Data Analysis – Algorithms and Tools (Summary of Tack 2)

16th International workshop on Advanced Computing and Analysis Techniques in physics research – ACAT 2014

Martin Spousta

Charles University in Prague

… Bridging the disciplines

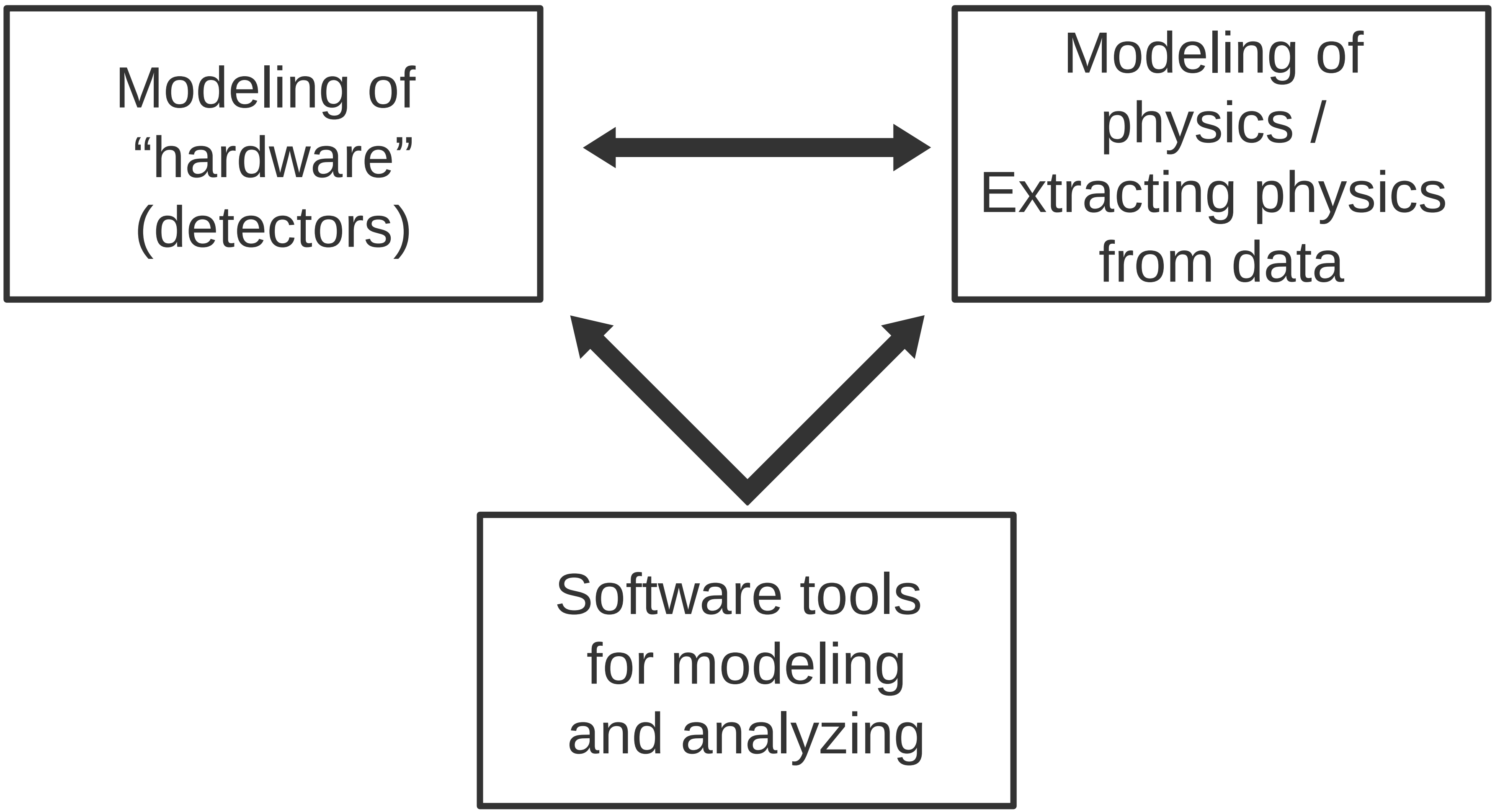
Disciplines

Modeling of “hardware” (detectors)

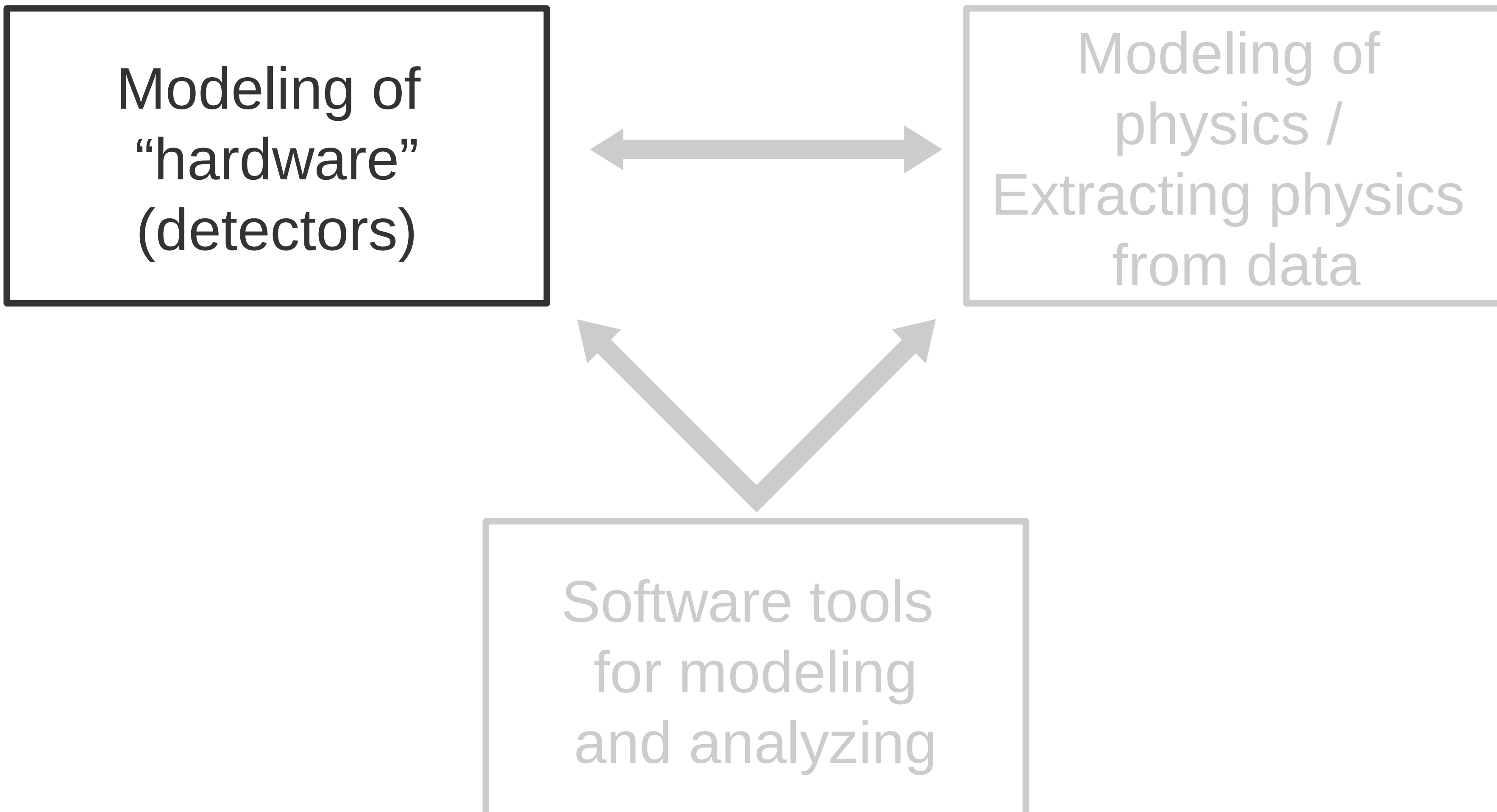
Modeling of physics / Extracting physics from data

Software tools for modeling and analyzing

Disciplines



Disciplines

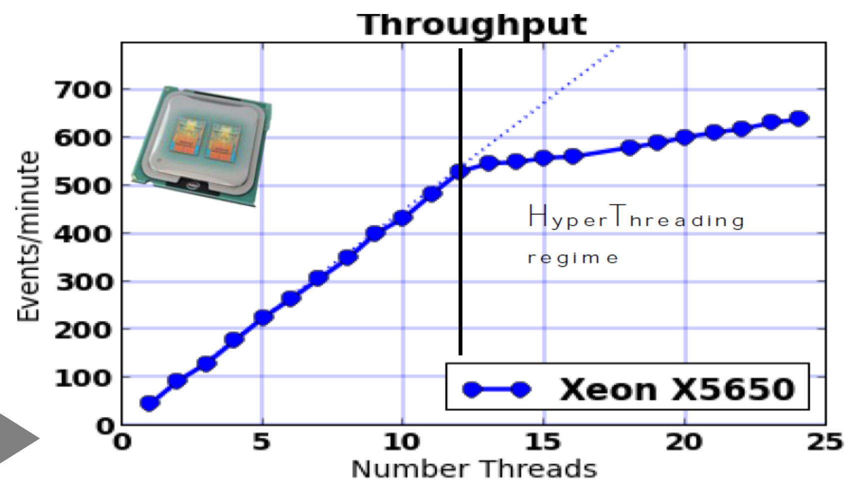


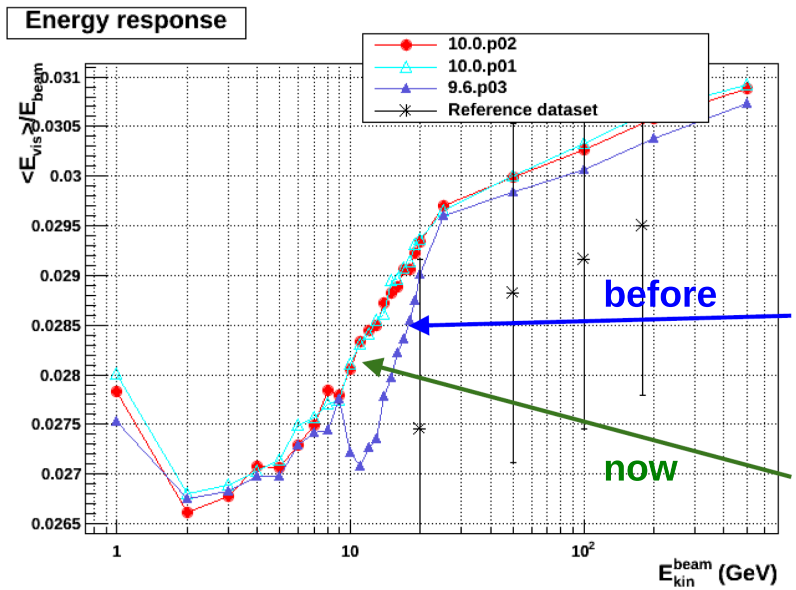
Geant 4

John Apostolakis

Geant4 9.4: **LHC run I, ATLAS, CMS**
Geant4 9.6: **ATLAS run II**
Geant4 10.0: **CMS run II**

