CMS public data: format and usage

Tom McCauley
thomas.mccauley@cern.ch
Fermilab

DPHEP/DASPOS Workshop
CERN
21 Mar 2013

CMS public data

The CMS experiment has allowed the release of the following data to the public for use in education and outreach

- 2000 events each of **J/ψ → μμ**, J/ψ → ee
- 2000 events each of Υ → μμ, Υ → ee
- 500 events each of **Z → μμ**, **Z → ee**
- 1000 events each of **W → μν**, **W → eν**
- 100,000 events each of **dimuon**, *dielectron*, and dijet events in the energy range 2-110 GeV
- 19 Higgs candidate events: *10* γγ, **1 2e2μ, 1 4e, 1 4μ**, 2 bb, 2 ττ, 2 WW in the mass range 120-130 GeV
- **(~50 1/pb single muons for top quark analysis)**

Bold: indicates datasets already delivered and/or in use

Italic: in production/preparation

http://cms.web.cern.ch/content/cms-public-data

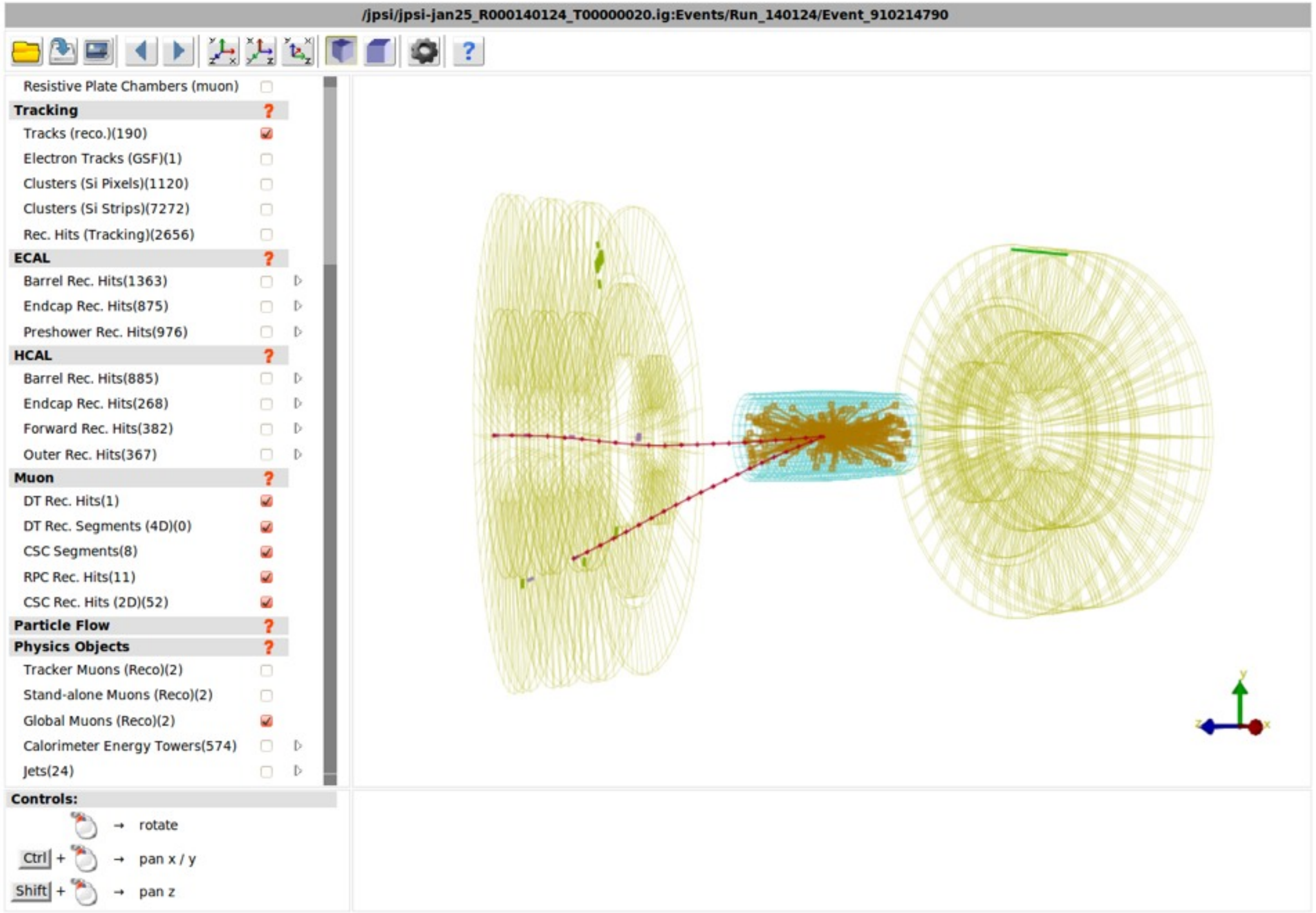
Data format

Basis format of released datasets is the **ig** format (used for iSpy event display http://cern/ch/ispy):

- Human-readable, text-based file format
- Originally developed for iSpy event display
- Simple .zip files with a flat directory structure with a separate directory per run: Events/Run_*/Event_*
- Each run directory contains one or more event files
