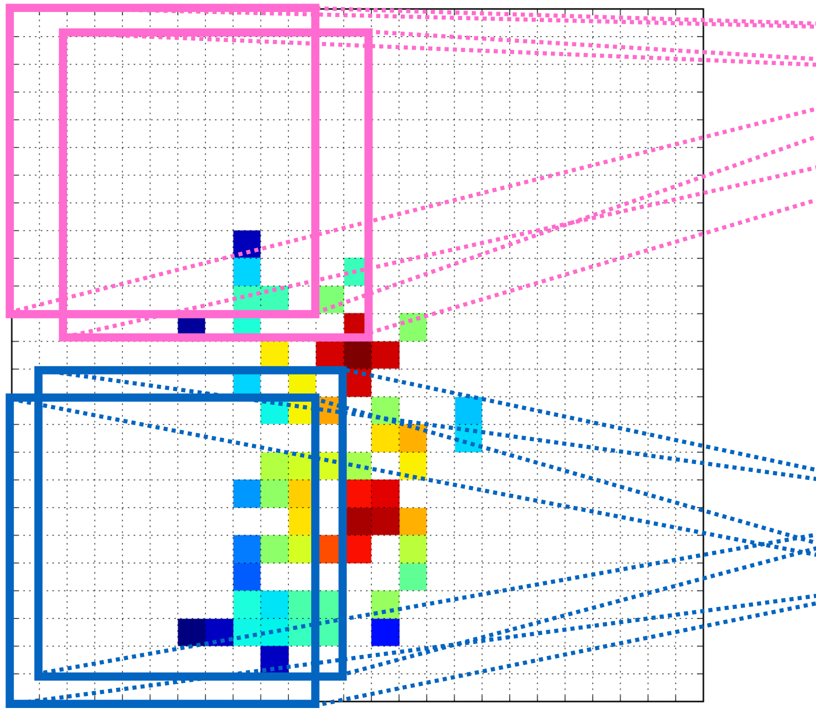
Data Science
Unique NP Challenges

Benjamin Nachman

Lawrence Berkeley National Laboratory

bpnachman.com
bpnachman@lbl.gov

@bpnachman bnachman



BERKELEY LAB

BIDS
BERKELEY INSTITUTE FOR DATA SCIENCE

SANPC
June 2024

Overview

Modern data science offers many tools that we can use directly.

Overview

Modern data science offers many tools that we can use directly.

But, there are also many aspects that industry **won’t solve for us**.

Overview

Modern data science offers many tools that we can use directly.

But, there are also many aspects that industry **won’t solve for us**.
Need Data Physicists for custom solutions to unique NP challenges.

Overview

Modern data science offers many tools that we can use directly.

But, there are also many aspects that industry **won’t solve for us**. Need Data Physicists for custom solutions to unique NP challenges.

Simulation(-based inference)

Non-image/text-based data

Norms for what is “physics” and recognition

Proprietary code/data

Many year-long experiments

Bespoke (legacy) software

Overview II: Ubiquity

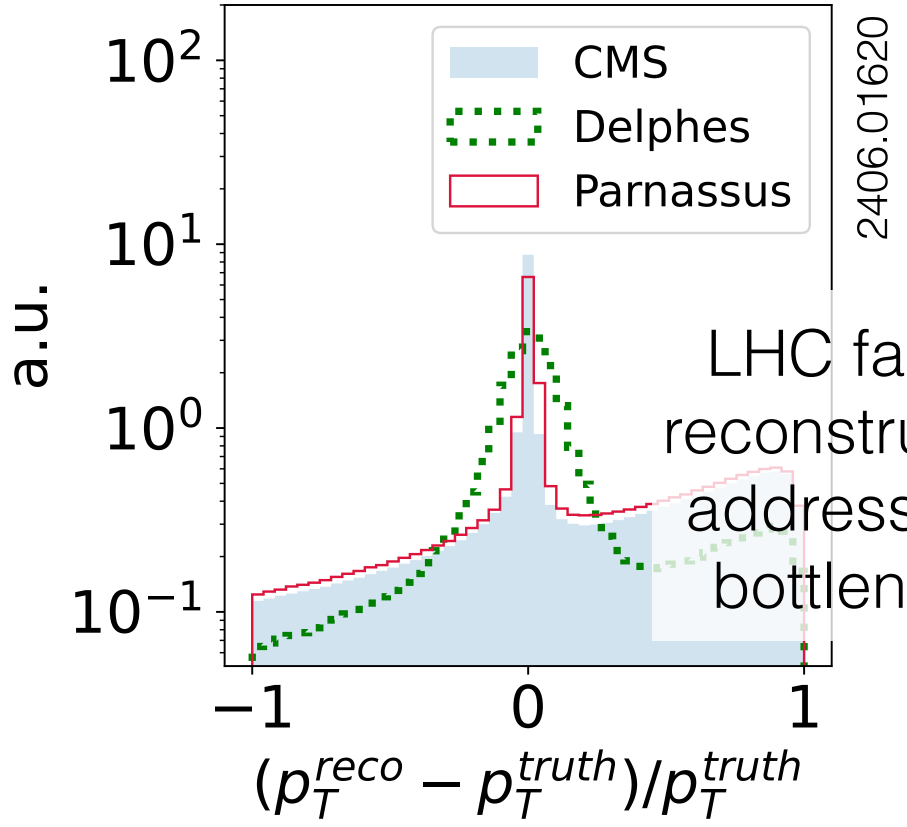
AI/ML is already playing a critical role in nearly all aspects of NP. There is no doubt that it will play a central role for the design, operations, and data analysis of future projects.

