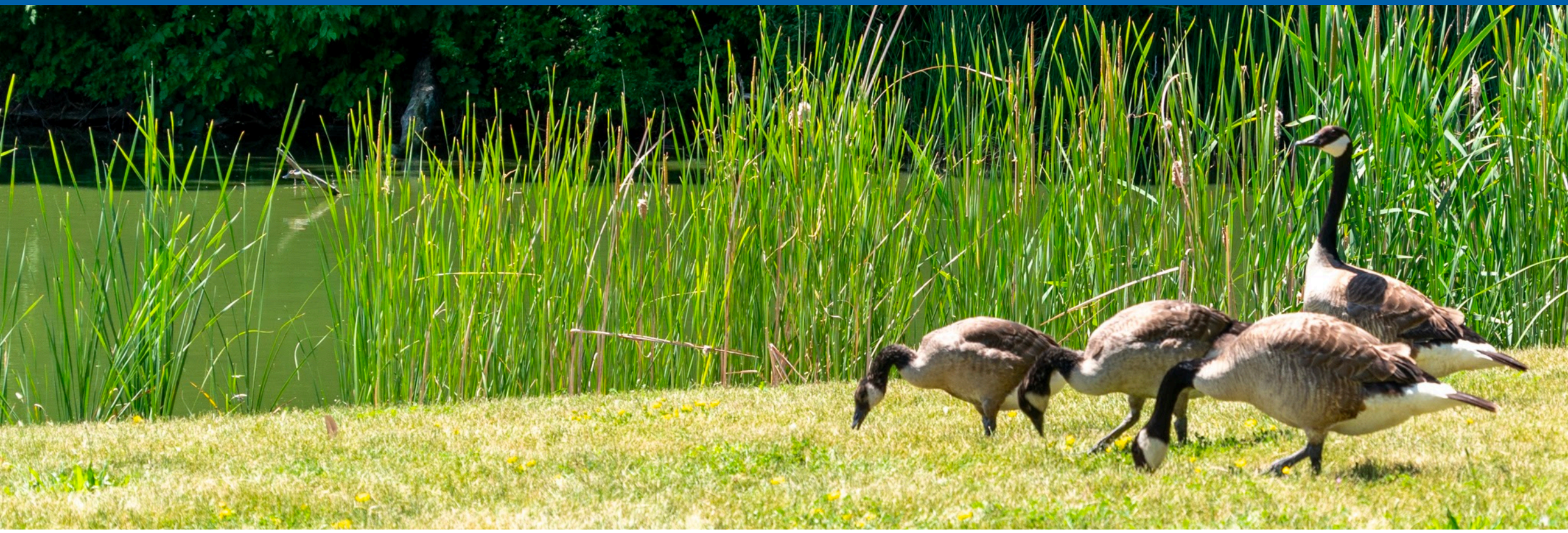


Image credit: Marguerite Tonjes

Potential directions for coffea & scikit-hep

Nick Smith

PyHEP.dev 2023

25 July, 2023

Hi! About me:

- Postdoc @ Fermilab
- Higgs physics
  - Boosted H(bb), H(cc)
  - Higgs combination & EFT
- CMS computing
  - Workflow & Data management operations
    - Operation lead 2020-2022 Rucio transition
  - Storage R&D: Ceph S3 object stores
- Coffea

@monalisatatertot

Coffea project

- A user interface to *columnar analysis*
  - Optimized array programming kernels build an **expressive *and* performant** language
  - Seamless integration with ML tools due to shared interface