- Toolkit for simulation of passage of particles through matter.
- New release 10.0. Updates in design:
  - multi-threading
  - strong reproducibility of events (simulation independent of history of previous running)
- Updates in physics modeling:
  - corrected energy response in QGSP_BERT + improving lateral shower shape
  - improved FTF (enable anti-nucleons)
  - new unique solid (USolid) library
  - improved radioactive decay

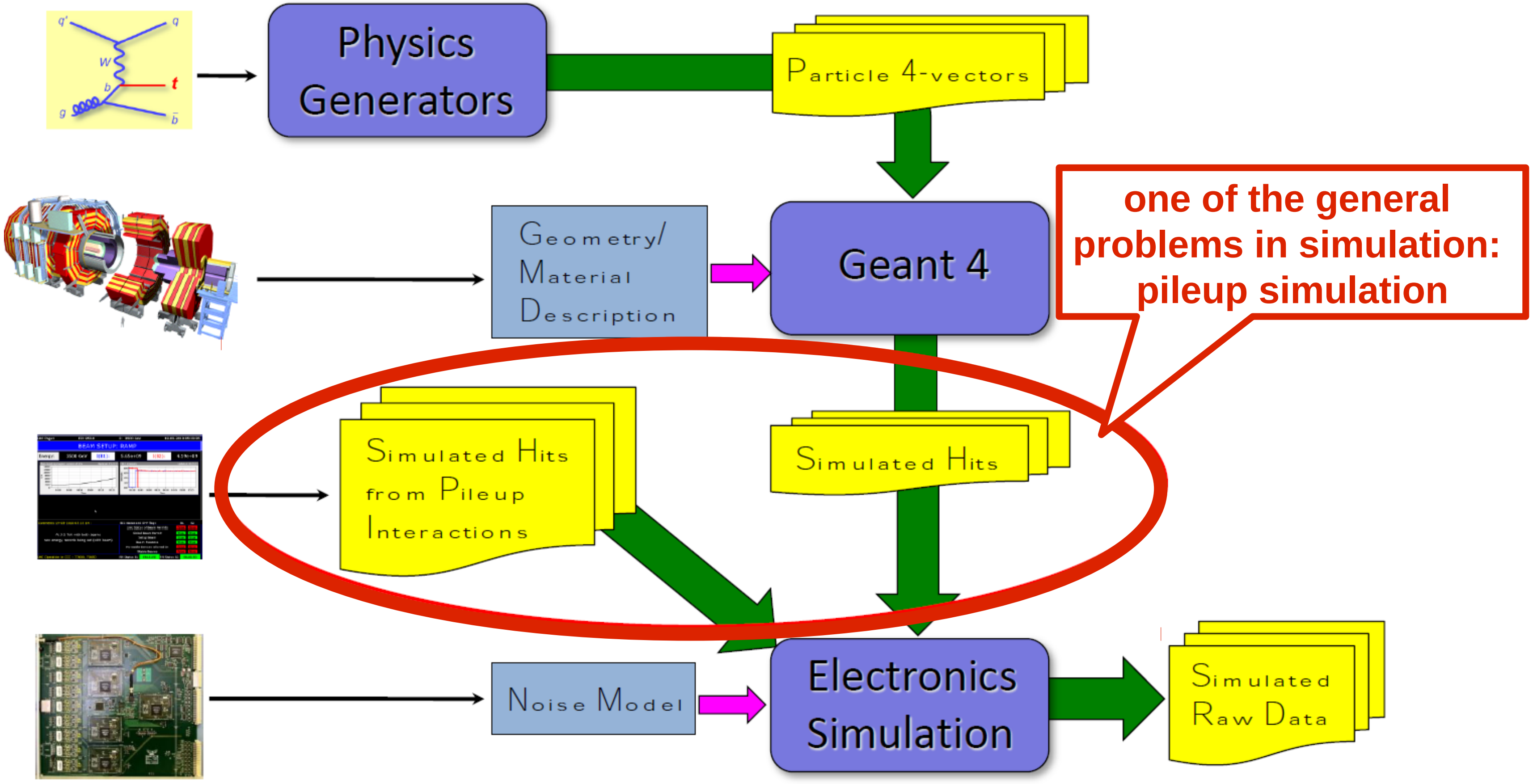




CMS Simulation Upgrade

David Lange

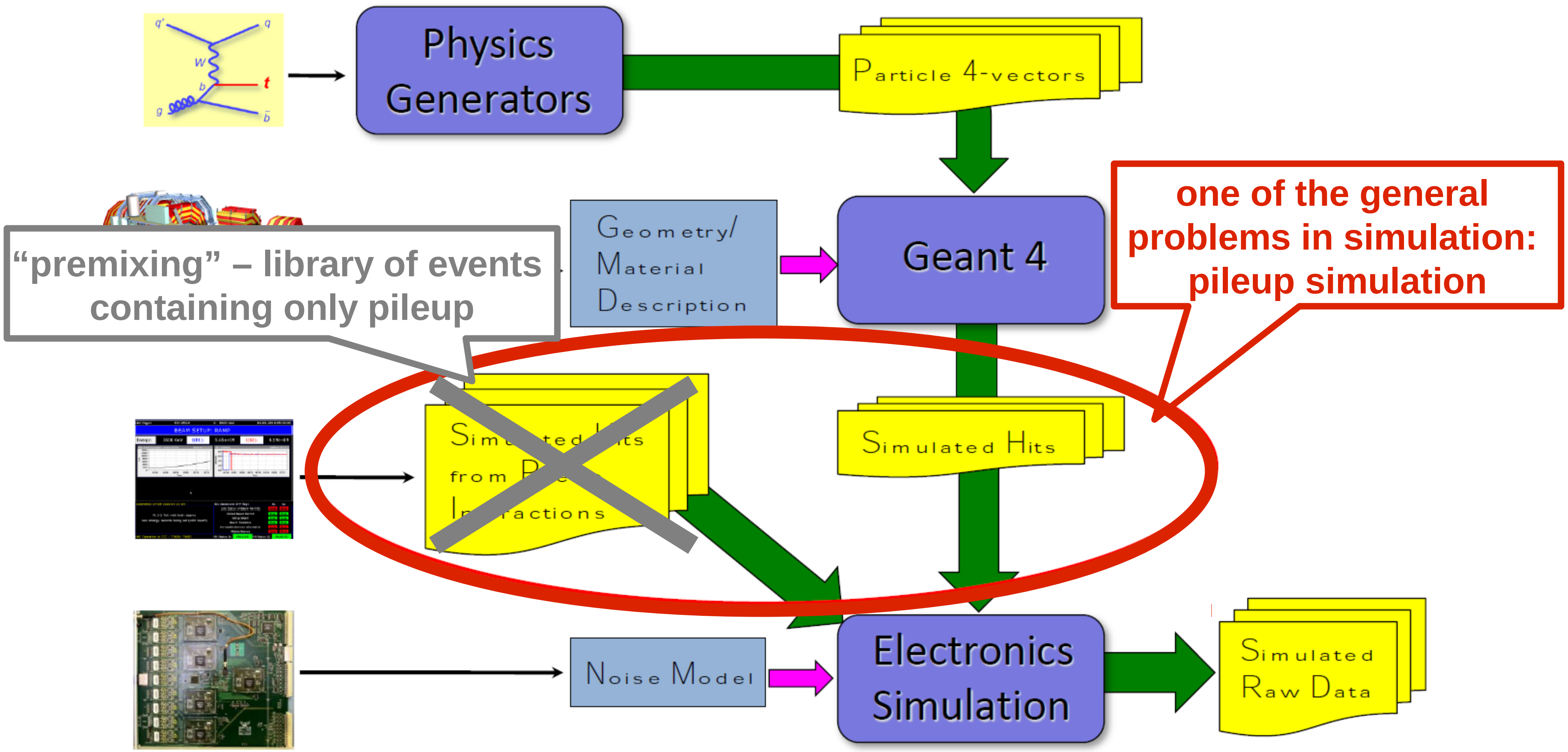
Geant4 9.4: LHC run I, ATLAS, CMS
Geant4 9.6: ATLAS run II
Geant4 10.0: CMS run II



CMS Simulation Upgrade

David Lange

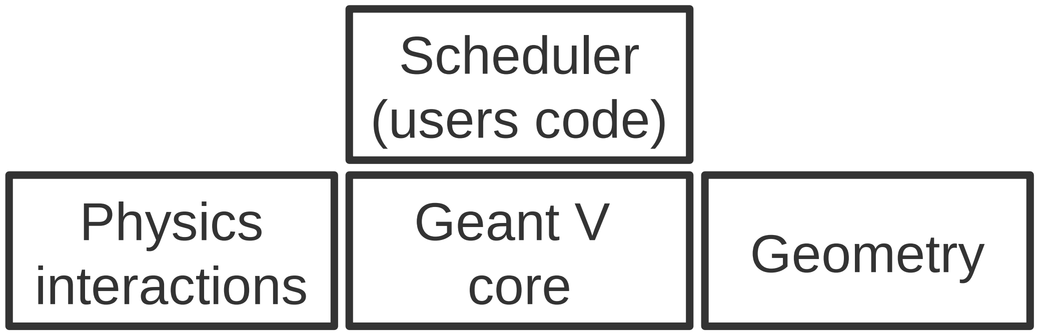
Geant4 9.4: LHC run I, ATLAS, CMS
Geant4 9.6: ATLAS run II
Geant4 10.0: CMS run II



Geant V

Andrei Gheata

- Motivation: speed up the particle transport simulation. Most simulation time spent on a few percents of volume.



… each of these component is a subject of R&D

- Features of new Geant:
  - Vectorization and locality (not a single track but a group of tracks are transported)
  - Multi-threading
  - Adding new entities to control the work flow
  - Optimizing simulation geometry
  - Possibility to include some fast simulation

Fast simulation

Andrei Gheata

- CPU in Run I dominated by MC productions
  - Geant4 robustness = success of Run I, but many physics analyses suffer from the lack of MC statistics
  - Fast simulation boosted some MC samples in Run I and will be indispensable in Run II
- Fast simulation = use parameterizations of the response or pre-generated samples
  - Generic: PGS, Delphes
  - Experiment specific: FastSim, Atlfast, ...
- Alternative approaches:
  - replace costly physics objects by pre-clustered ones (e.g. frozen showers)
  - filtering and selective parameterizations, fast tracking
  - ...