1. Facility — *accelerator design, operations; magnet training, …*
2. Detectors — *detector design, construction (e.g. QA/QC), operations, data acquisition, …*
3. Data analysis — *theory, simulation, reconstruction, statistical analysis, …*

one word here doesn't do it justice!

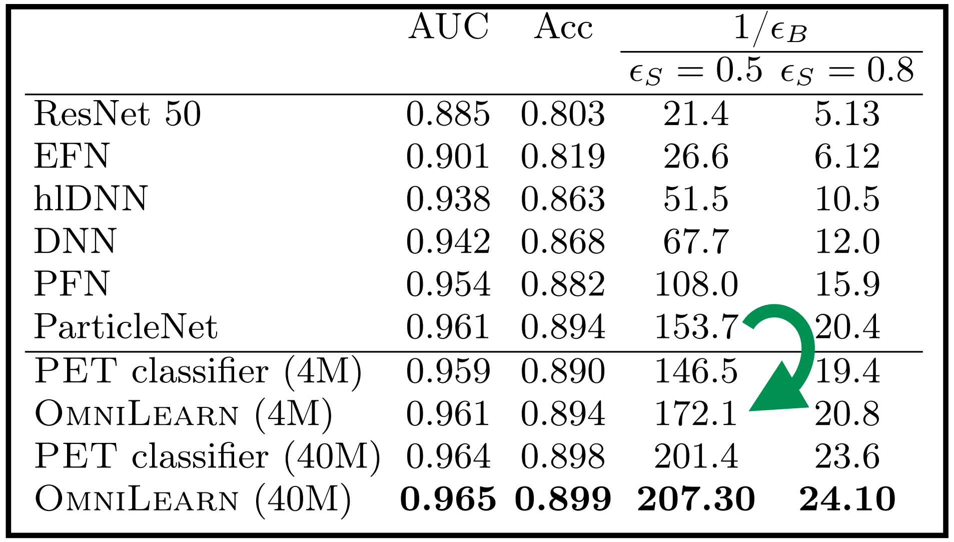
Overview III: Big science for less $$

These are just personal examples - just the tip of the iceberg!



2406.01620

LHC fast simulation & reconstruction together - address computational bottleneck in one go!