Fermilab

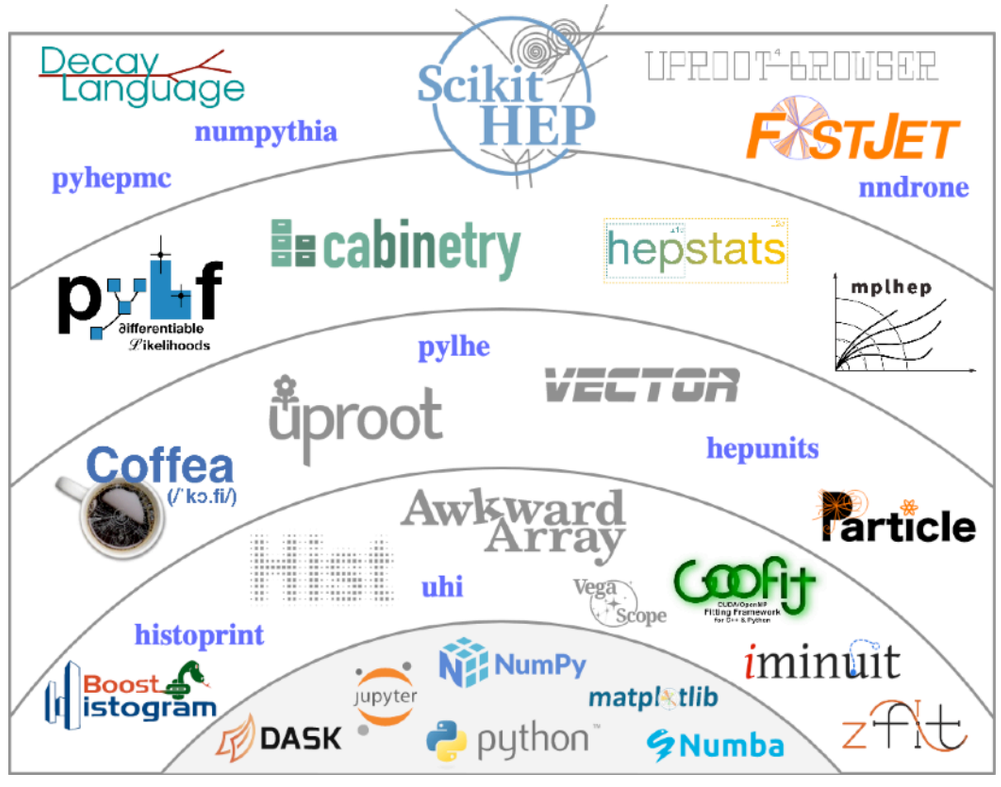
Coffea project

- A user interface to *columnar analysis*
  - Optimized array programming kernels build an **expressive *and* performant** language
  - Seamless integration with ML tools due to shared interface
- An incubator for rapid prototyping
  - Fill in missing pieces of ecosystem
  - Good abstractions are factored out

Fermilab

Coffea project

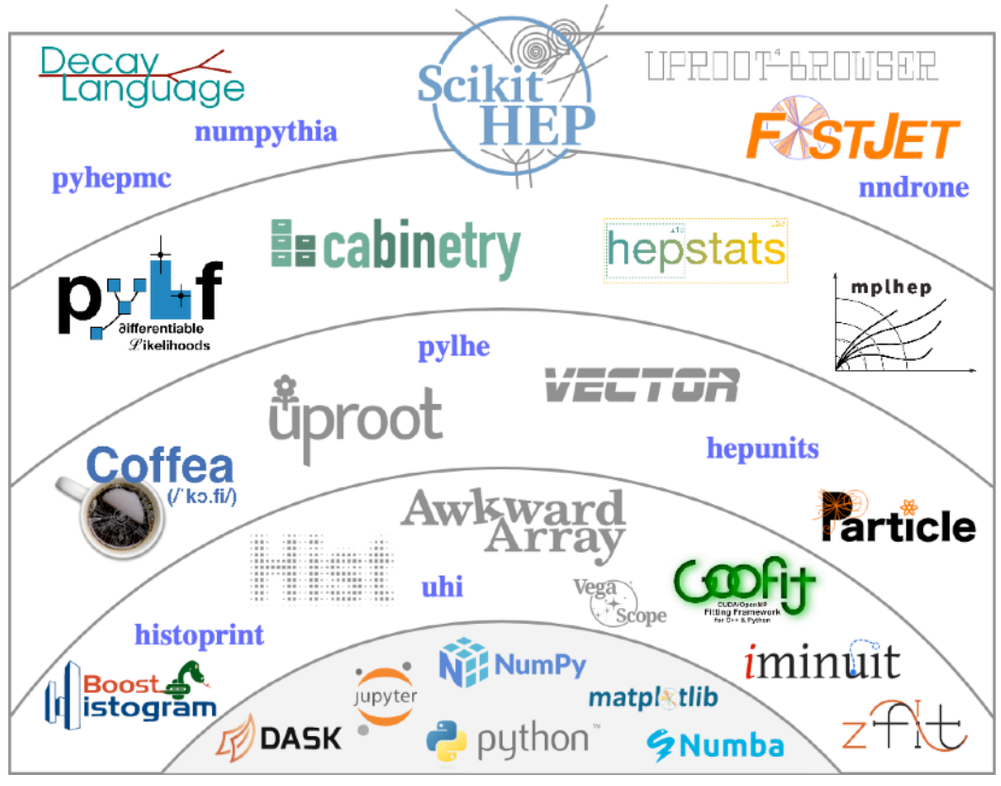
- A user interface to *columnar analysis*
  - Optimized array programming kernels build an **expressive *and* performant** language
  - Seamless integration with ML tools due to shared interface
- An incubator for rapid prototyping
  - Fill in missing pieces of ecosystem
  - Good abstractions are factored out
- A minimum viable product
  - Already used in several CMS publications
  - In use by ATLAS, ProtoDUNE collaborators
  - Early feedback builds ecosystem roadmap
    - Vibrant contributor community