Sample	Full G4	Fast G4	Atlfast2
Minimum bias	551	246	31.2
$t\bar{t}$	1990	757	101
Jets	2640	832	93.6
Photons + jets	2850	639	71.4
$W^{\pm} \to e^{\pm}\nu_e$	1150	447	55.1
$W^{\pm} \to \mu^{\pm}\nu_{\mu}$	1030	438	57.0
Heavy ion	56k	21.7k	3050

Simulation times in kSI2K seconds

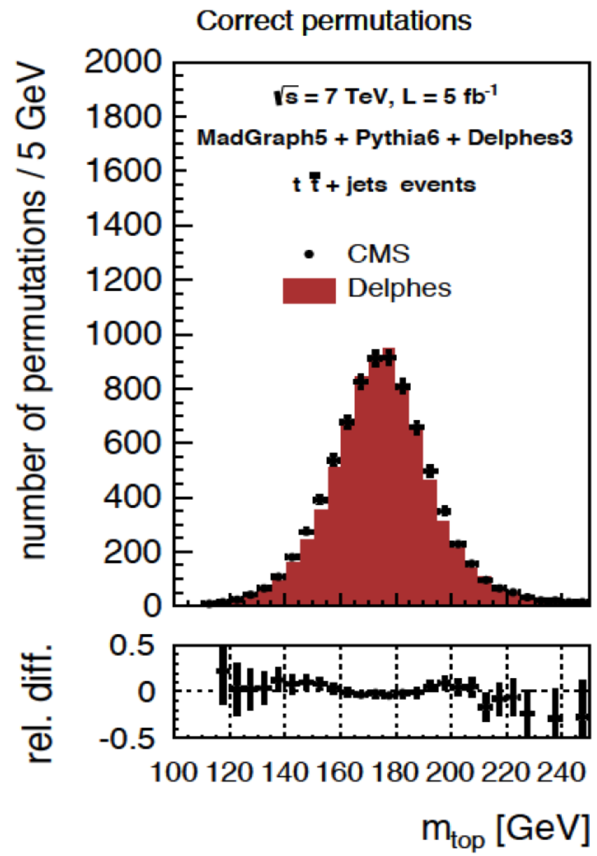
- Digitization and tracking = bottlenecks of fast simulation => generally, need a combination of both, fast simulation and full simulation

Delphes 3

Alexandre Mertens

- Tool for fast detector simulation based on parameterizations
- Since 2007:
  - community based, modularity, interfaces (e.g. MadGraph or HepMC, … FastJet, ROOT output)
  - used by phenomenologists (Snowmass)
- Some attractive features:
  - particle flow algorithm
  - pile-up
  - propagation in magnetic field, calorimeters
  - b-tagging, jet substructure

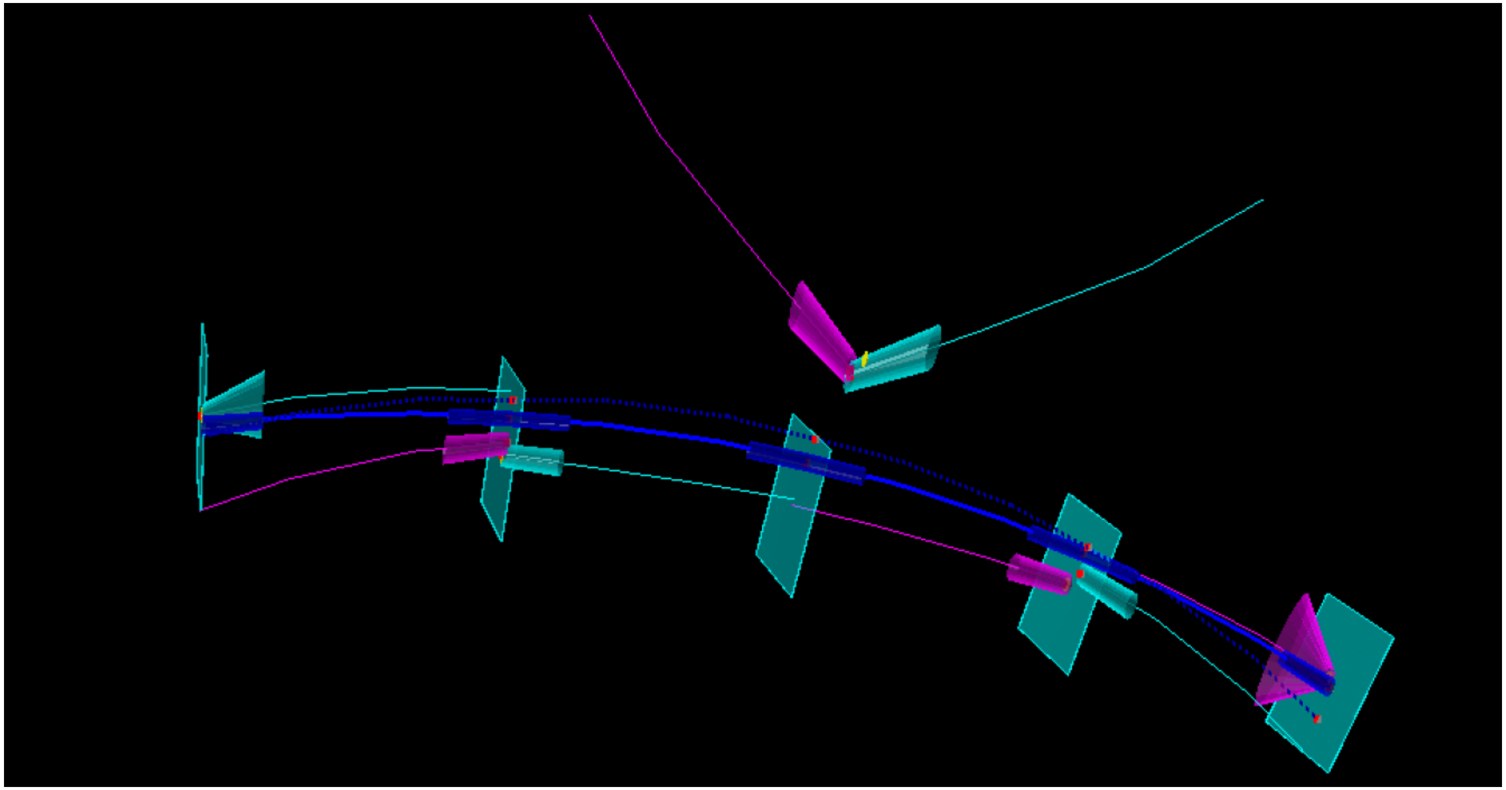
Full simulation (G4):	**10-100 s/ev**
Fastsim of ATLAS, CMS:	**1s/ev**
Delphes:	**10 ms/ev**



- Can mimic ATLAS and CMS
- Only software allowing to simulate realistic HL-LHC environment

Johannes Rauch

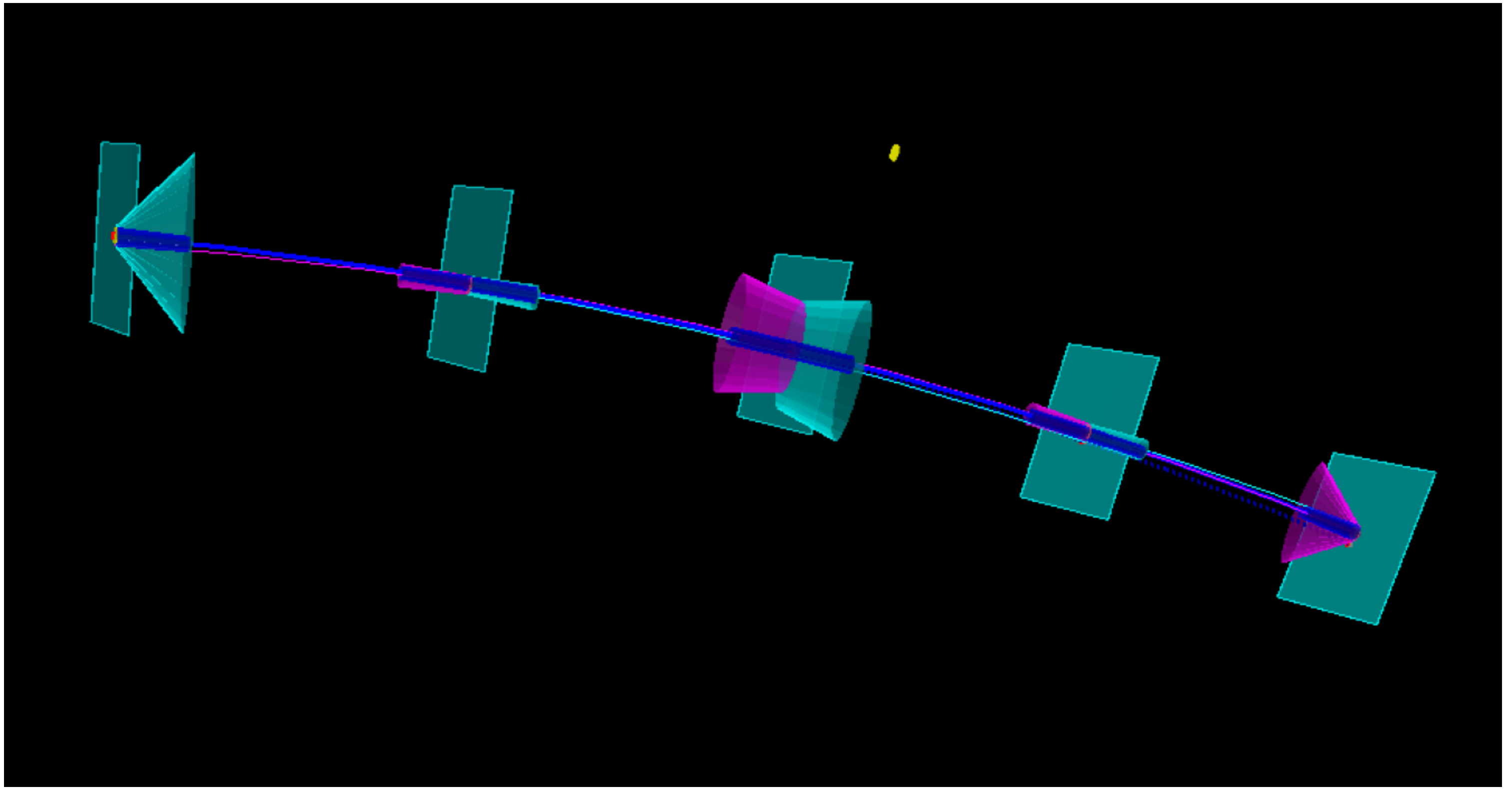
Genfit



- Generic Track-Fitting Toolkit
  - Modular track fitting framework and simulation of a tracker
  - Can use Standard Kalman Fitter (linearizing around predictions) or improve using Deterministic Annealing Filter (–> better handling of out-layers)

Genfit

Johannes Rauch

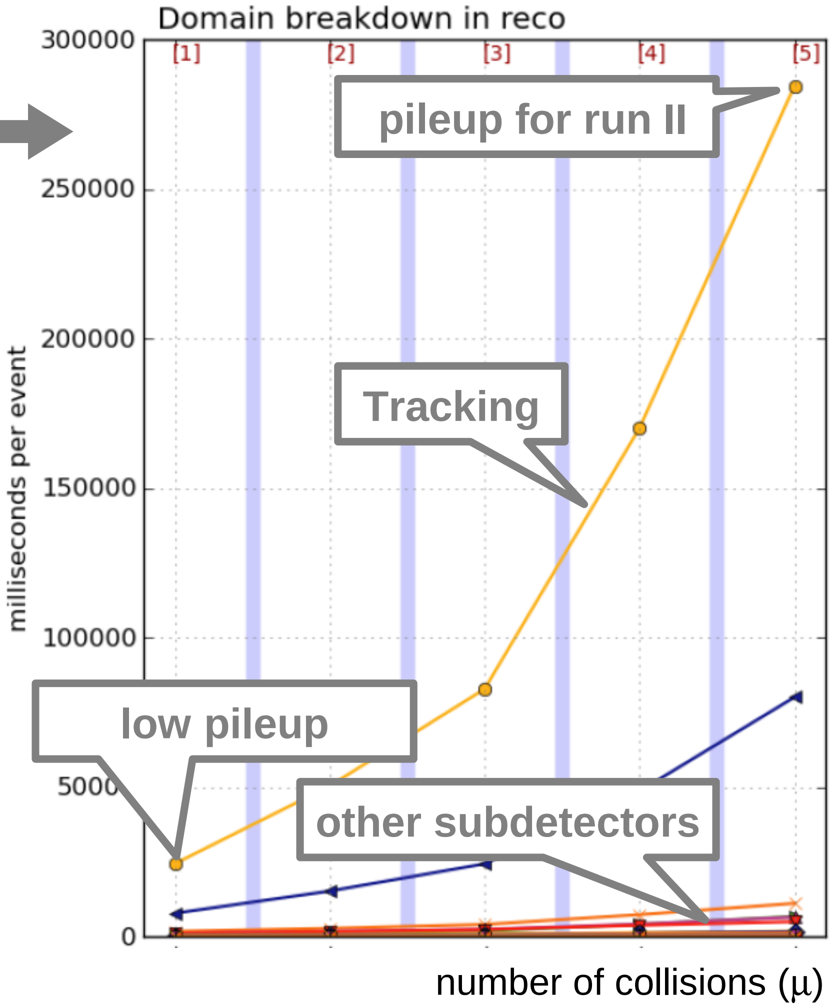


- Open source (http://sourceforge.net/projects/genfit)
- May be a very useful tool for experiments to be built and already in use (e.g. Belle II, PANDA)
- Questionable speed

