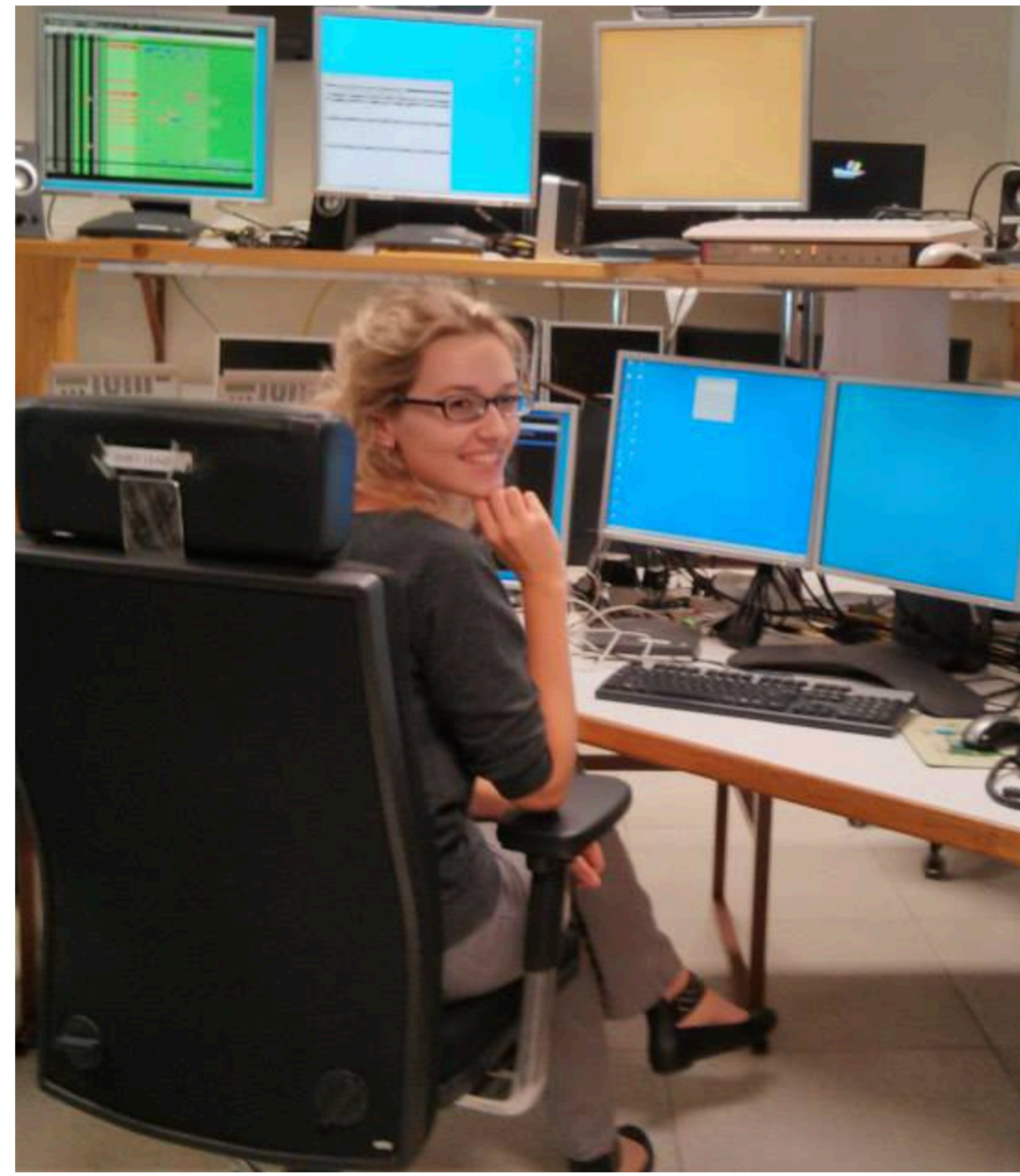
The Last Presentation

as a doctoral student at LBC
by

Ana Trisovic

About me

- Graduated at Faculty of Computing in Belgrade, Serbia and Faculty of Engineering
- Internship with Microsoft Development Center in 2013
- Technical student with LHCb in 2013-2014 working on event display for the LHCb Masterclass exercise
- PhD Candidate at University of Cambridge and doctoral student at LHCb from September 2014