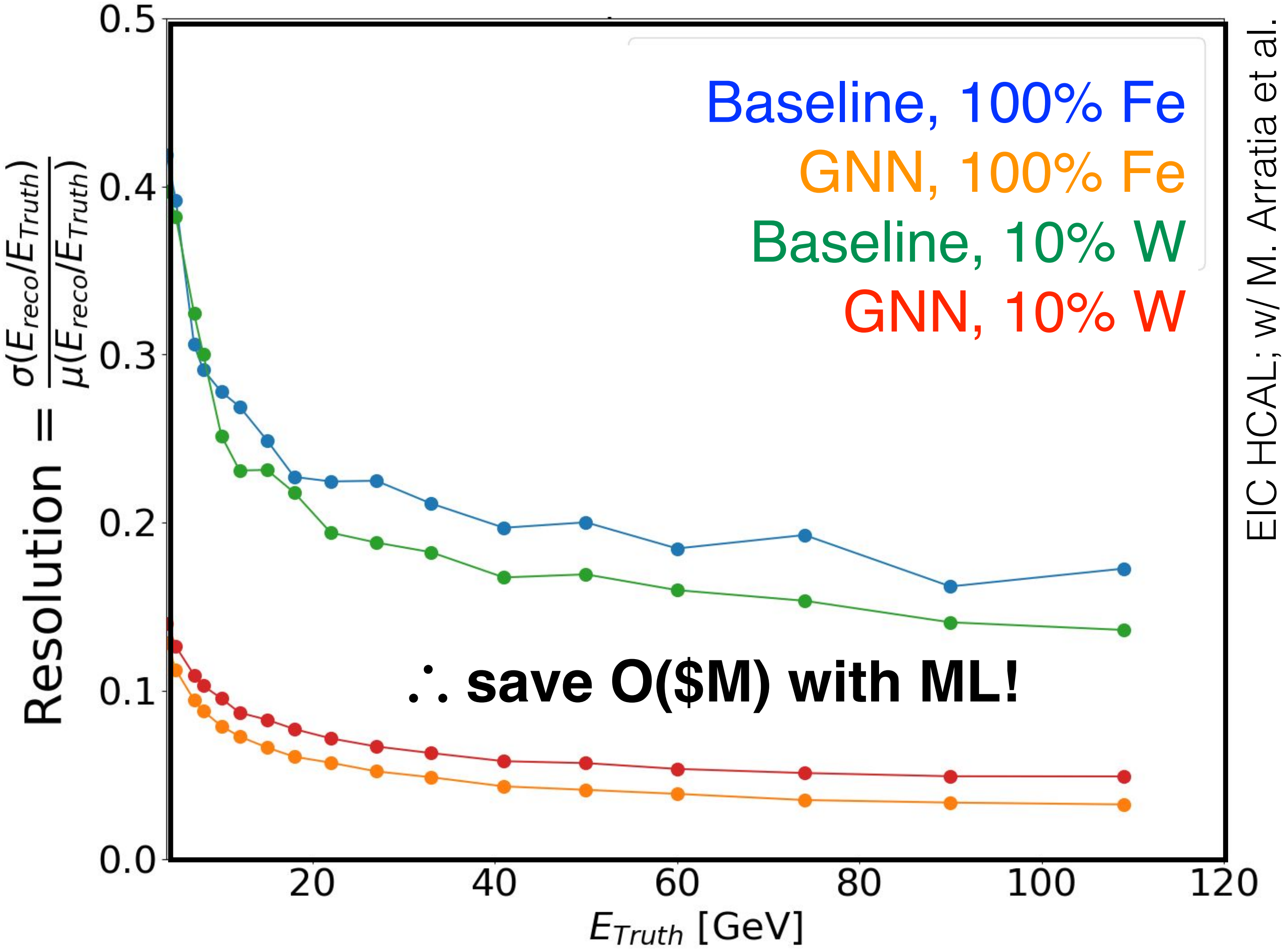


	AUC	Acc	$1/\epsilon_B$	
			$\epsilon_S = 0.5$	$\epsilon_S = 0.8$
ResNet 50	0.885	0.803	21.4	5.13
EFN	0.901	0.819	26.6	6.12
hlDNN	0.938	0.863	51.5	10.5
DNN	0.942	0.868	67.7	12.0
PFN	0.954	0.882	108.0	15.9
ParticleNet	0.961	0.894	153.7	20.4
PET classifier (4M)	0.959	0.890	146.5	19.4
OmniLearn (4M)	0.961	0.894	172.1	20.8
PET classifier (40M)	0.964	0.898	201.4	23.6
OmniLearn (40M)	**0.965**	**0.899**	**207.30**	**24.10**

2404.16091

LHC top tagging:
avoid expensive simulations -
fine tune a foundation model!

Codesign of EIC calorimeter
definitive answer: do we need W?



∴ save O($M) with ML!

EIC HCAL; w/ M. Arratia et al.

Open data != data preservation

We all agree that we want to be able to analyze our unique (and expensive) data in perpetuity.

Open data != data preservation

We all agree that we want to be able to analyze our unique (and expensive) data in perpetuity.

Funding for a project ends when the project ends (by definition)

Open data != data preservation

We all agree that we want to be able to analyze our unique (and expensive) data in perpetuity.

Funding for a project ends when the project ends (by definition)

How to fund modernization of data (and simulation)?