Fermilab

Coffea project

- A user interface to *columnar analysis*
  - Optimized array programming kernels build an **expressive *and* performant** language
  - Seamless integration with ML tools due to shared interface
- An incubator for rapid prototyping
  - Fill in missing pieces of ecosystem
  - Good abstractions are factored out
- A minimum viable product
  - Already used in several CMS publications
  - In use by ATLAS, ProtoDUNE collaborators
  - Early feedback builds ecosystem roadmap
    - Vibrant contributor community

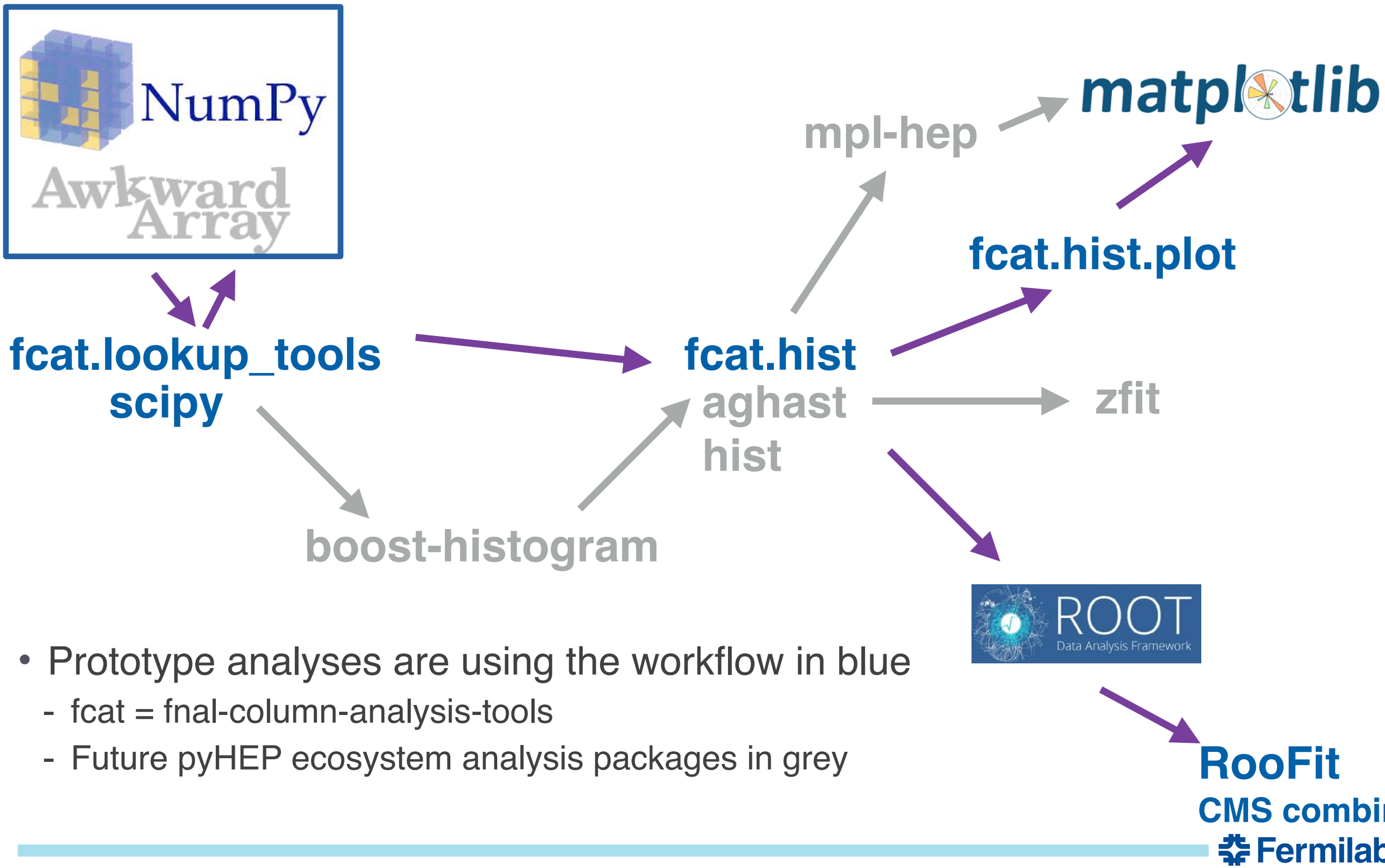


We might be in the business of putting ourselves out of business

Fermilab

Coffea in 2019 (*HOW* we started)