New ATLAS Tracking

Nicholas Styles

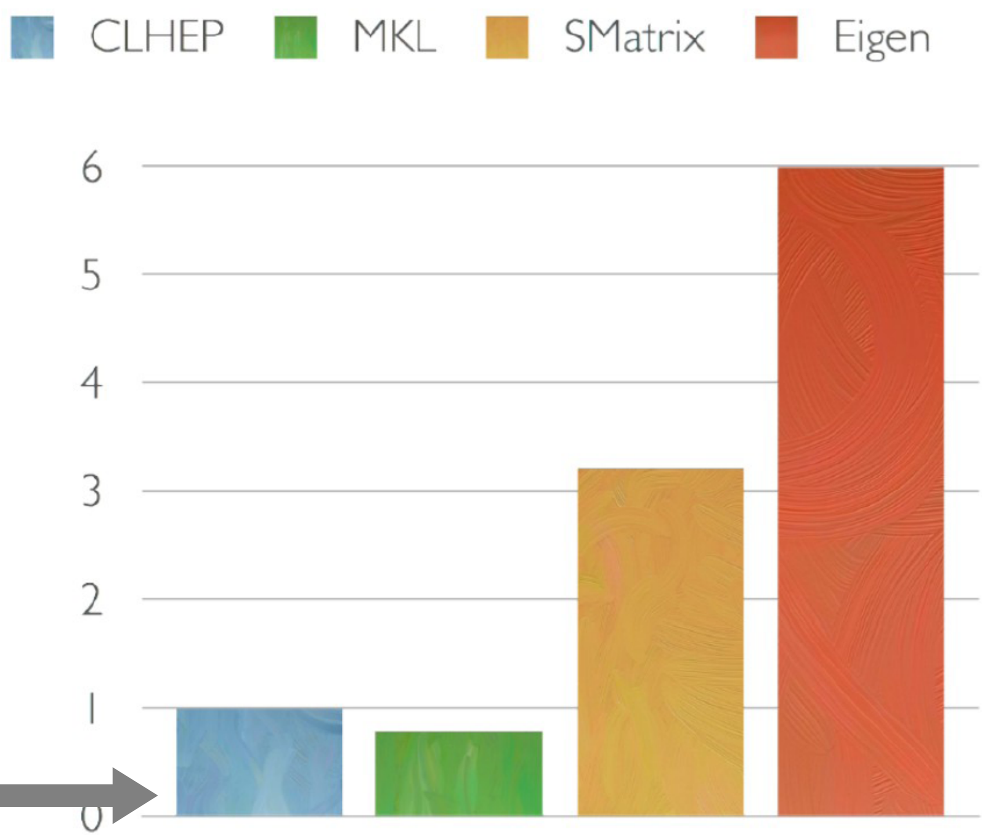
- New hardware environment:
  - high pile-up during LHC run II
  - IBL (insertable b-layer) = additional layer
- Features:
  - move to templeted class design
  - use curvilinear coordinates
  - better handling of information on covariances
- Software design:
  - move to new data format xAOD
  - CLHEP replaced by Eigen



New ATLAS Tracking

Nicholas Styles

- New hardware environment:
  - high pile-up during LHC run II
  - IBL (insertable b-layer) = additional layer
- Features:
  - move to templeted class design
  - use curvilinear coordinates
  - better handling of information on covariances
- Software design:
  - move to new data format xAOD
  - CLHEP replaced by Eigen

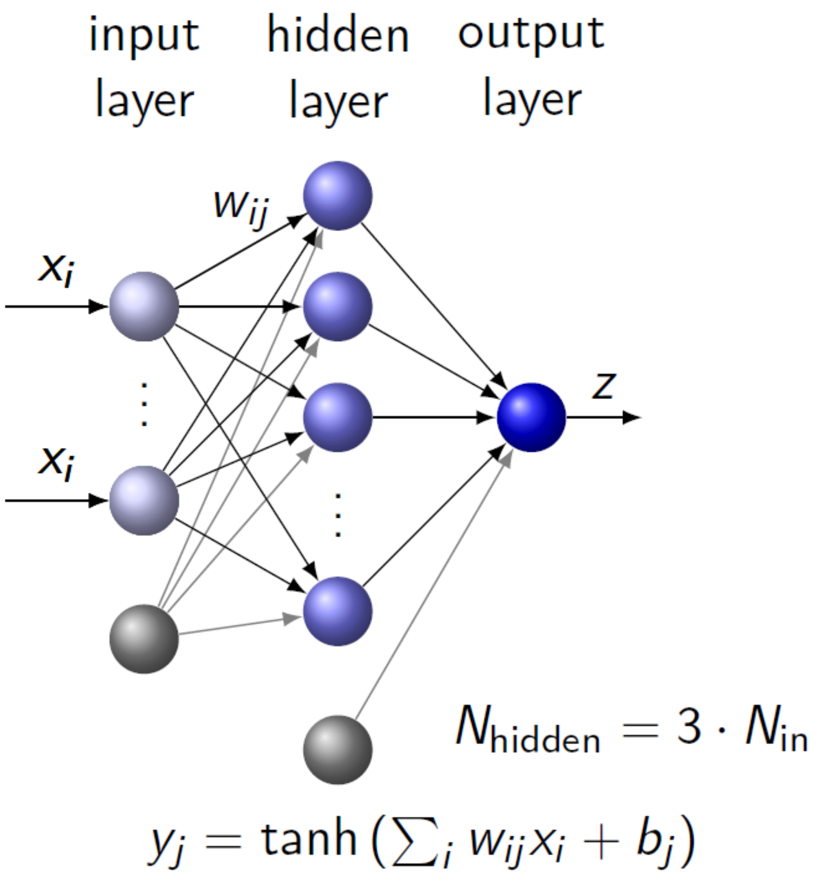


Speed-up WRT CLHEP for multiplication of rectangular (3x5) matrices

Neural network for vertexing at Belle II.

Sara Neuhaus

- Use neural network for z-vertex triggering at Belle II – larger backgrounds (luminosity upgrade $8 \cdot 10^{35}$ $cm^{-2}s^{-1}$– superKEKB in 2016):
  - topological and time information from Central Drift Chamber
  - 1 hidden layer, 1 neuron, fully forward connected MLP
  - training with back-propagation algorithm
  - output is scaled z vertex



- Firmware implementation (Virtex 7 FPGA board with fast external memory)
- Estimated resolution: 1.3 – 2.3 cm (original goal 2 cm)

Robust tracking with neural network for JLab

Cristiano Fanelli

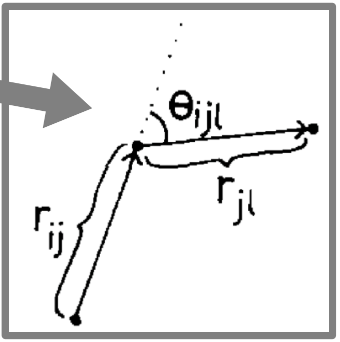
- New robust tracking for JLab experiments investigating hadron structure using polarized e^- beams at up to 12 GeV (new setup: lager pile-up, new tracking with Gas Electron Multiplier (GEM))
- Neural network implemented within a Mean Field theory framework to get the association of hits into tracks

Energy being minimized

Natural distance measure

$$E = -\frac{1}{2} \sum \delta_{jk} \frac{(\cos \theta_{ijl})^m}{r_{ij} + r_{jl}} S_{ij} S_{kl} + \ \ldots$$

neurons, connections between two points in subsequent GEM planes

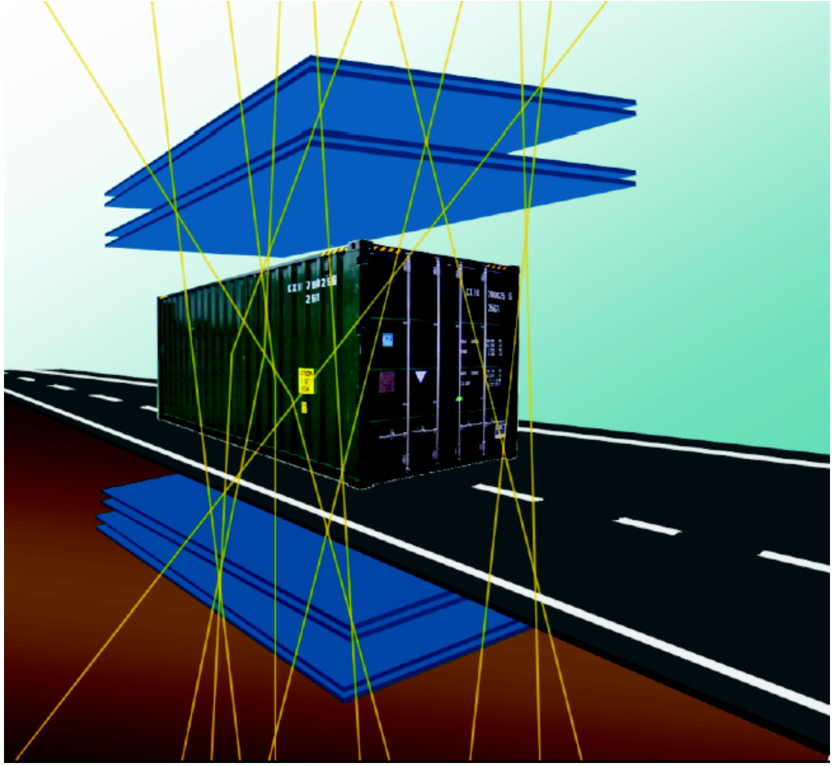


- Kalman + Rauh-Tung-Striebel fitters used to get the fits of tracks
- Improved resolution from ~80 μm (design) to 10 μm x computation time?