Ana Trisovic 1.0

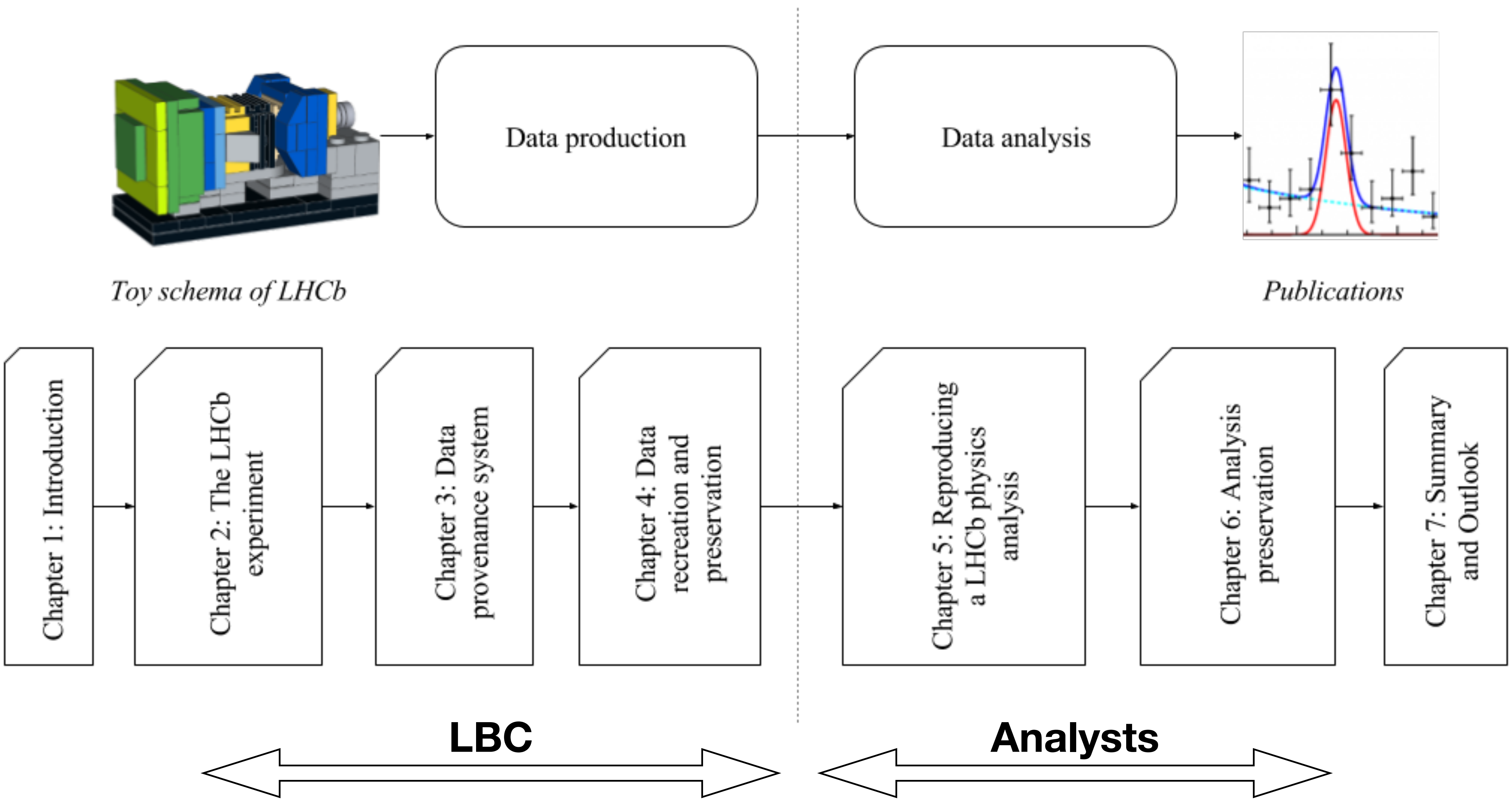
Agenda

- My work
- My CERN experience
- My feedback to the group

PhD project synopsis

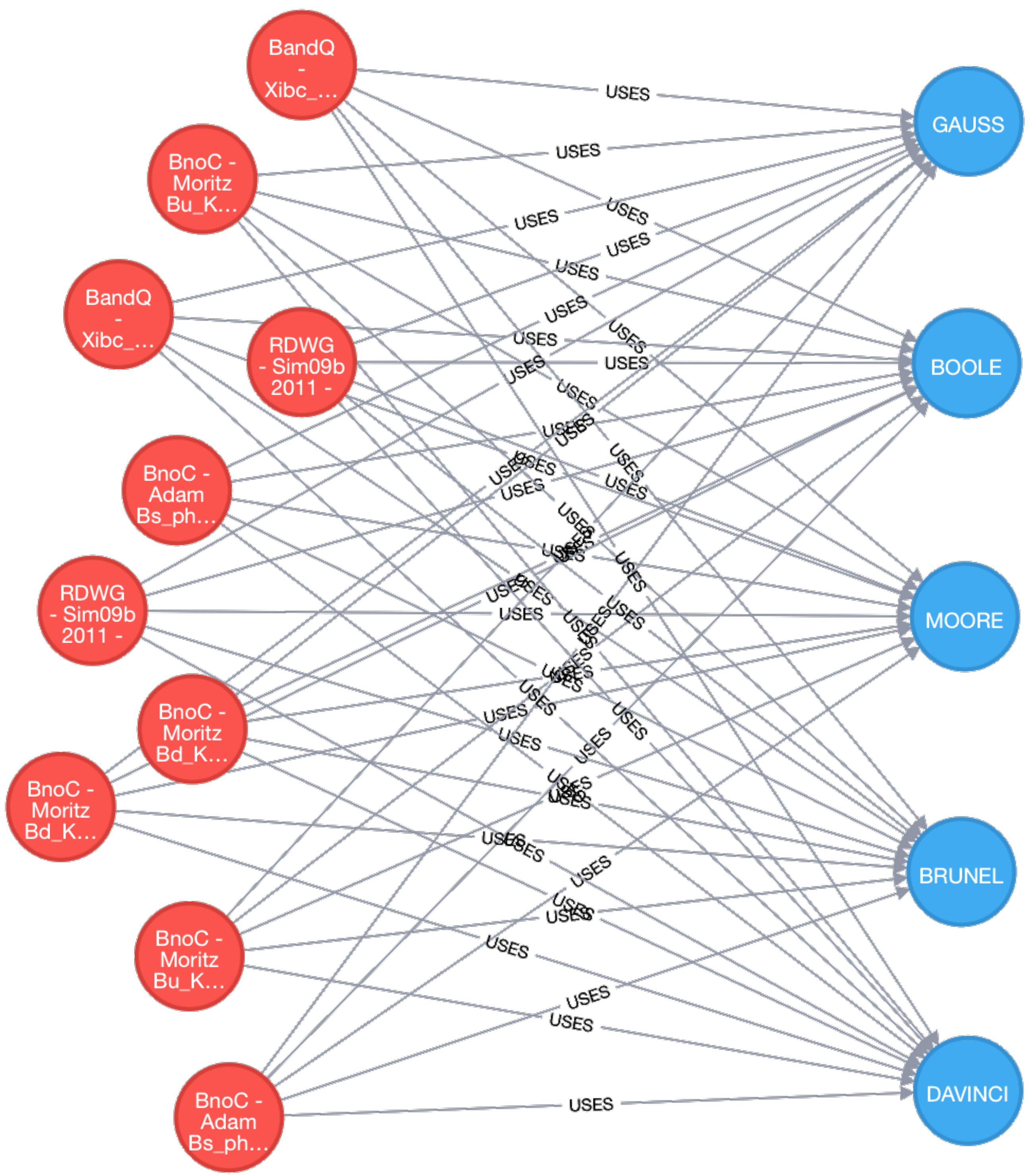
- Scientific preservation
  - Experimental data & Monte Carlo
  - Software & documentation
  - Conditions database
  - Analyses & publications
- Research reproducibility
  - Research validation
  - New measurements on old data
  - Enforced by scientific journals and funding agencies
- Good practice in analysis methodology
- Open Data