Open data != data preservation

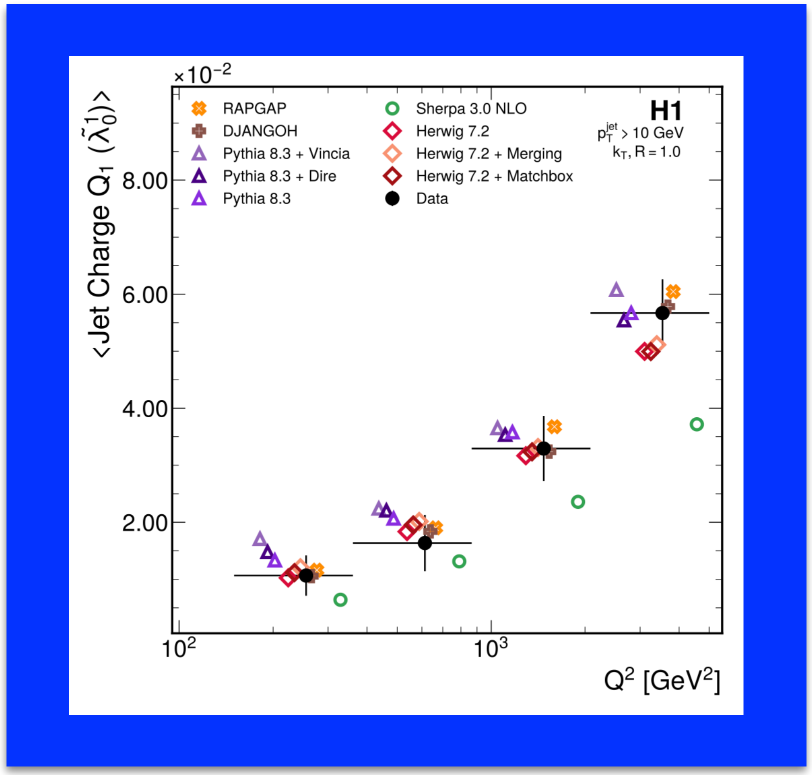
We all agree that we want to be able to analyze our unique (and expensive) data in perpetuity.

Funding for a project ends when the project ends (by definition)

How to fund modernization of data (and simulation)?

Success story: HERA

Failure story: many…



Results in the era of AI

Bread and butter: binned differential cross sections

HEPData

HEP & NP

Repository for publication-related ~~High Energy Physics~~ data

Results in the era of AI

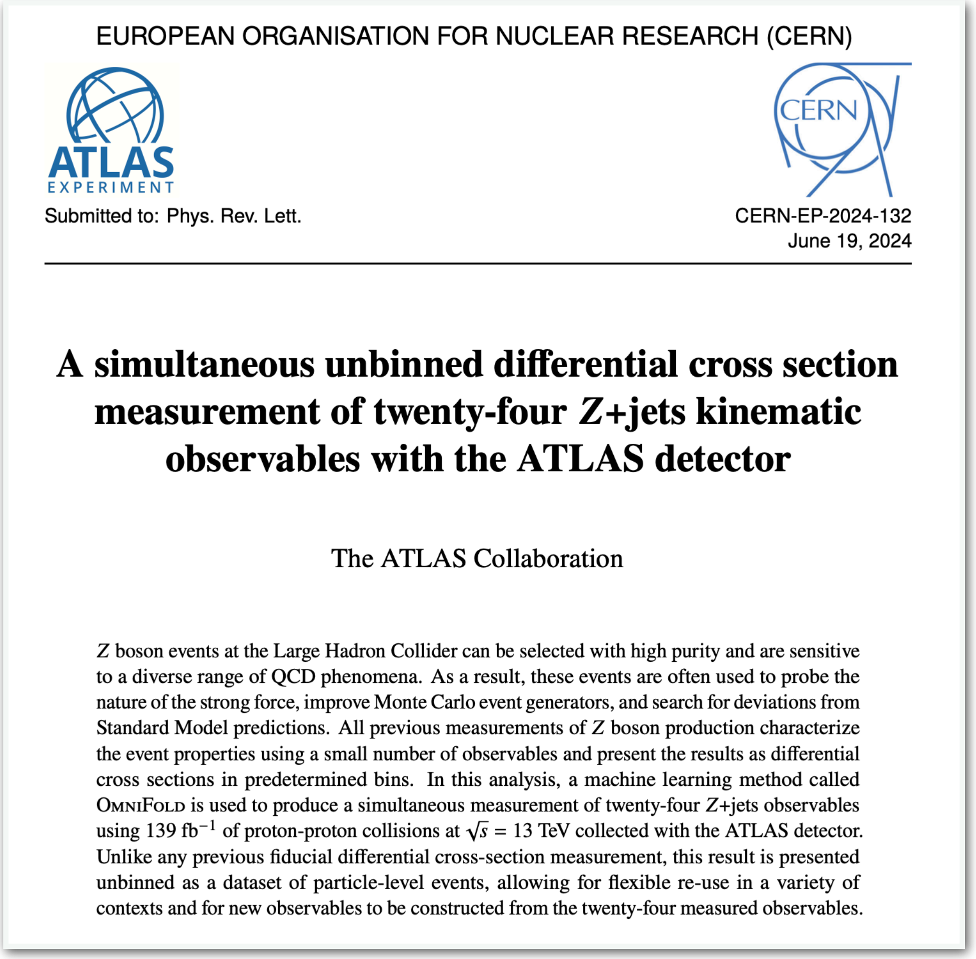
Bread and butter: binned differential cross sections

What about high-dimensional data products?
(e.g. the results are neural networks)

Results in the era of AI

Bread and butter: binned differential cross sections

What about high-dimensional data products?
(e.g. the results are neural networks)



EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH (CERN)

Submitted to: Phys. Rev. Lett.