Muon Portal Project

M. Bandieramonte

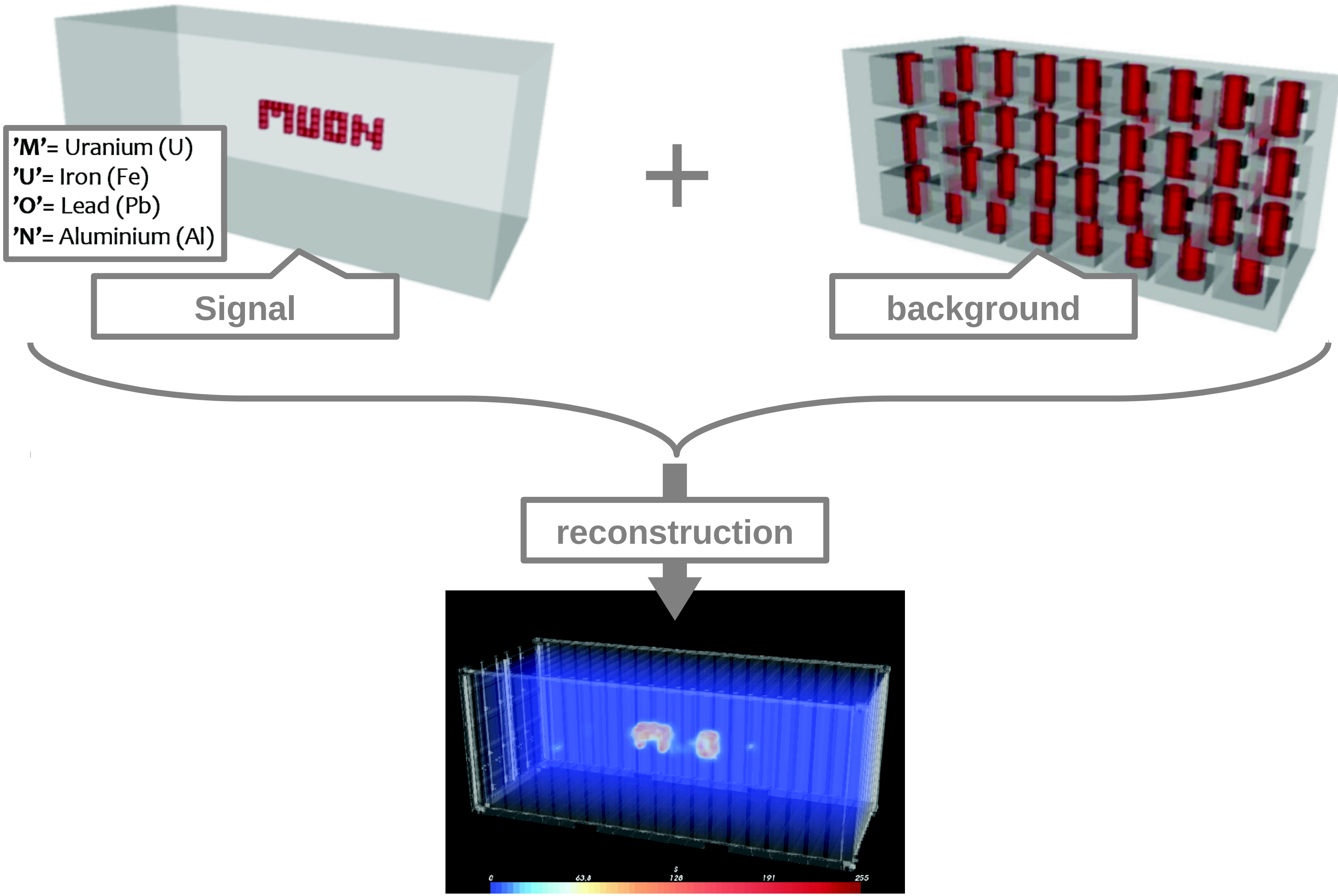
- Use non-destructive muon 3D tomography to scan the containers to detect weapons or radioactive material (200 M containers / year worldwide)
- Muon scattering strongly depends on the proton number of material
- Reconstruction:
  - basic approach: POCA (Point of closest approach between incoming and outgoing tracks) – neglects multiple scattering, poor resolution
  - advanced tool: Density based clustering with Friends-of-Friends percolation algorithm (used in cosmology) … acts as a filter, may be combined with other algortithms



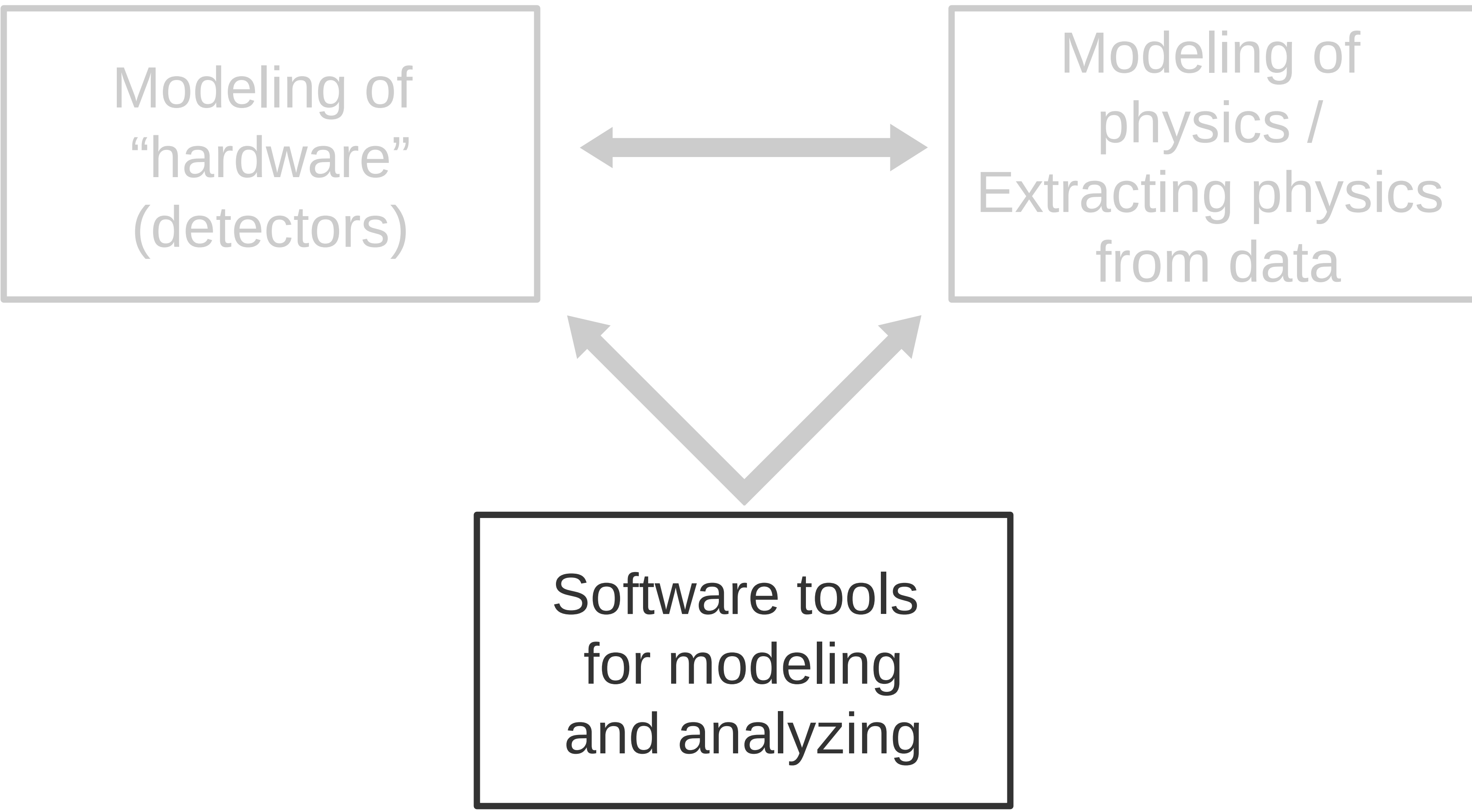
- Estimated start up ~ March 2015

Muon Portal Project

M. Bandieramonte



Disciplines



ROOT 6

Axel Naumann

- ROOT is an object-oriented program and library taylored for particle physics data analysis but used in many other fields, e.g. astronomy, data mining
- ROOT 5 – great tool but some criticism:

From Wikipedia :-)

> ## Criticisms [edit]
>
> Criticisms of ROOT include its difficulty for beginners, as well as various aspects of its design and implementation. Frequent causes of frustration include extreme code bloat, heavy use of global variables,[1] and a perverse class hierarchy.[2] From time to time these issues are discussed on the ROOT users mailing list.[3][4] While scientists dissatisfied with ROOT have in the past managed to work around its flaws,[5] some of the shortcomings are slowly being addressed by the ROOT team. The CINT interpreter, for example, has been replaced by the CLING inerpreter,[6] and numerous bugs are fixed with every release.

… many of these statements to be removed with ROOT 6

Release plan

- ROOT 6.00 published May 30, 2014 - require C++11 now
- ROOT 6.02 scheduled for end September; targeted to LHC frameworks for Run 2
- ROOT 6.04 scheduled for early 2015, plans:

ROOT 6

Axel Naumann

- ROOT is an object-oriented program and library taylored for particle physics data analysis but used in many other fields, e.g. astronomy, data mining
- ROOT 5 – great tool but some criticism:

From Wikipedia :-)

Criticisms [edit]

Criticisms of ROOT include its difficulty for beginners, as well as various aspects of its design and implementation. Frequent causes of frustration include extreme code bloat, heavy use of global variables,[1] and a perverse class hierarchy.[2] From time to time these issues are discussed on the ROOT users mailing list.[3][4] While scientists dissatisfied with ROOT have in the past managed to work around its flaws,[5] some of the shortcomings are slowly being addressed by the ROOT team. The CINT interpreter, for example, has been replaced by the CLING inerpreter,[6] and numerous bugs are fixed with every release.