- The event files themselves are JSON (JavaScript Object Notation) files (easily used for web applications)

Data format

- .ig files are extracted from CMS data (using "CMSSW" framework) including necessary physics (*e.g.* tracks, reconstructed particles, hits) and graphical information (*i.e.* positions, *etc.* in global coordinates)
- csv and JSON-formatted summary files (*i.e.* containing just 4-vector-type information) easily generated
- Files can be easily reorganized to suit the needs of new exercises
- Format is flexible, extensible, and self-documenting
- C++ and python APIs available
- Easily (trivially) handled in ruby (JavaScript)



/jpsi/jpsi-jan25_R000140124_T00000020.ig:Events/Run_140124/Event_910214790

- Resistive Plate Chambers (muon) □

Tracking ?

- Tracks (reco.)(190) ☑
- Electron Tracks (GSF)(1) □
- Clusters (Si Pixels)(1120) □
- Clusters (Si Strips)(7272) □
- Rec. Hits (Tracking)(2656) □

ECAL ?

- Barrel Rec. Hits(1363) □ ▷
- Endcap Rec. Hits(875) □ ▷
- Preshower Rec. Hits(976) □ ▷

HCAL ?

- Barrel Rec. Hits(885) □ ▷
- Endcap Rec. Hits(268) □ ▷
- Forward Rec. Hits(382) □ ▷
- Outer Rec. Hits(367) □ ▷

Muon ?

- DT Rec. Hits(1) ☑
- DT Rec. Segments (4D)(0) ☑
- CSC Segments(8) ☑
- RPC Rec. Hits(11) ☑
- CSC Rec. Hits (2D)(52) ☑

Particle Flow ?

Physics Objects ?