ncsmith-how2019-columnar

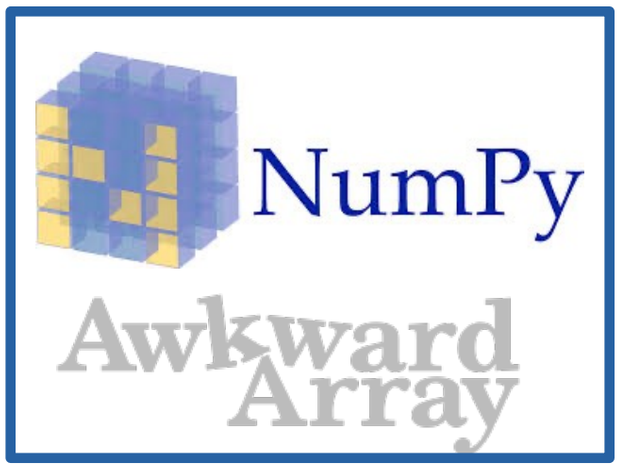


- Prototype analyses are using the workflow in blue
  - fcat = fnal-column-analysis-tools
  - Future pyHEP ecosystem analysis packages in grey

Fermilab

Coffea in 2019 (*HOW* we started)

ncsmith-how2019-columnar



fcat.lookup_tools

scipy

boost-h

mpl-hep

matplotlib

- Coffea lookup_tools allowed CMS publications to happen
  - Correctionlib abstracts
- Scipy usage is mostly a training issue
  - Consider parton, same exact spline as LHAPDF
- Users unsure how to glue torch/triton/etc.
  - Some boilerplate: coffea.ml_tools
- Object (e.g. 4-vector) façade missing from this diagram!
  - PyHEP 2020 NanoEvents demo youtube
  - Coffea to scikit-hep/vector transition 🚧

- Prototype analyses are using the workflow in blue
  - fcat = fnal-column-analysis-tools
  - Future pyHEP ecosystem analysis packages in grey

RooFit

CMS combine

Fermilab

Coffea in 2019 (*HOW* we started)

ncsmith-how2019-columnar

IRIS-HEP topical: new histogram tools

- Most mpl-hep wishes came true
  - Upstreamed `ax.stairs`
  - Style sheets for all 4 LHC exp.
  - Convenient 1D & 2D APIs for *pre-binned data*
- Boost-histogram & hist well-established
  - coffea.hist deprecated
- aghast
  - UHI *protocol* solves inter-op
  - Serialization: BH 🚧

mpl-hep → matplotlib

fcat.hist.plot

fcat.hist

aghast → zfit

hist

ROOT Data Analysis Framework

RooFit

CMS combine

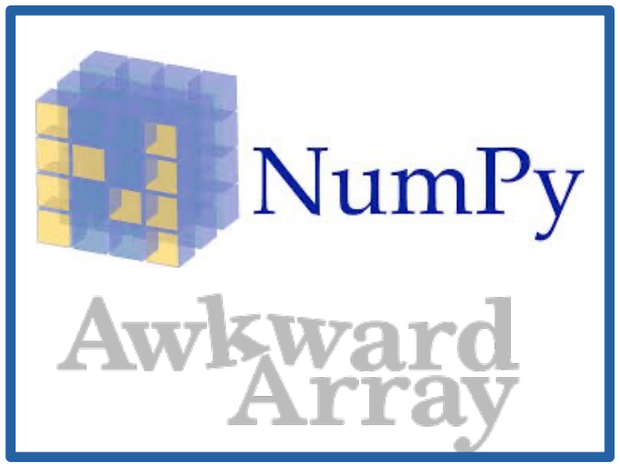
e workflow in blue

ackages in grey

Fermilab

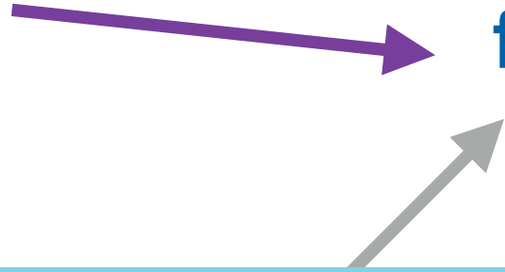
Coffea in 2019 (*HOW* we started)

ncsmith-how2019-columnar



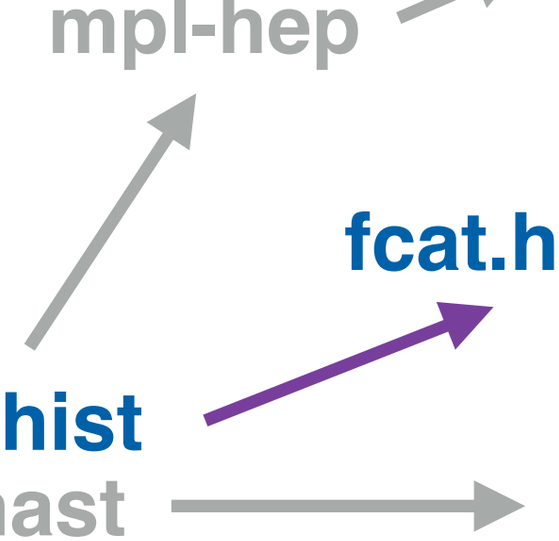
fcat.lookup_tools

scipy



fcat.hist

aghast
hist



mpl-hep

matplotlib

fcat.hist.plot

zfit

- Statistical tools: many options
  - RooFit, GooFit, zfit, pyhf, …
  - Coffea is CMS-heavy = combine
- Separate model-building from inference?
  - a la RooFit/RooStats