… many of these statements to be removed with ROOT 6

Release plan

- ROOT 6.00 published May 30, 2014 - require C++11 now
- ROOT 6.02 scheduled for end September; targeted to LHC frameworks for Run 2
- ROOT 6.04 scheduled for early 2015, plans:

ROOT 6

Axel Naumann

- Interpreter CINT –> CLING (uses CLANG + LLVM infrastructure)
  => interpreting from #include without dictionaries (!)
  => clang diagnostics
  => C++14 support
- New TTreeReader simplifies the reading from trees
- New TFormula uses in time compiler
  e.g. TF1(“CosICan”, [](double* x, double*p) { return p[0]*cos(x[0]); }, 0., 1., 1))
- Graphics to LaTeX
- Transparency and shading
- Graphics User Interface: Improved axis zooming, guides for objects placement
  …

We need discussions and feedback - else we just do what we want! ;-)

Clad – automatic differentiation

Vassil Vassilev

- How to differentiate:
  - Numerical differentiation (precision losses, ...)
  - Symbolic differentiation – e.g. operator overloading (too slow)

=> Extend the symbolic differentiation to the compiler as a module

```
#include "clad/Differentiator/Differentiator.h”

double pow2(double x) {return x*x;}

// The body will be generated by clad:
double pow2_dx(double);

int main()
{
  // Differentiate pow2. Clad will define a fun-
  // ction named pow2_dx(double) with the  deri-
  // vative, ready to be called.
  clad::differentiate(pow2, 0);
  printf("Result is %f\n", pow2_dx(4.2));
  return 0;
}
```

The most simple example

Clad – automatic differentiation

Vassil Vassilev

- To be done – e.g. integrate Clad in ROOT6 via cling C++ interpreter => boost the performance of minimizations and fitting
- Useful examples: https://github.com/vgvassilev/clad/blob/master/demos/

```
#include "clad/Differentiator/Differentiator.h”

double pow2(double x) {return x*x;}

// The body will be generated by clad:
double pow2_dx(double);

int main()
{
  // Differentiate pow2. Clad will define a fun-
  // ction named pow2_dx(double) with the  deri-
  // vative, ready to be called.
  clad::differentiate(pow2, 0);
  printf("Result is %f\n", pow2_dx(4.2));
  return 0;
}
```

The most simple example

Tools for Higgs Discovery

Verkerke, Wouter

- Higgs discovery = result of collaborative statistical analysis of many signal and control samples

RooFit

Toolkit for modeling of expected distribution of events

Mathematical concept		RooFit class
variable	x	RooRealVar
function	$f(x)$	RooAbsReal
PDF	$f(x)$	RooAbsPdf
space point	$\vec{x}$	RooArgSet
integral	$\int_{x_{min}}^{x_{max}} f(x)\,dx$	RooRealIntegral
list of space points		RooAbsData

HistFactory

Allows structured model building for binned likelihood template models (common in LHC analyses)

RooStat

Provides a wide set of statistical tests that can be performed on RooFit models

Tools for Higgs Discovery

Verkerke, Wouter

- Higgs discovery = result of collaborative statistical analysis of many signal and control samples

Toolkit for modeling of expected distribution of events

RooFit

HistFactory

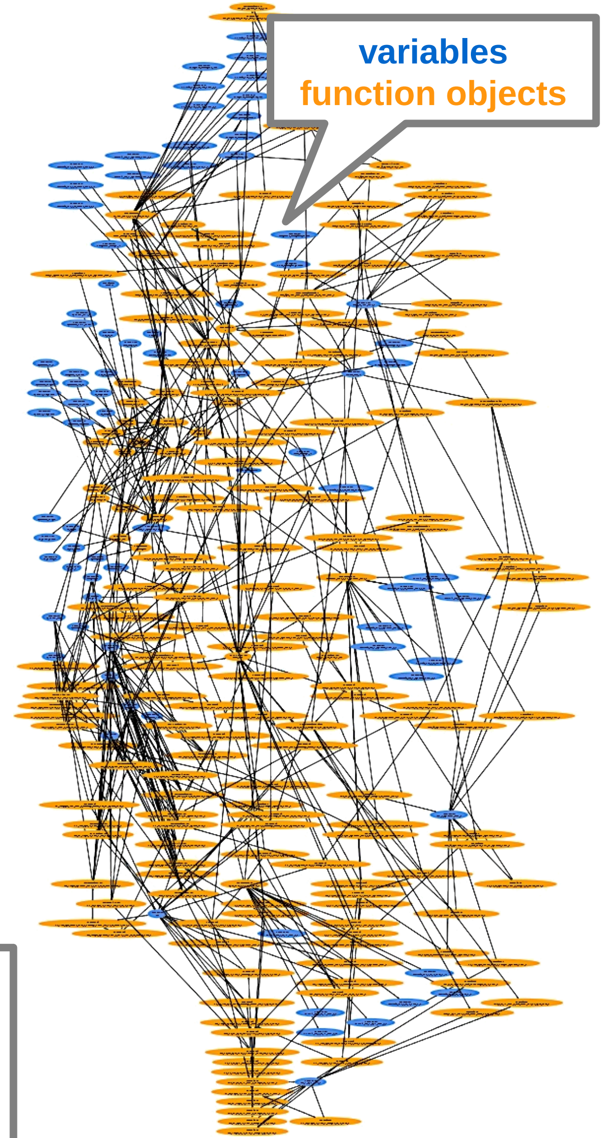
Allows structured model building for binned likelihood template models (common in LHC analyses)

Concept that persists complete likelihood models!

RooWorkspace

RooStat

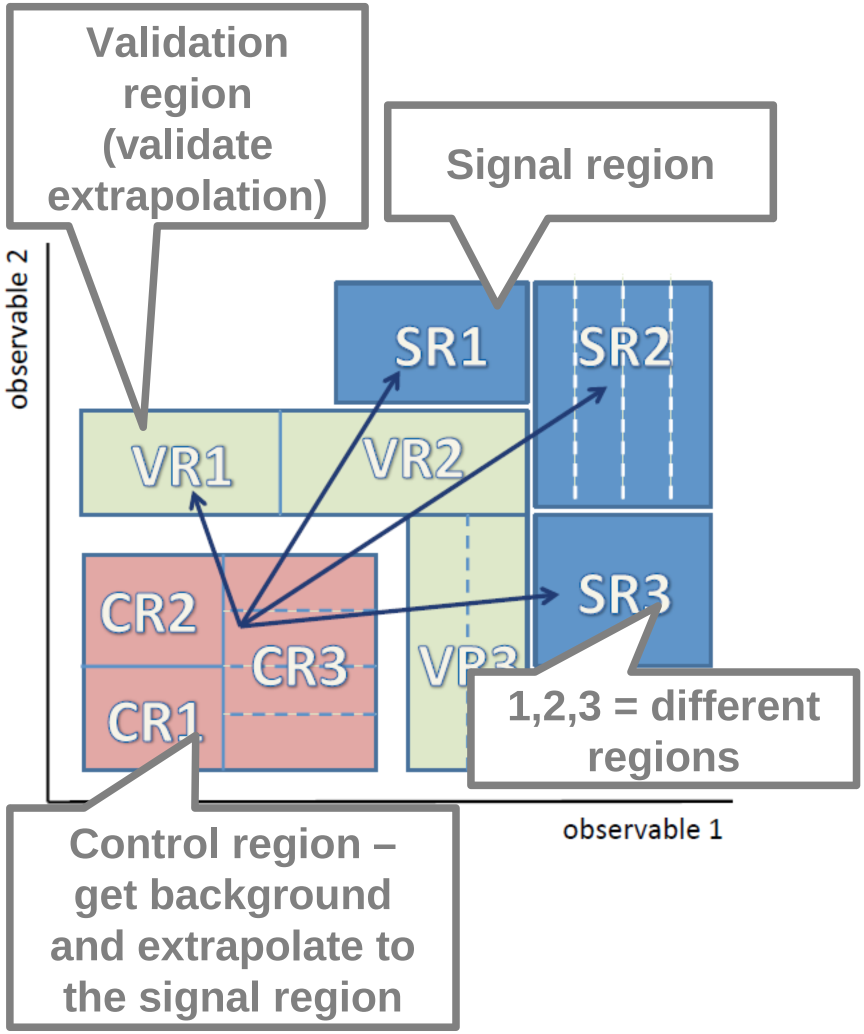
Provides a wide set of statistical tests that can be performed on RooFit models



HistFitter

Jeanette Lorenz

- Framework developed on top of RooFit + HistFactory & RooStat
- Extensions in:
  - programmable framework to build and test complex data models
  - construct and fit PDFs and provide a statistical interpretation
  - built in concepts of control, validation and signal regions with rigorous treatment of extrapolation
  - book-keeping of multiple models, tools for graphical representation
- Public release of the tool in ~ month



ATLAS data Model for Run II

Marcin Nowak

Run I:

Event reconstruction → AOD (Analysis Object Data) ~PB → DPD (Derived Physics Data) ~PB

Legible only inside complicated ATLAS software called ATHENA => unusable users analyses

Each subgroup creates DPDs = ntuples readable in ROOT => issue with size of all the data

Run II:

Event reconstruction → xAOD = AOD+DPD ~PB

Data object legible in ROOT. Can be skimmed both in ROOT or in ATHENA

ATLAS data Model for Run II

Marcin Nowak

```
#include “xAODRootAccess/Init.h”
#include “xAODRootAccess/TEvent.h”
#include “xAODMuon/MuonContainer.h”

int main()
{
  xAOD::Init();
  TFile* file = TFile::Open(“xAOD.root”, “READ”);
  xAOD::TEvent event;
  event.readFrom(file);

  for( Long64_t entry=0; entry < event.getEntries(); ++entry
  {
     event.getEntry(entry);
     const xAOD::MuonContainer* muons = 0;
     event.retrieve(muons, “Muons”);
     std::cout << “1st muon pT = “ << muons->at(0)->pt() << std::endl;
  }
  return 0;
}
```

Simple access to events

Simple access e.g. to kinematics of objects in containers

Disciplines

Modeling of “hardware” (detectors)

Modeling of physics / Extracting physics from data

Software tools for modeling and analyzing

Multivariate data analysis

Helge Voss

- HEP is not the “state of the art” in MVA
- Overcame the standard problems (systematics, what clasiffier, what variables, …) and try to get closer to and profit from the latest methods used worldwide out of HEP

Jiří Franc

- Compare different MVA techniques within inclusive top pair production cross-section measurement in D0 at Tevatron:

Econometry

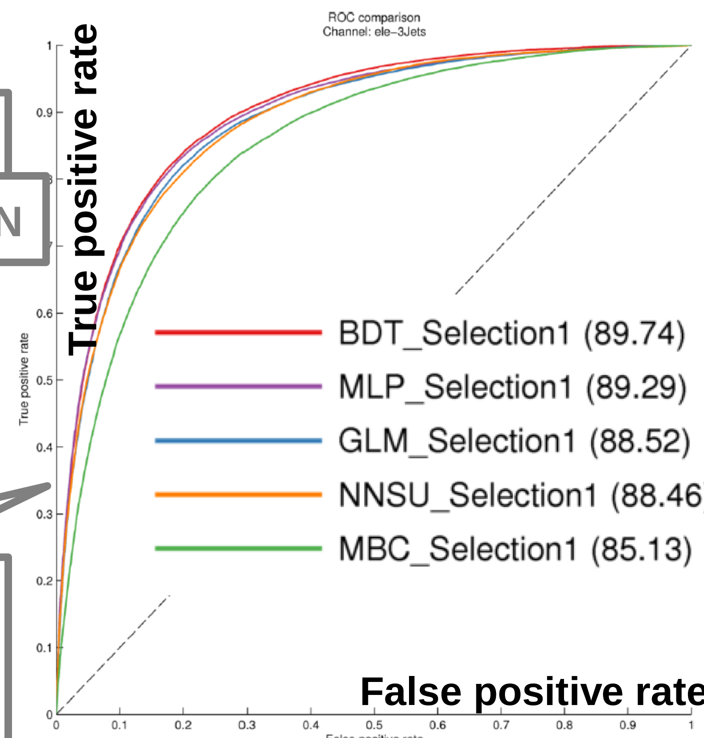
Acoustics

Novel NN

  – Generalized Linear Models (GLM)
  – Model based clustering (MBC)
  – Neural networks with switching units (NNSU)
  – Boosted Decision Tree (BDT) from TMVA
  – Multilayer Perceptrons (MLP) from TMVA

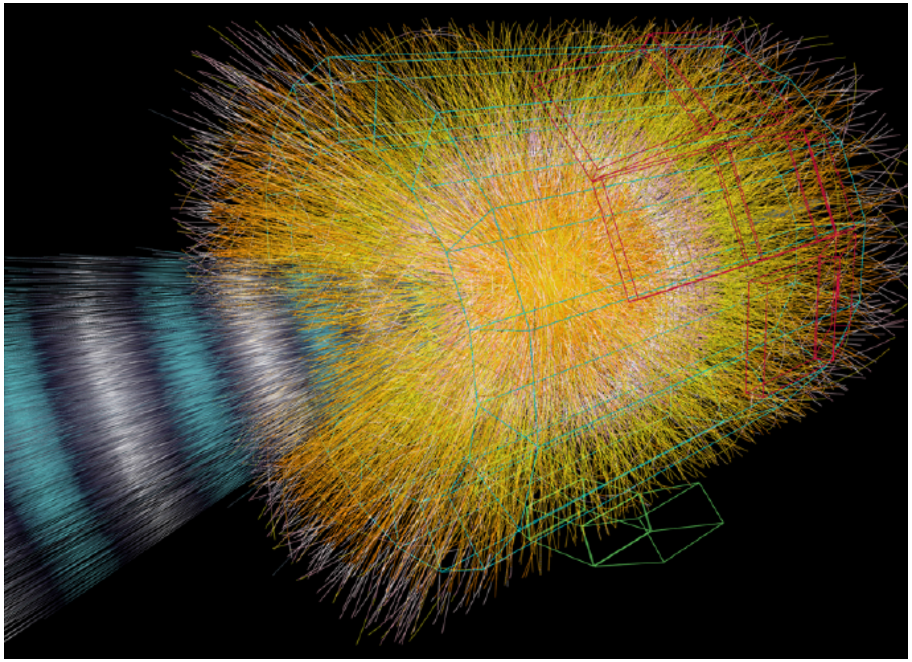
HEP “standards”

Appropriate use and good training => all show similar performance!