CERN-EP-2024-132
June 19, 2024

A simultaneous unbinned differential cross section measurement of twenty-four Z+jets kinematic observables with the ATLAS detector

The ATLAS Collaboration

Z boson events at the Large Hadron Collider can be selected with high purity and are sensitive to a diverse range of QCD phenomena. As a result, these events are often used to probe the nature of the strong force, improve Monte Carlo event generators, and search for deviations from Standard Model predictions. All previous measurements of Z boson production characterize the event properties using a small number of observables and present the results as differential cross sections in predetermined bins. In this analysis, a machine learning method called OmniFold is used to produce a simultaneous measurement of twenty-four Z+jets observables using 139 fb^{-1} of proton-proton collisions at $\sqrt{s} = 13$ TeV collected with the ATLAS detector. Unlike any previous fiducial differential cross-section measurement, this result is presented unbinned as a dataset of particle-level events, allowing for flexible re-use in a variety of contexts and for new observables to be constructed from the twenty-four measured observables.

https://gitlab.cern.ch/atlas-physics/public/sm-z-jets-omnifold-2024

Data physicists

Who will address experiment agnostic, cross-cutting methodology for NP?

We don’t need many of these people, but we do need some and they require specialized skills.

Their impact will be huge. We will be able to save a lot of money and for a given budget/detector, achieve much better science.

They are physicists. They are not theorists and they are not experimentalists. They are also not computer scientists or software engineers (although we need those too!)

Simulation - theory or experiment?

Another lost group - simulation developers!

Some of this work is theory and some is experiment,
but most is neither!

How do we fund long-term development, maintenance,
and user support of critical tools like Geant4?
Huge impact → opportunity for US leadership?

N.B. very natural for national labs! Difficult for university
groups (but maybe can change with incentives?)

Forward-proofing code

We need code preservation in addition to data preservation!

Critical need: improve literacy with modern open source software stack (version control, CI/CD, containers, …).

Embrace automation with AI

e.g. can LLMs help us automatically migrate all software efficiently? can they help us with automated documentation?

ROOT versus SciPy

One way to ensure code preservation is to use code everyone is using.

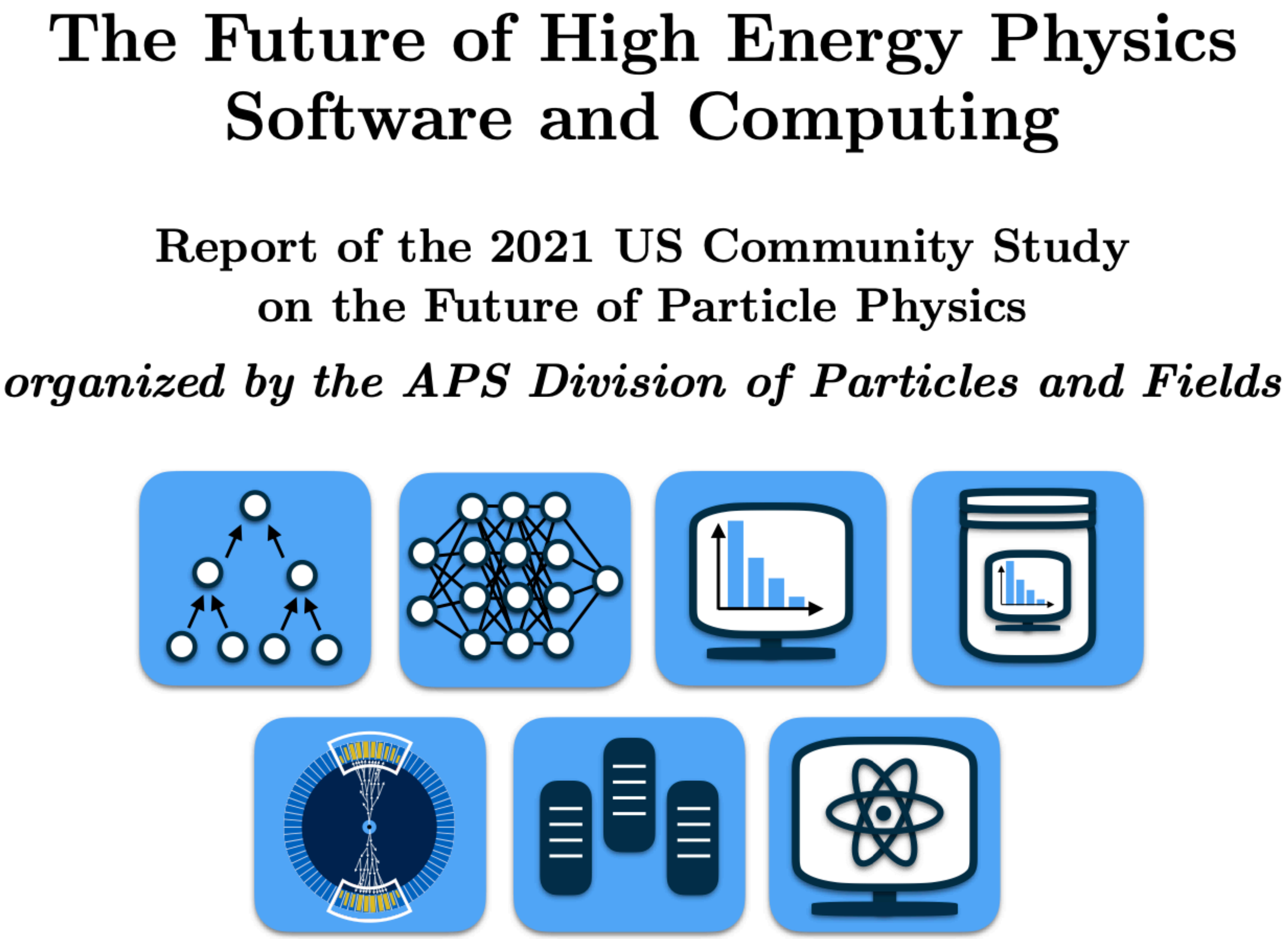
ROOT and other bespoke tools are fantastic and in many ways, were ahead of their time.