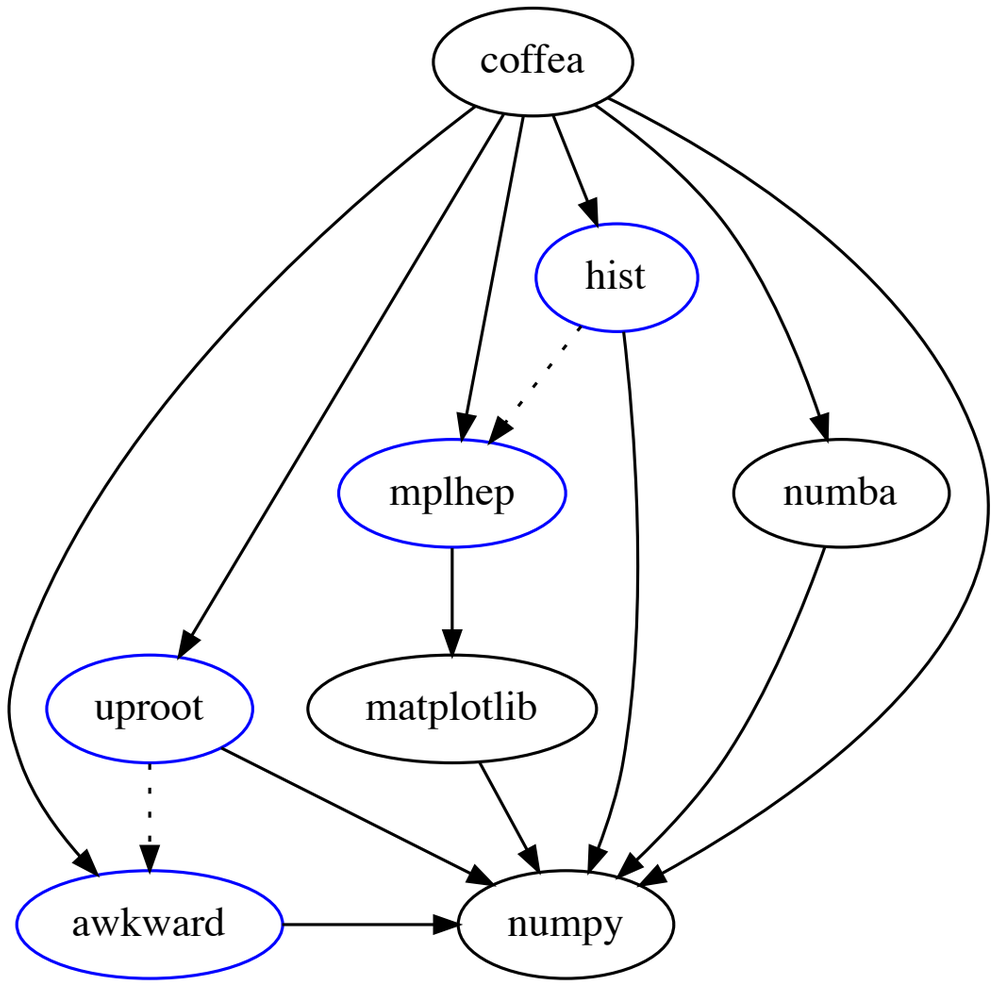
RooFit

CMS combine

Fermilab

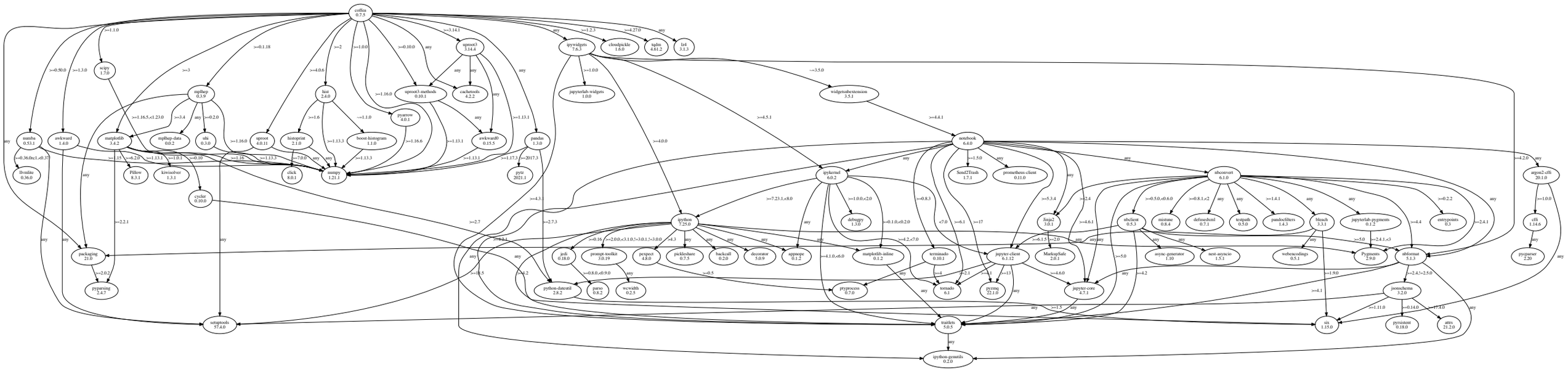
Coffea in 2021

- Accurate but abridged dependencies
  - Blue is scikit-hep



Coffea in 2021

- Accurate ~~but abridged~~ dependencies
  - *We live in a society*



Fermilab

Transitions

- Awkward 0.x ➜ 1.x
  - Oof
- coffea.hist ➜ hist
  - Smoother, 1-1 rosetta helped
- Awkward 1.x ➜ 2.x
  - Eager-mode: seamless it seems
  - ak.virtual to dask-awkward: to be seen IMO
- coffea.nanoevents.methods.vector ➜ vector
  - Really overdue
- coffea.processor ➜ ?

Death to processors?

- Dask obviates much of processor, the rest needs a long-term home
- NanoEvents object façade
- Dataset mangling tools
  - Though we didn't do much here to begin with
- Coffea accumulators
  - Move to hist? Dask-histogram knows tree reduction

```
22      class Addable(Protocol):
23          def __add__(self: T, other: T) -> T:
24              ...
25
26
27      Accumulatable = Union[Addable, MutableSet, MutableMapping]
28
29
30      def add(a: Accumulatable, b: Accumulatable) -> Accumulatable:
```

Key directions for me this week

- Good APIs / protocols for interoperability attract users
  - Lateral movement, tool discovery
  - Examples:
    - `hist.logpdf(data: Hist, model: ImplementsCDFProtocol) -> Callable`
      - Or better to set goal: template fraction fit in two lines?
    - Use `particle` for pdgId repr in NanoEvents