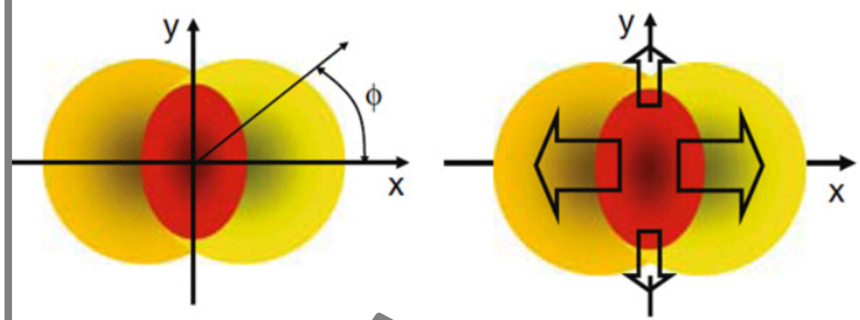


Challenging environment of heavy ion collisions

Scott Pratt



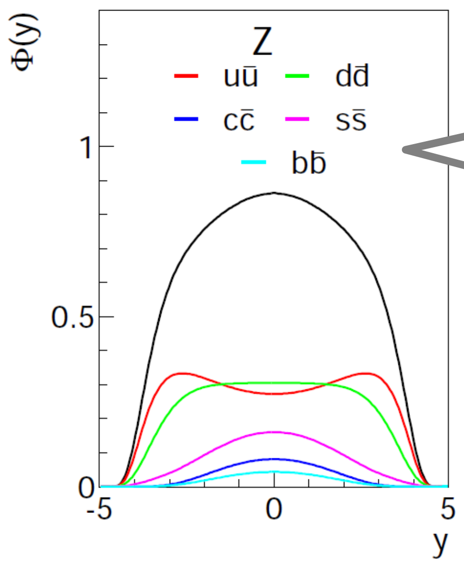
$$v_2 \equiv \langle \cos 2\phi \rangle$$



Basic observables like elliptic flow ("global event shape") are strongly model dependent

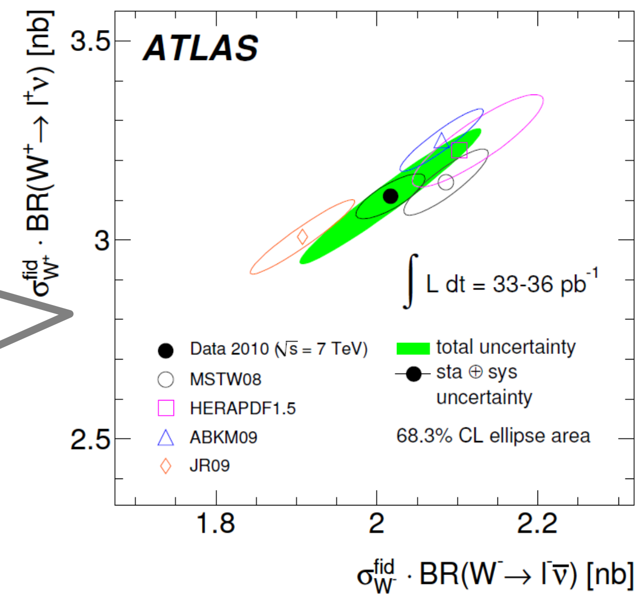
- How to extract information from the complex collision?
- Markov chain Monte Carlo method for extracting the viscosity and equation state of the Quark-Gluon plasma created in the collision.

- Parton distribution functions essential for doing the physics at LHC



Strong dependence of observable quantities on PDFs

Large uncertainties in the knowledge of PDF (e.g. gluon PDF at small-x …)

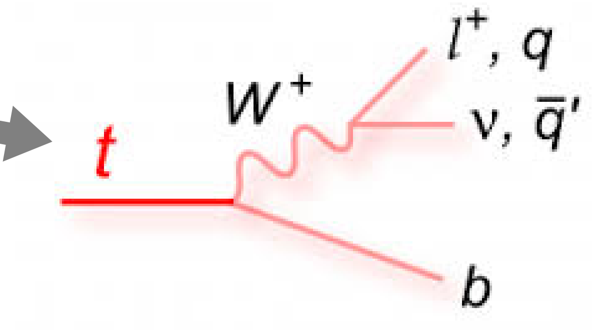


- QCD fit framework which allows namely:
  - extracting PDFs from data
  - comparing theory predictions (e.g. nNLO pQCD σ ⊗ nNLO PDFs) with data
- PDF in standard LHAPDF grids, gridding tools like APPLgrid allowing a fast use of nNLO pQCD σ)
- More than 15 papers with LHC data using the framework

Matrix element method at CMS

Camille Beluffi

- Typical problem: reconstruction of particles from chain decays.



- Methods: - cut based analysis
  - MVAs
  - Matrix Element method
- ME establishes direct link between theory and event reconstruction.
- Discriminant is built in terms of probability that the event with reconstructed kinematics x matches the hypothesis α:

pQCD matrix element

$$P(x^{vis}|\alpha) = \frac{1}{\sigma_\alpha} \int dx_1 dx_2 f(x_1) f(x_2) \int d\phi |M(p)_\alpha|^2 W(p^{vis}, p)$$

PDFs

Transfer function from simulation

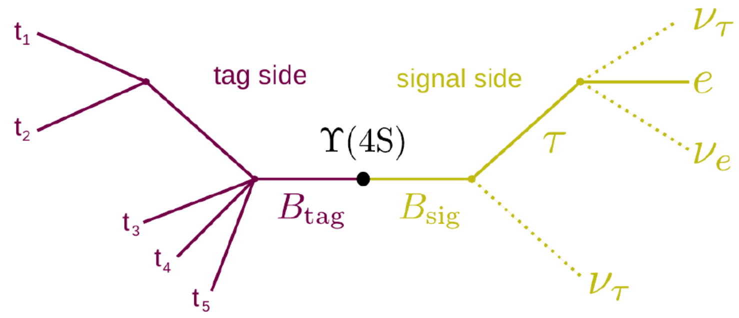
+ Maximize the amount theoretical information in discrimination
+ No complicated training as for MVAs
+ Versatile (CMS: Higgs search, spin, Tevatron: top mass, ...)

– Can be slow
– Only at Leading Order

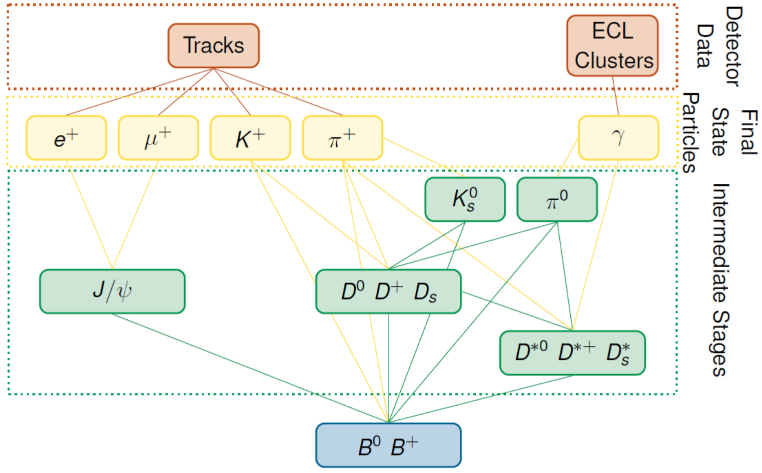
Hierarchical B meson reconstruction at Belle II

C. Pulvermacher

- $\Upsilon(4S)$ to two B mesons to study CP violation, (beyond) SM physics, ...
- Complicated decay structure with different objects in final state objects. Reconstructing one of B meson allows getting 4-momentum of the other B without explicit reconstruction of the full decay chain.



- Hierarchical reconstruction separated into stages:
  - Stage of final particles: multivariate classifiers can improve PID
  - Stage of intermediate decays: combine particles into different decay each with own classifier
  - O(100) classifiers handle O(1000) exclusive B decay modes



- Hierarchical reconstruction implemented in an automated framework which needs only minimal interventions by users

Other “advanced techniques”

Alex Mott

- Identifying the Higgs boson with Quantum Computer
  - formulate the Quadratic Unconstrained Binary Optimization problem in speech of Hamiltonian and use the Adiabatic Quantum Annealing
  - can find an optimal classifier using Quantum Annealer
  - applied on H –> $\gamma\gamma$

Evan Sangaline

- Pridix: Particle Identification without modeling the response
  - uses minimization of generalized Kullback-Leibler diverge
  - generic – can be used across experiments
  - very good performance reported!
  - https://github.com/sangaline/pidrix

Other “advanced techniques”