We should have a serious conversation about how much we need to depend* on legacy tools with a relatively small user base.

Doesn’t necessarily need to be exclusive or!

*Should NP **contribute** to the development of e.g. SciPy?

Community feedback from HEP



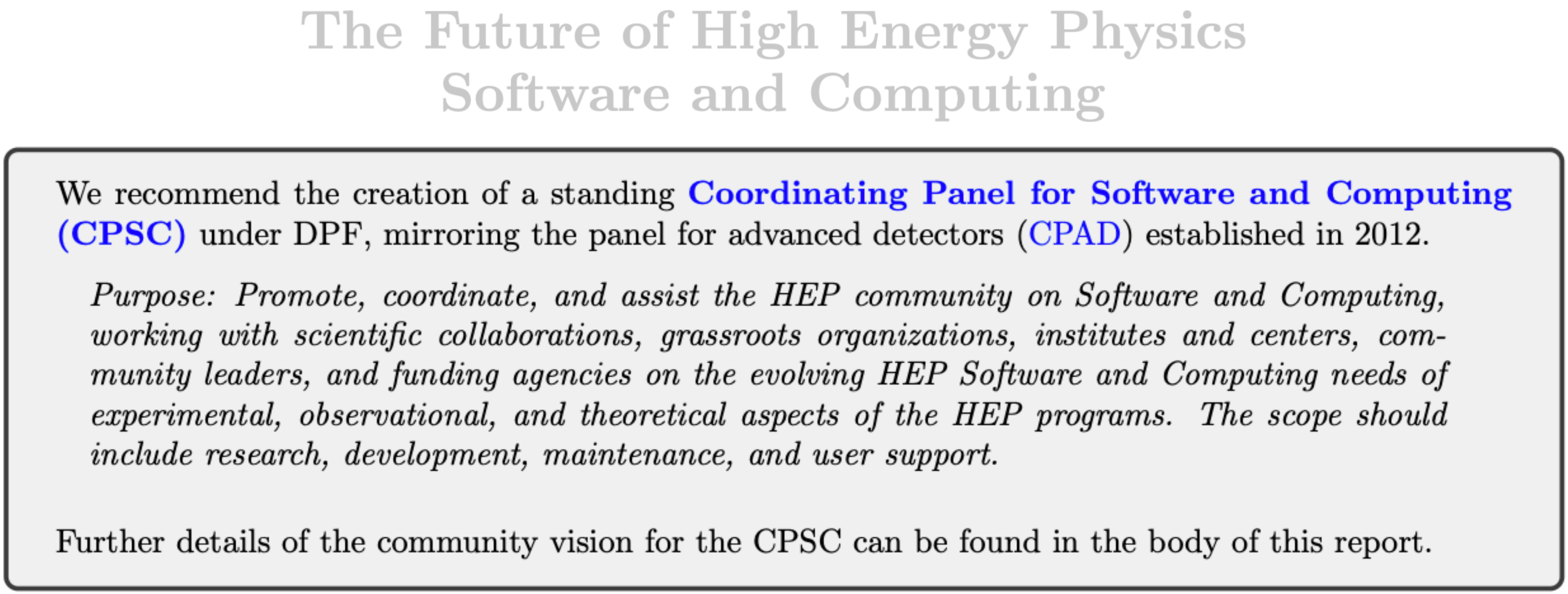
The Future of High Energy Physics Software and Computing

