Corralling Heterogeneous Systems (and users)

G. Watts/UW Seattle
S2I2 Princeton Workshop

2017-05-02

Competing Interests

Production
(MC Generation, Reconstruction, Skimming, Production Ntuple Making, etc.)
Big Teams

Analysis
(Spin on AOD's, plotting, ML, statistics, etc.)
Small Team

Competing Interests

Production
(MC Generation, Reconstruction, Skimming, Production Ntuple Making, etc.)
Big Teams

Analysis
(Spin on AOD's, plotting, ML, statistics, etc.)
Small Team

Different tools for different stages

Competing Interests

Production
(MC Generation, Reconstruction, Skimming, Production Ntuple Making, etc.)
Big Teams

Analysis
(Spin on AOD's, plotting, ML, statistics, etc.)
Small Team

Some common things – like the data-model (ROOT)

Which can define a great deal of how an analyzer interacts with the system

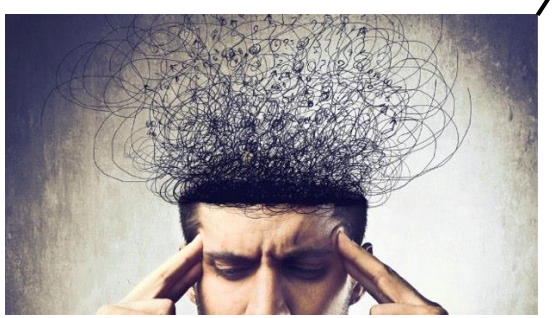
Who is the tail? Who is the dog?

Analyzer: Publish Paper with measurement or search

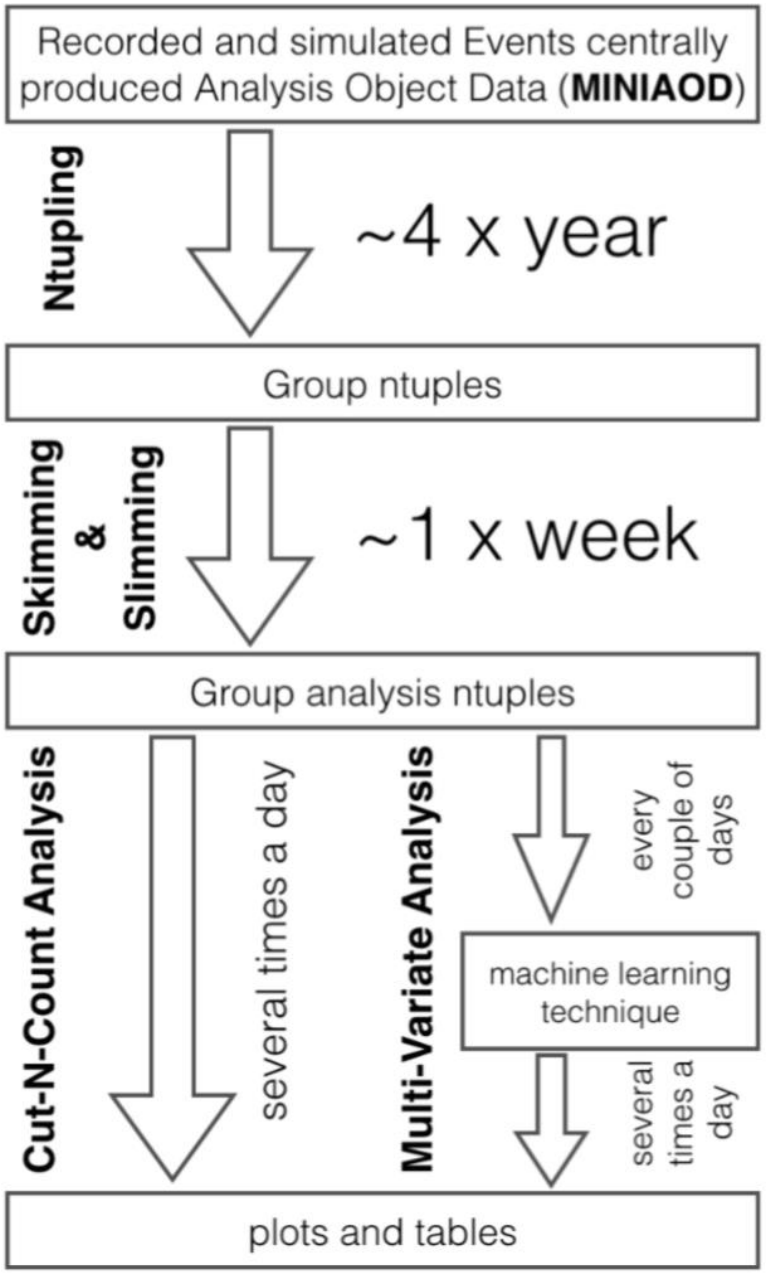
Do not care how, as long as it is right!