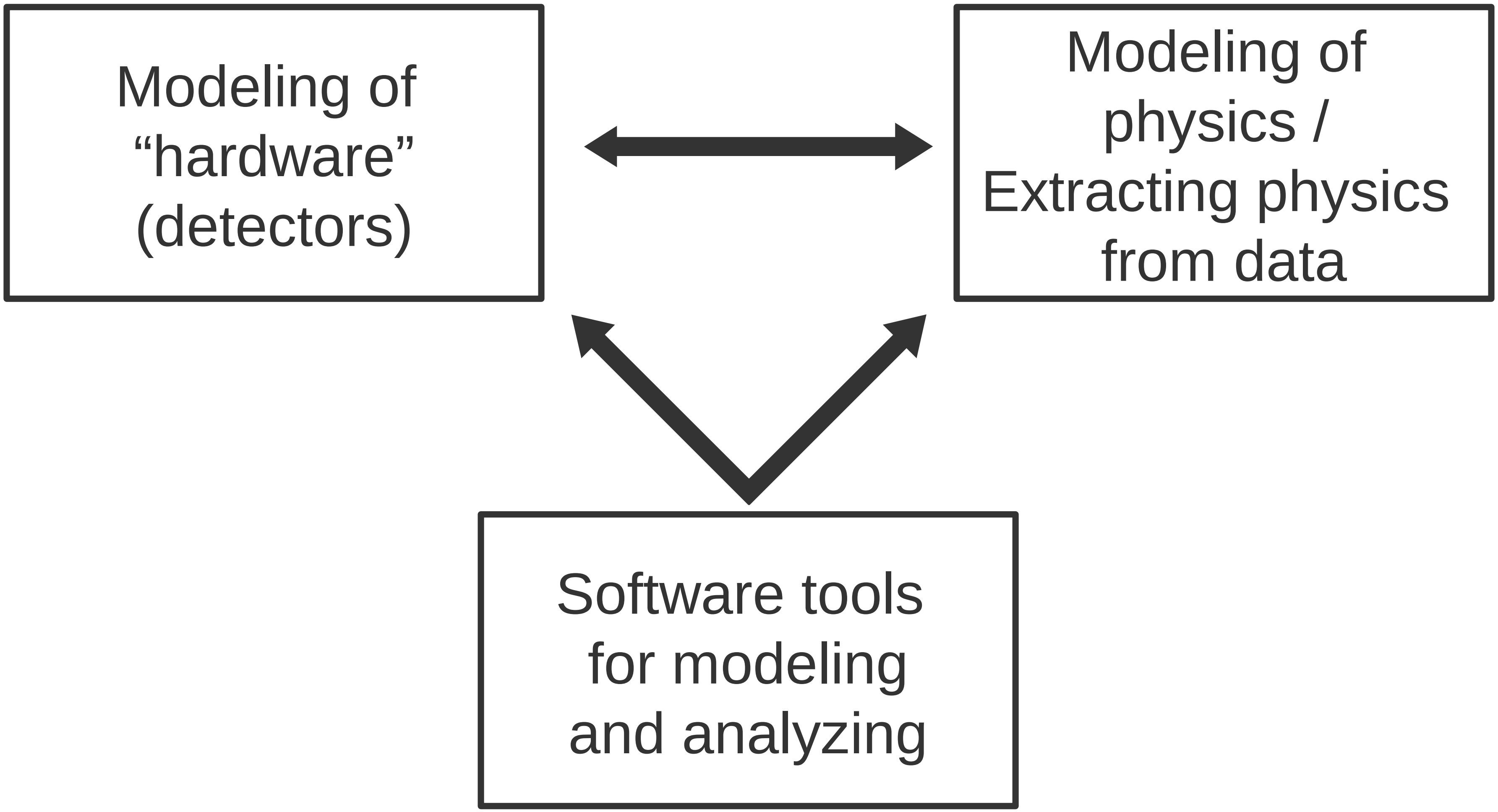
Nikolay Gagunashvili

- Density mixture unfolding
  - Mixture Density Model used for representation of unknown true distribution
  - cross-validation approach used to define optimal parameters of the unfolding
  - good performance demonstrated on non-trivial signal distribution convoluted with gaussian resolution
  - ☺ can be used as multi-dimensional unfolding
  - ☹ public implementation needed!

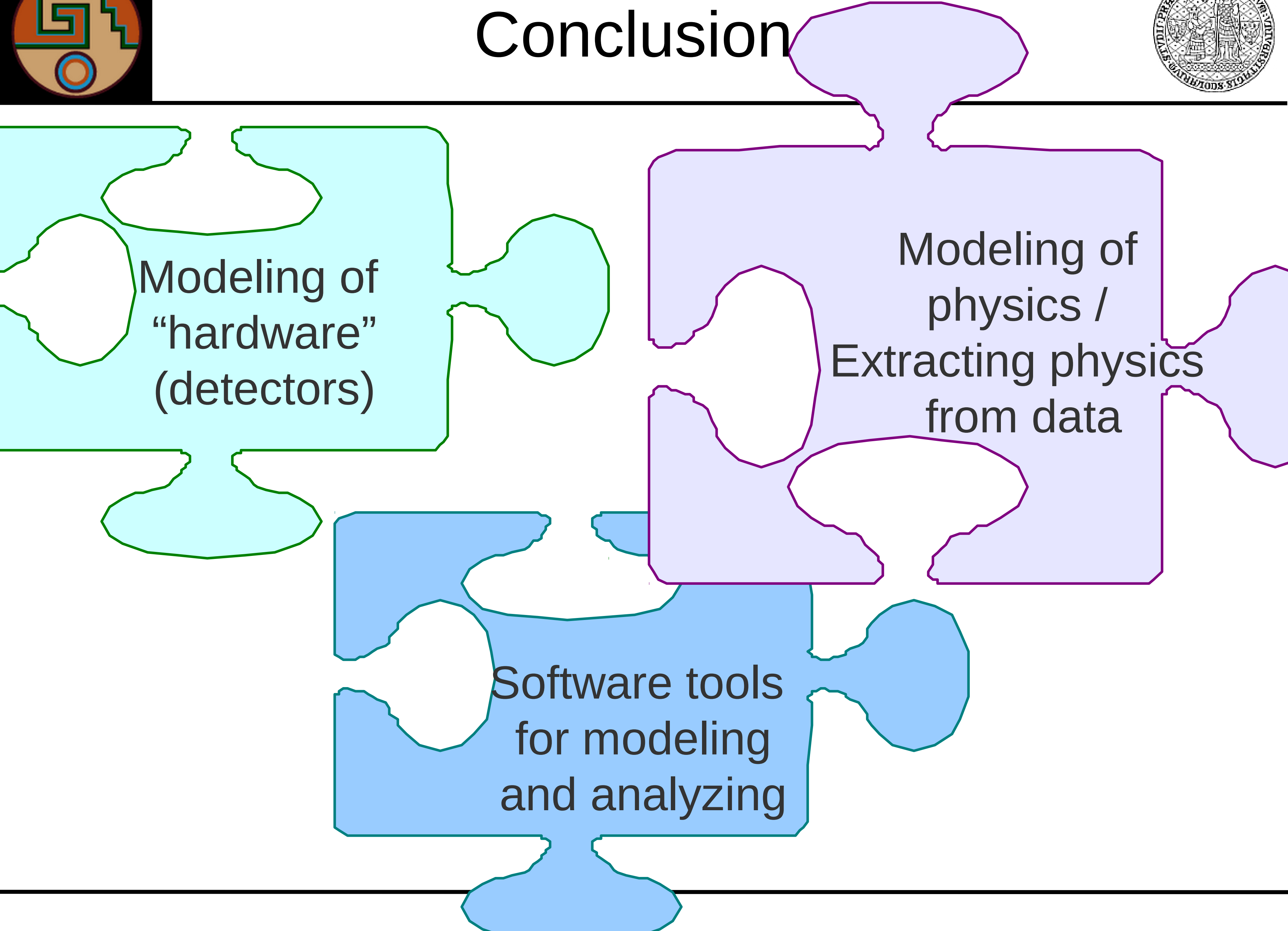
Vladislav Matousek

- Overview of deconvolution methods for γ-ray spectroscopy
  - Richardson-Lucy Algorithm
  - Gold deconvolution algorithm
  - Richardson-Lucy deconvolution
  - Maximum A Posteriori Deconvolution Algorithm
  - …
  - some implemented in ROOT (see TSpectra classes)

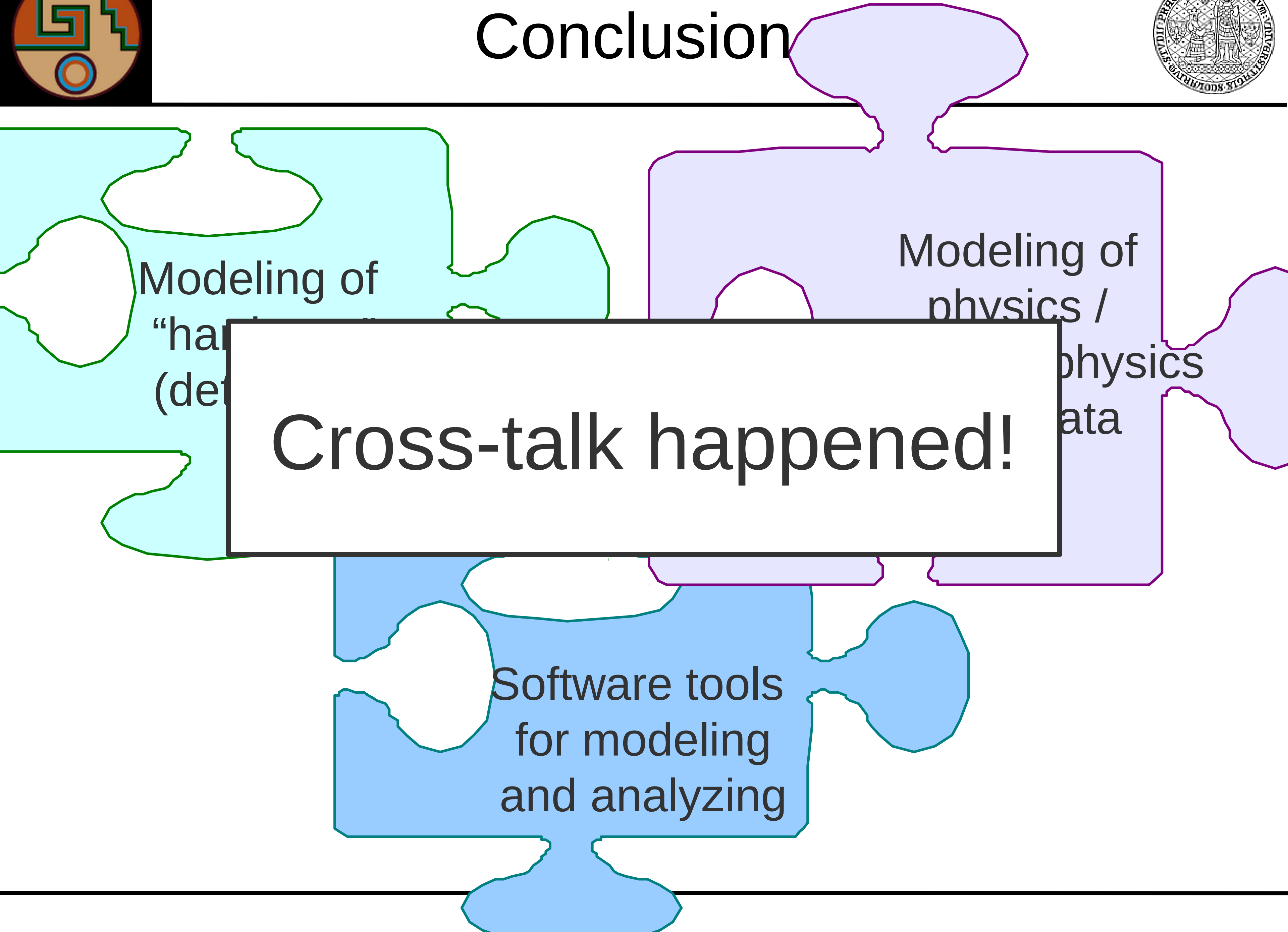
Disciplines



Conclusion


Modeling of “hardware” (detectors)
Modeling of physics / Extracting physics from data
Software tools for modeling and analyzing

Conclusion


Modeling of
Modeling of
physics /
Cross-talk happened!
Software tools
for modeling
and analyzing

Backup