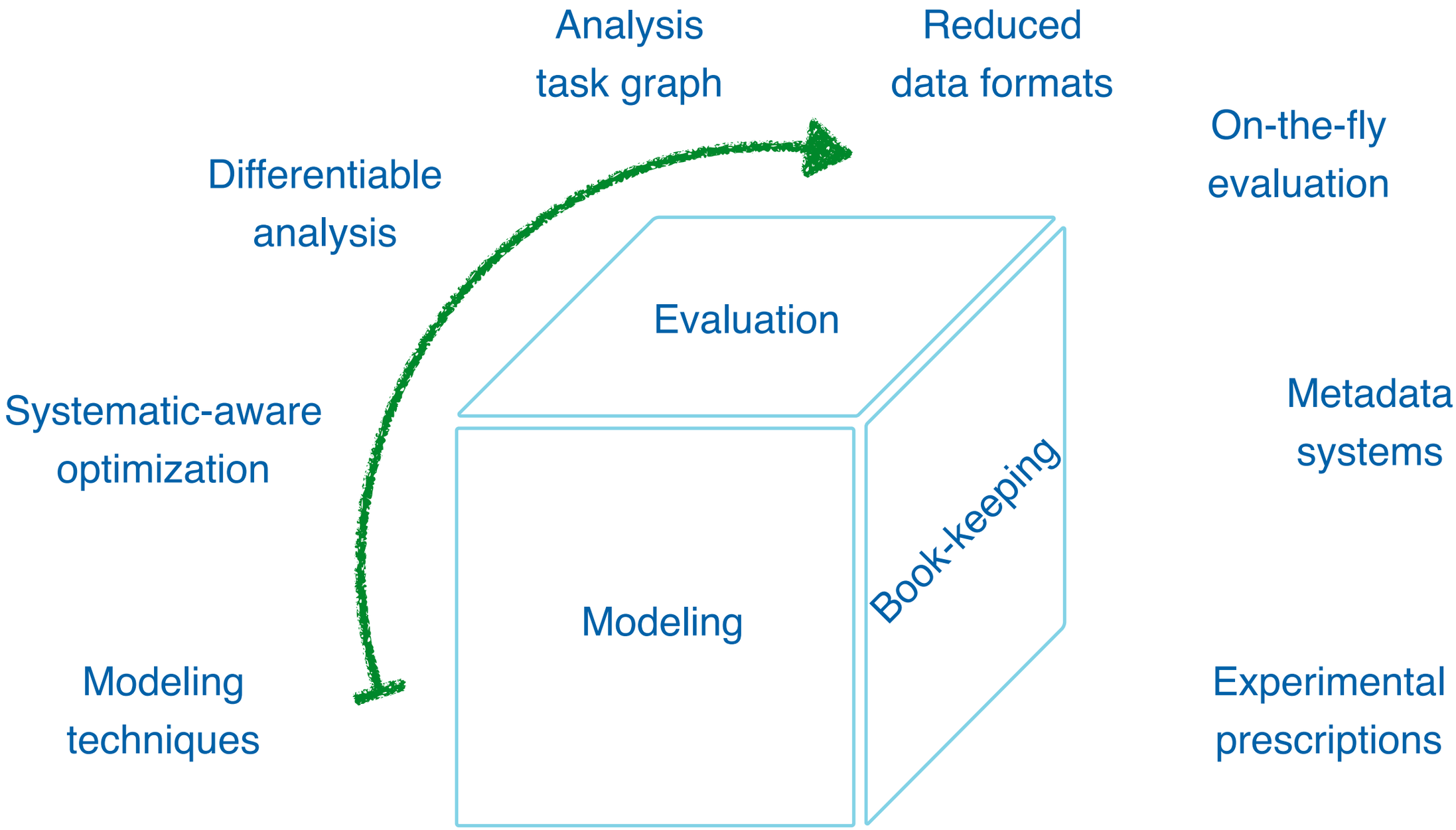
Key directions for me this week

- Good APIs / protocols for interoperability attract users
  - Lateral movement, tool discovery
  - Examples:
    - `hist.logpdf(data: Hist, model: ImplementsCDFProtocol) -> Callable`
      - Or better to set goal: template fraction fit in two lines?
    - Use `particle` for pdgId repr in NanoEvents
- It's all about data delivery
  - Reliable xrootd: uproot-fsspec project
  - Task graph enables virtual data
    - Columnservice 2.0 / S3 / etc.

Key directions for me this week

- Good APIs / protocols for interoperability attract users
  - Lateral movement, tool discovery
  - Examples:
    - `hist.logpdf(data: Hist, model: ImplementsCDFProtocol) -> Callable`
      - Or better to set goal: template fraction fit in two lines?
    - Use `particle` for pdgId repr in NanoEvents
- It's all about data delivery
  - Reliable xrootd: uproot-fsspec project
  - Task graph enables virtual data
    - Columnservice 2.0 / S3 / etc.
- Statistics is a cool hobby made less fun by systematics
  - HS3, jaxfit
  - CMS Higgs Combination workspace: 10 GB RAM, 30h to minimize
    - Help!
  - API for systematics?