Report of the 2021 US Community Study on the Future of Particle Physics

organized by the APS Division of Particles and Fields

https://arxiv.org/pdf/2210.05822

Community feedback from HEP



The Future of High Energy Physics
Software and Computing

We recommend the creation of a standing **Coordinating Panel for Software and Computing (CPSC)** under DPF, mirroring the panel for advanced detectors (CPAD) established in 2012.

> *Purpose: Promote, coordinate, and assist the HEP community on Software and Computing, working with scientific collaborations, grassroots organizations, institutes and centers, community leaders, and funding agencies on the evolving HEP Software and Computing needs of experimental, observational, and theoretical aspects of the HEP programs. The scope should include research, development, maintenance, and user support.*

Further details of the community vision for the CPSC can be found in the body of this report.

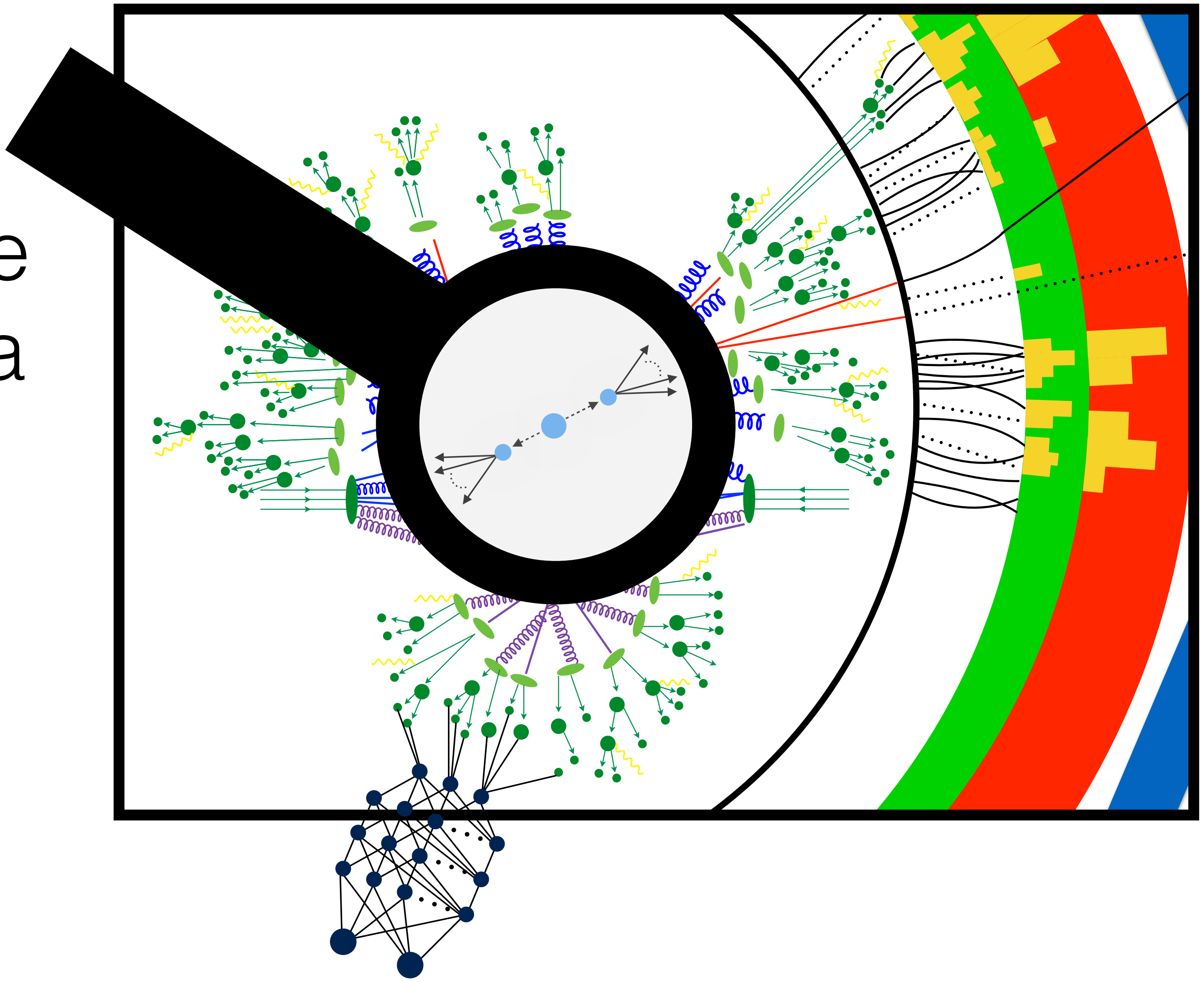
This is happening!