Environment:

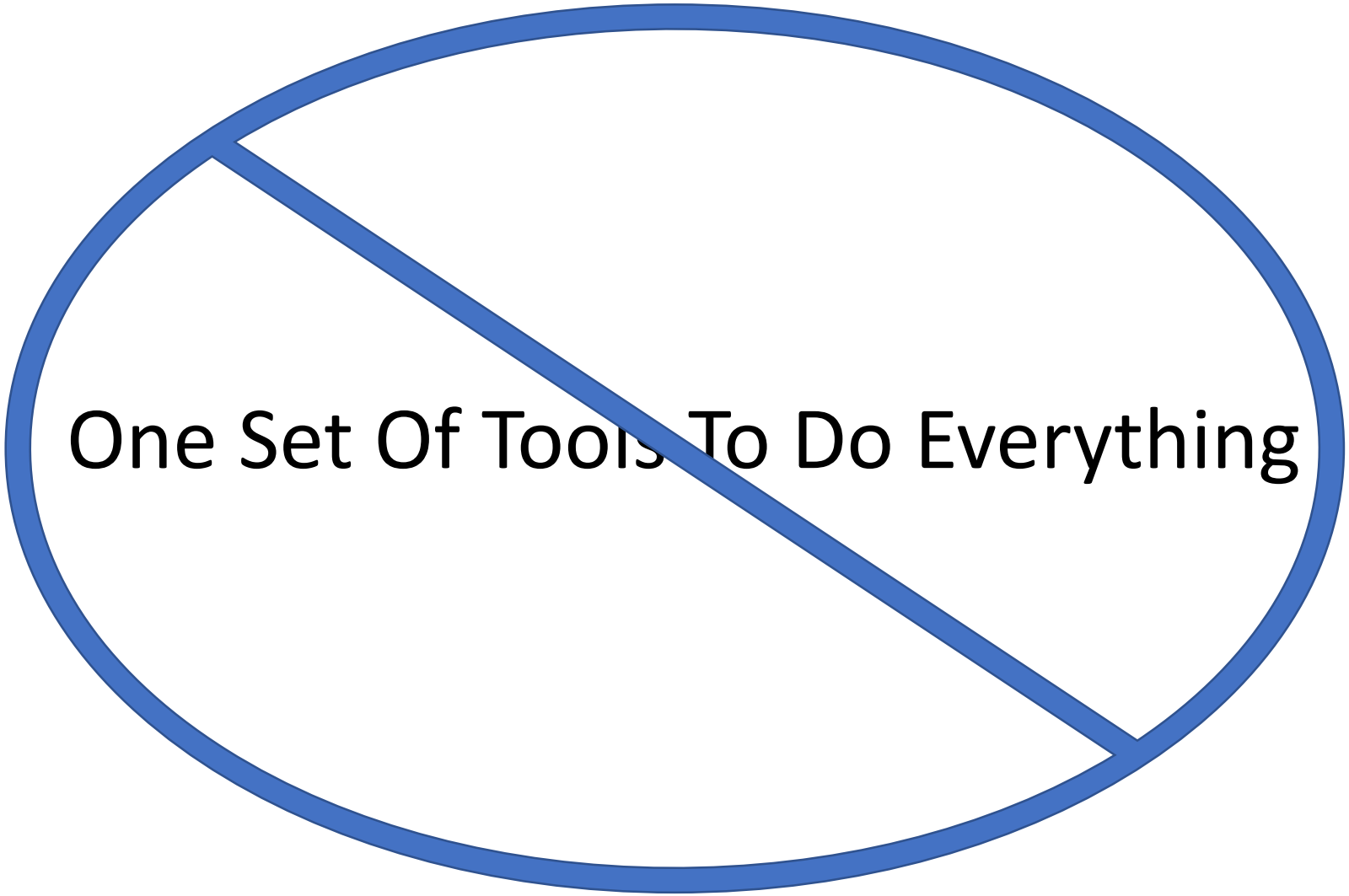
- Increasingly heterogeneous/chaotic system
- Distributed analysis
- Laptop analysis
- Large/small datasets



(Analyzer)

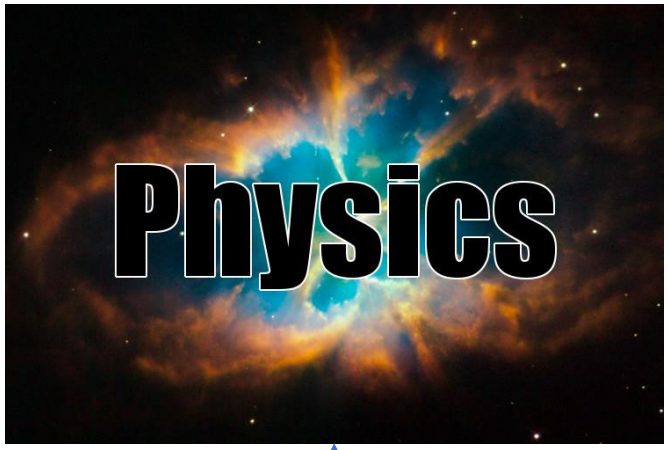


“Hey – can any of your analysis run on a super computer?”