Topics in systematics that I won't have time for, probably

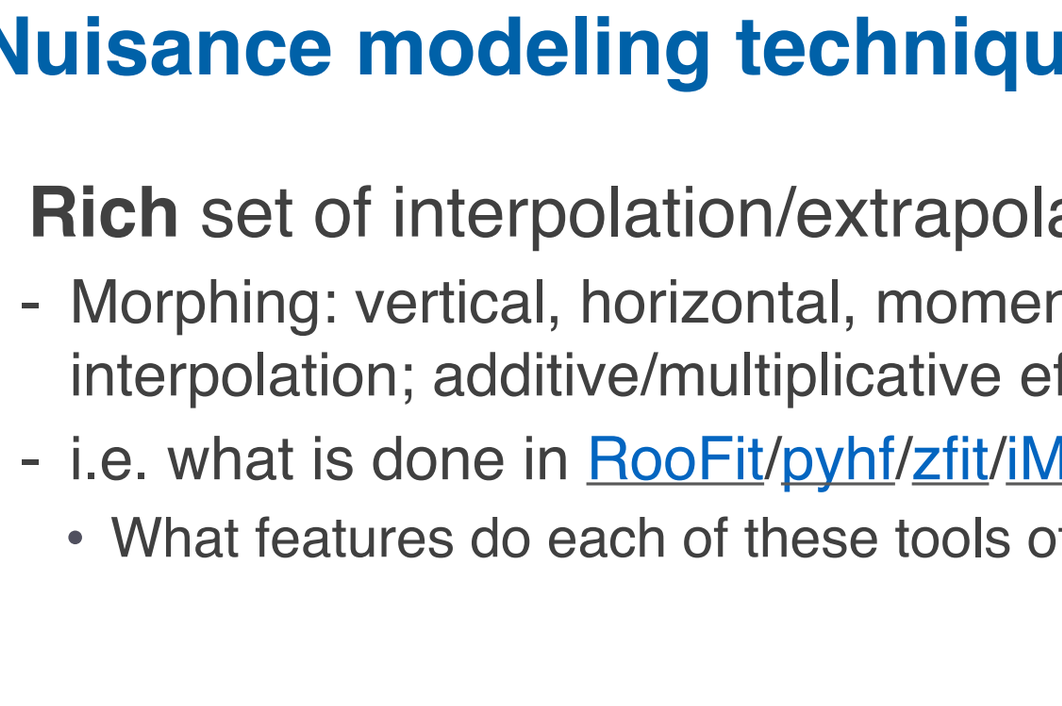


Fermilab

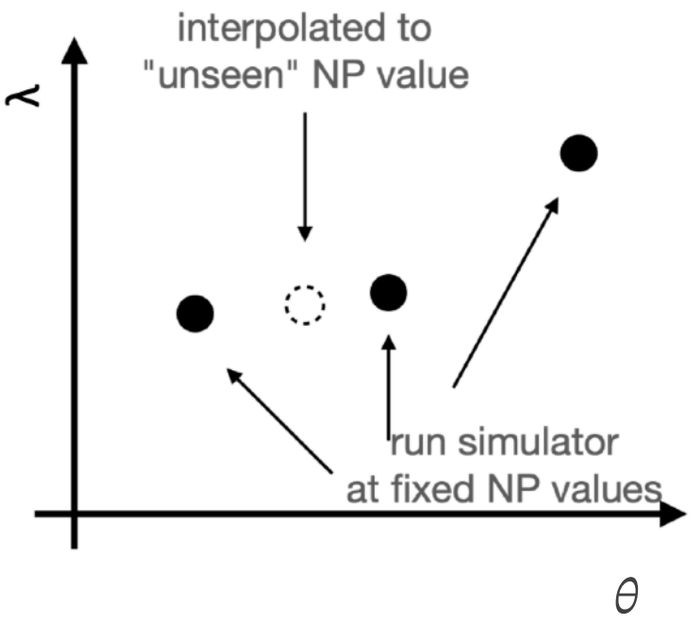
Nuisance modeling techniques

- **Rich** set of interpolation/extrapolation techniques at end-stage
  - Morphing: vertical, horizontal, moment; splines; gaussian process; asymmetric shift interpolation; additive/multiplicative effects; MC stat uncertainty, BB-lite; …
  - i.e. what is done in RooFit/pyhf/zfit/iMinuit/combine/etc.
    - What features do each of these tools offer? Nobody has it all!

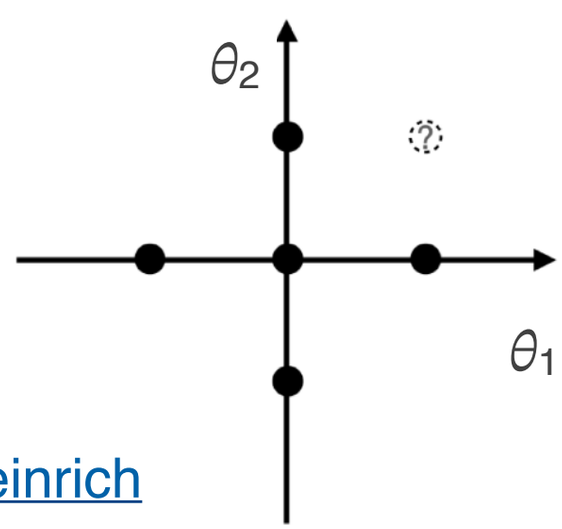
Visualization of bin-by-bin linear interpolation of distribution



W. Verkerke



Combining effects



L. Heinrich

Fermilab

Nuisance modeling techniques

- Simpler taxonomy of techniques to get inputs to fitting tools?