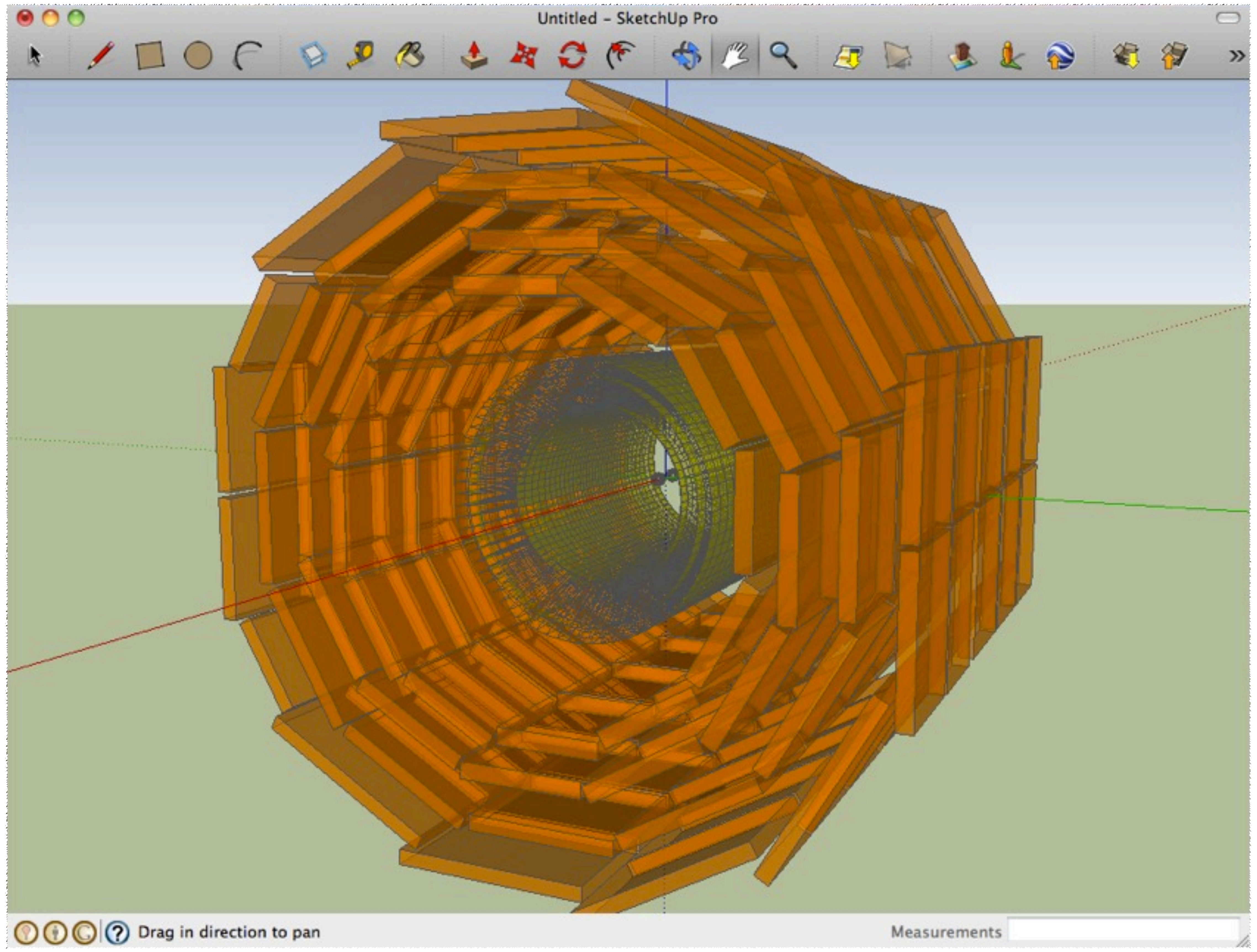
- Tracker Muons (Reco)(2) □
- Stand-alone Muons (Reco)(2) □
- Global Muons (Reco)(2) ☑
- Calorimeter Energy Towers(574) □ ▷
- Jets(24) □ ▷

Controls:

→ rotate

Ctrl + → pan x / y

Shift + → pan z



ig file geometry read with ruby plugin in SketchUp

Browser-based event display written in JavaScript using jQuery and pre3d which reads ig files: used in masterclasses and elab

Also: prototype C# API for use in Unity

http://www.i2u2.org/elab/cms/event-display

.ig event file format specifics (I)

- The event file format defines three categories of objects: *Types, Collections, and Associations*

- *Types* describe what is contained in the file (*i.e.* the “schema”):

```
“Types”:
 {
   "Event_V2": [["run", "int"],["event", "int"],["ls", "int"],["orbit",
"int"],["bx", "int"],["time", "string"],["localtime", "string"]],
   "Tracks_V2": [["pos", "v3d"],["dir", "v3d"],["pt", "double"],["phi",
"double"],["eta", "double"],["charge", "int"],["chi2", "double"],
["ndof","double"]],
   "TrackDets_V1": [["detid", "int"],["front_1", "v3d"],["front_2", "v3d"],
["front_3", "v3d"],["front_4", "v3d"],["back_1", "v3d"],["back_2", "v3d"],
["back_3", "v3d"],["back_4", "v3d"]]
 }
```

.ig event file format specifics (II)