https://arxiv.org/pdf/2210.05822

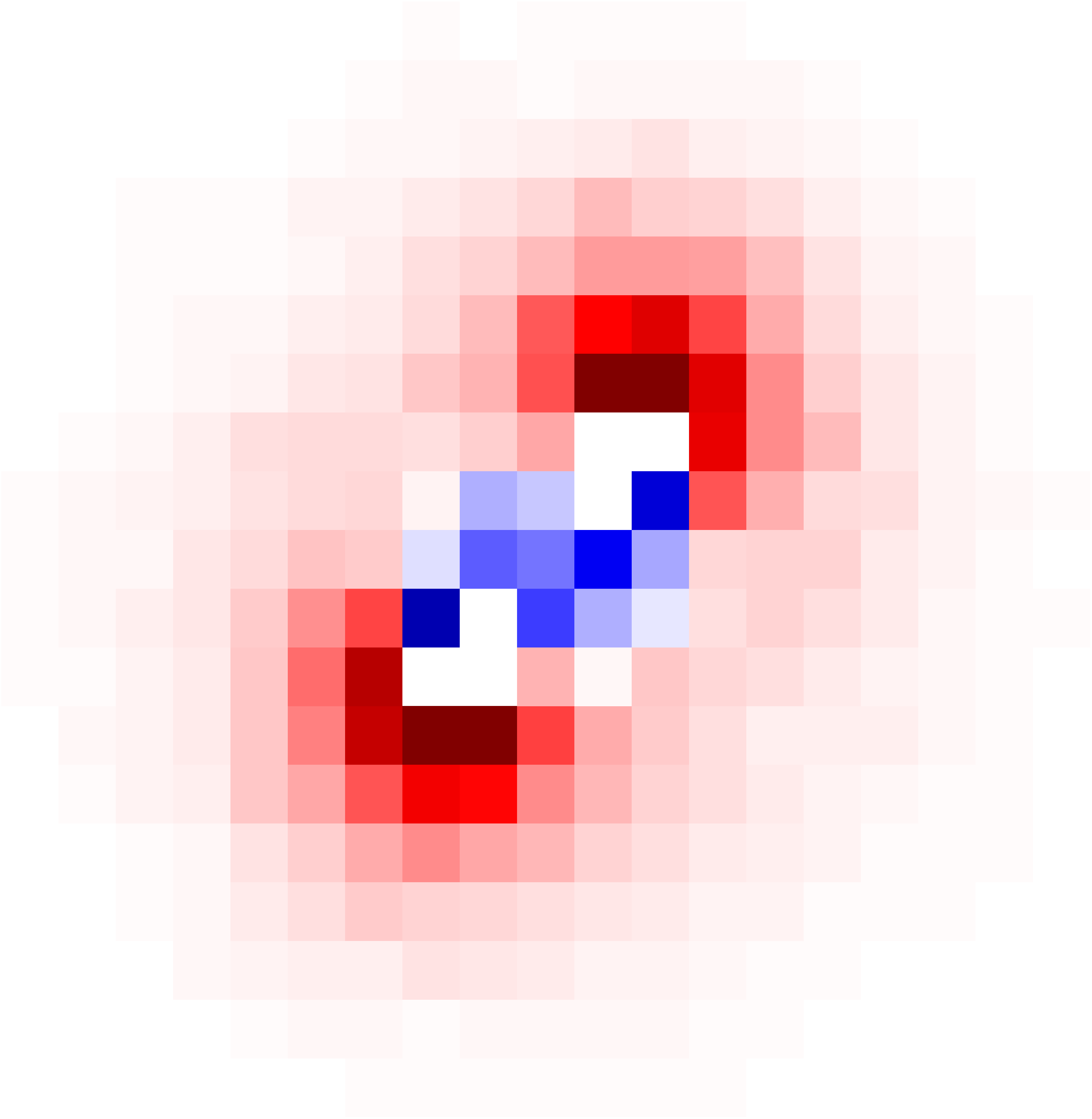
Outlook

I am inspired to be part of this conversation!

This is an exciting time, where we are at a cross roads - data science has a comparable impact to instrumentation on NP science.



Will we be ready now, tomorrow, and beyond ?



Fin.