  - This is the dominant analysis-stage computation expense (process billions of events)
- Posit three basic techniques
  - I think all of these can be done unbinned as well
  - Just need functions w(x,θ) and Δ(x,θ)

$$\int_{\text{bin}} P(x)\mathrm{d}x \approx \frac{\sigma}{N} \sum_{x_i \sim P(x)}^{N} 1(x_i \in \text{bin})$$

(nominal)

$$\int_{\text{bin}} P(x|\theta=\theta_1)\mathrm{d}x \approx \frac{\sigma}{N} \sum_{x_i \sim P(x|\theta=\theta_1)}^{N} 1(x_i \in \text{bin})$$

(alternative sample, e.g. 2-point)

$$\int_{\text{bin}} P(x|\theta=\theta_1)\mathrm{d}x \approx \frac{\sigma}{N} \sum_{x_i \sim P(x)}^{N} w(x_i, \theta=\theta_1) 1(x_i \in \text{bin})$$

(reweight, e.g. efficiency)

$$\int_{\text{bin}} P(x|\theta=\theta_1)\mathrm{d}x \approx \frac{\sigma}{N} \sum_{x_i \sim P(x)}^{N} 1(x_i + \Delta(x_i, \theta=\theta_1) \in \text{bin})$$

(shift, e.g. energy scale)