- *Collections* describe the specific instances of each object described in *Types*:

```
“Collections”:
 {
  "Event_V2": [[146944, 200013589, 254, 66398865, 2259, "2010-Sep-30 00:58:26.130584 GMT",
"Wed Sep 29 19:58:26 2010 CDT"]
],
  "Tracks_V2": [[[0.000948177, 0.000181534, 0.0468811], [-0.642776, -0.00388211, 0.766045],
1.52506, -3.13555, 1.01068, -1, 7.021, 16],[[0.000930731, 0.00027696, 0.0467901],
[-0.760437, -0.123472, 0.637565], 1.01404, -2.98063, 0.754061, 1, 13.6519, 12]],
  "TrackDets_V1": [[302058008, [-0.0435299, -0.00451072, 0.0635923], [-0.0435936,
-0.00454519, 0.128392], [-0.040811, 0.011414, 0.128403], [-0.0407473, 0.0114485, 0.0636035],
[-0.0438106, -0.00446177, 0.0635921], [-0.0438743, -0.00449624, 0.128392], [-0.0410917,
0.011463, 0.128403], [-0.041028, 0.0114975, 0.0636033]],
[302125084, [-0.0759849, 0.0120135, 0.130731], [-0.075936, 0.0121173, 0.19553], [-0.0776272,
-0.00399411, 0.195558], [-0.0776762, -0.00409795, 0.130758], [-0.0757015, 0.0119837,
0.13073], [-0.0756525, 0.0120876, 0.19553], [-0.0773438, -0.00402386, 0.195557],
[-0.0773927, -0.0041277, 0.130758]]
 }
```

.ig event file format specifics (III)

- *Associations* encode relationships between instances of event data (e.g. a specific Tracks_V2 instance given by two indices is associated to these TrackDets_V1 given by their two indices):

```
“Associations”:
{
 “TrackDetMatches_V1”: [[[1,0], [2,0]], [[1,1], [2,1]]]
}
```

For more information see
http://cern.ch/ispy/ig-specs.htm

CMSSW to ig (with C++ API)

```
IgCollection &tracks = storage->getCollection("Tracks_V2");
IgProperty DIR   = tracks.addProperty("dir", IgV3d());
IgProperty PT  = tracks.addProperty("pt", 0.0);
IgProperty PHI = tracks.addProperty("phi", 0.0);
IgProperty ETA = tracks.addProperty("eta", 0.0);

for ( reco::TrackCollection::const_iterator track = collection->begin(), trackEnd = collection->end();
      track != trackEnd; ++track )
{
  IgCollectionItem item = tracks.create();

  item[DIR] = IgV3d((*track).px(),(*track).py(),(*track).pz());
  item[PT]  = (*track).pt();
  item[PHI] = (*track).phi();
  item[ETA] = (*track).eta();
}
```

Main users

Several organizations make use of the data and create and run educational programs for high school students such as e-Labs and masterclasses (and thus far drive the requirements):

- I2U2 (Interactions in Understanding the Universe): an "educational virtual organization" http://www.i2u2.org
- QuarkNet: provides programs and training for teachers and students in particle physics in the USA http://quarknet.fnal.gov
- IPPOG (International Particle Physics Outreach Group): network of particle physicists and educators http://ippog.web.cern.ch/ippog

All three organizations collaborate and share material

Dataset requirements

- Main users to be students, supervised by teachers, studying the data in the context of masterclasses and e-Labs
- The physics content of the datasets should be readily useable in exercises but not restrictive in scope, to allow for flexibility and surprises
- The data format must be easy to use and not require complicated software in order to read and analyze
- The exercises based on the datasets shouldn’t be too difficult but shouldn’t be trivial either

Masterclasses

- Masterclasses: on certain days in the year thousands of students from all over the world travel to nearby universities and research laboratories to listen to lectures, analyze real LHC data, and interact with other groups via videoconference.

- In 2012 there were 35 CMS masterclasses (18 in US) with thousands of students (~ 9000 over ALICE, ATLAS, and CMS in 32 countries).

- This years' classes running now, from Feb 25 - Mar 22 (one CMS class today @CERN)

- Current exercise: W+:W-, Z & J/psi invariant mass

http://www.physicsmasterclasses.org/index.php?cat=schedule

http://www.quarknet.us/library/index.php/Schedules_2013

Science Hack Days

- “A Hack Day is a 48-hour-all-night event that brings together designers, developers, scientists and other geeks in the same physical space for a brief but intense period of collaboration, hacking, and building 'cool stuff' ”*

- Two recent events in 2012, in San Francisco and Nairobi, used CMS data for data hacks (see SHD site for more info)

- IMHO: this is a nice example of unexpected benefits and applications that come from open and easy-to-use formats as **these events are self-organized**

http://cms.web.cern.ch/news/cms-public-data-activity-scoops-prize-nairobi

*http://www.sciencehackday.com

Thank you

More information and links:

http://cms.web.cern.ch/content/cms-public-data

Thanks to CMS collaboration and special thanks to Giulio Eulisse