One Set Of Tools To Do Everything

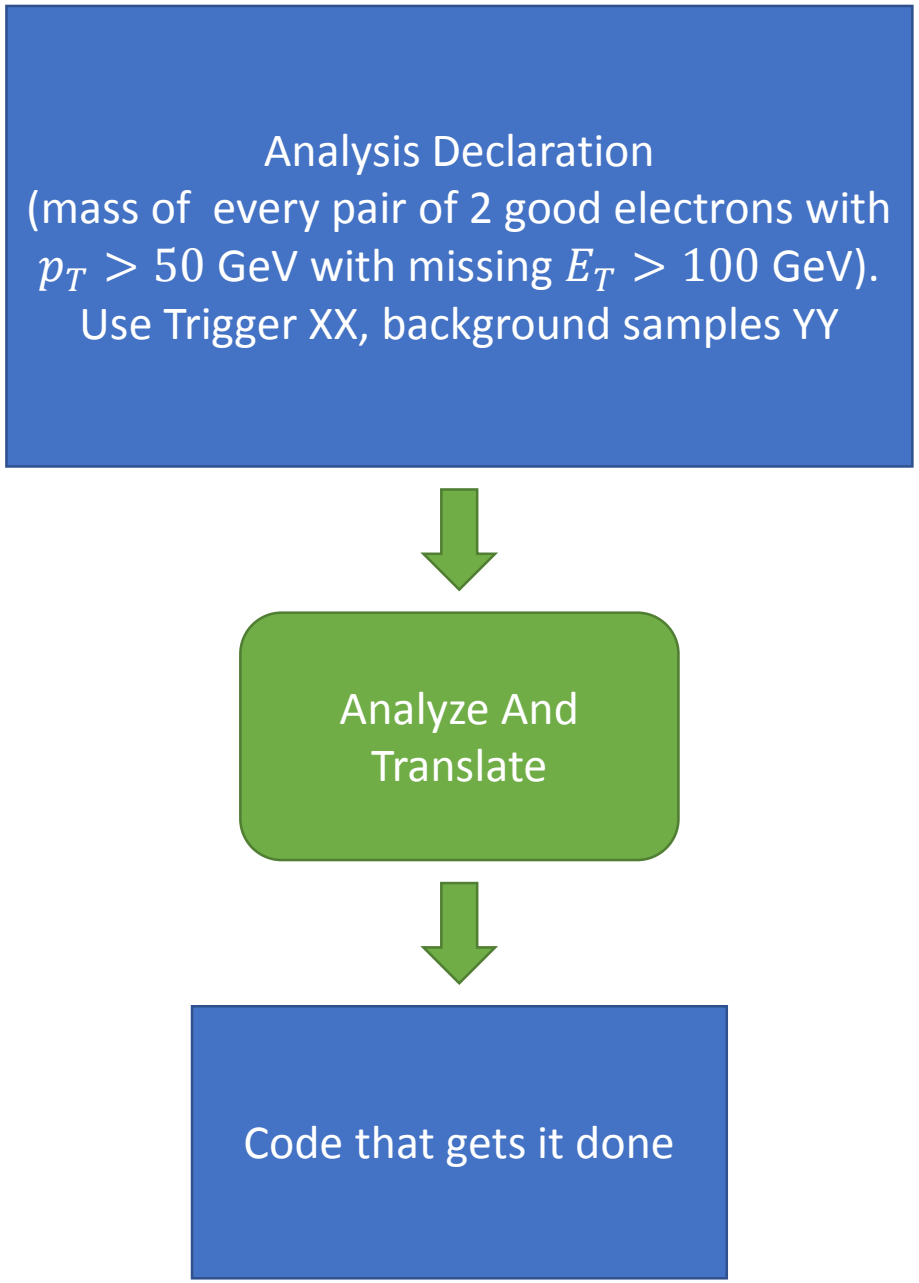
Even how you implement your analysis will change between environments!

And you might want parts of your analysis to run in different places!



This distance is getting too large for graduate students and postdocs!

justgetitdone



Separation of concerns

Concentrate on common tools for each platform

Keep physics near physics

I have an idea of how to develop this sort of thing… But…

/cvmfs/atlas.cern.ch/repo/ATLASLocalRootBase/x86_64/root/5.34.14-x86_64-slc6-gcc4.7/include/TMVA/Factory.h:134: **undefined reference** to `TMVA::Factory::AddTree(TTree*, TString const&, double, TCut const&, TMVA::Types::ETreeType)'

ATLAS Protected environment

In the wild environment

Abstractions always leak

Text Based (check in to a source control system, can do differencing)
Take advantage of tools that exist in the larger world (ML)
Corral data in some sort of a reasonable way
Support interactive exploration

TSelector

1000’s of lines of auto generated code

DataFrame

5 or 6 lines…