(some) Solutions
Thesis outline

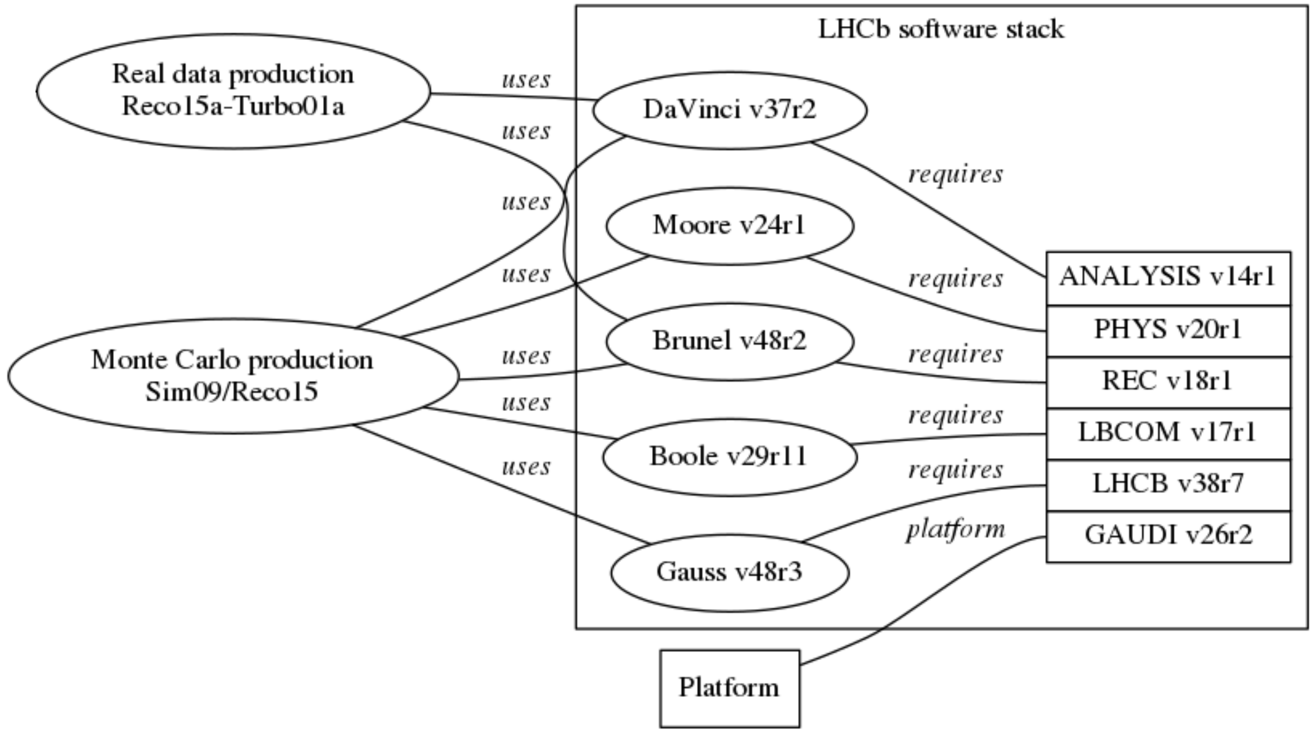


Chapter 3

Preservation database



- Keeps dependencies in software and links to data and simulation productions
- Implemented in neo4j 3



REST API

- REST API implemented on a web server with neo4j db
- Get production request by ID: http://lbc-preserve.cern.ch/getProduction?id=3137
- Get production request by BKK path: http://lbc-preserve.cern.ch/getProductionBKK?bkk=MC/2011/Beam3500GeV-2011-MagDown-Nu2-EmNoCuts/Sim05a/Trig0x40760037Flagged/Reco12a/Stripping17NoPrescalingFiltered/11114001

```
atrisovi@lxplus072: ~ > curl http://lbc-preserve.cern.ch/getProduction?id=3137
{"NumberOfEvents": "2500000", "upTime": "2011-05-03T08:28:20", "bkkpath": "MC/2010/Beam3500GeV-Oct2010-MagUp-Nu2,5/Sim
01/Trig0x002e002aFlagged/Reco08/11874021", "RequestType": "Simulation", "EventType": "11874021", "year": "2010", "p4St
ep": "{u'ApplicationName': u'Brunel', u'Usable': u'Obsolete', u'StepId': 11319, u'ApplicationVersion': u'v37r8p5', u'E
xtraPackages': u'AppConfig.v3r89;SQLDDDB.v5r44', u'StepName': u'Reco08', u'ProcessingPass': u'Reco08', u'SystemConfig'
: None, u'mcTCK': None, u'DDDB': u'head-20101206', u'Visible': u'Y', u'OptionsFormat': None, u'OptionFiles': u'$APPCON
FIGOPTS/Brunel/DataType-2010.py;$APPCONFIGOPTS/Brunel/MC-WithTruth.py;$APPCONFIGOPTS/Brunel/earlyData.py', u'CONDDB':
u'sim-20101210-vc-mu100', u'isMulticore': u'N', u'DQTag': None}", "p1Step": "{u'ApplicationName': u'Gauss', u'Usable':
 u'Obsolete', u'StepId': 12218, u'ApplicationVersion': u'v39r2', u'ExtraPackages': u'AppConfig.v3r89;DecFiles.v23r5;SQ
LDDDB.v5r44', u'StepName': u'Sim01', u'ProcessingPass': u'Sim01', u'SystemConfig': None, u'mcTCK': None, u'DDDB': u'he
ad-20101206', u'Visible': u'Y', u'OptionsFormat': None, u'OptionFiles': u'$APPCONFIGOPTS/Gauss/Beam3500GeV-mu100-MC10-
nu2,5.py;$DECFILESROOT/options/@{eventType}.opts;$LBPYTHIAROOT/options/Pythia.py', u'CONDDB': u'sim-20101210-vc-mu100'
, u'isMulticore': u'N', u'DQTag': None}", "ProPath": "Sim01/Trig0x002e002aFlagged/Reco08", "RequestID": "3137", "p3Ste
p": "{u'ApplicationName': u'Moore', u'Usable': u'Yes', u'StepId': 11339, u'ApplicationVersion': u'v10r2', u'ExtraPacka
ges': u'AppConfig.v3r89;SQLDDDB.v5r44', u'StepName': u'Trig0x002e002aFlagged', u'ProcessingPass': u'Trig0x002e002aFlag
ged', u'SystemConfig': None, u'mcTCK': None, u'DDDB': u'head-20101206', u'Visible': u'Y', u'OptionsFormat': None, u'Op
tionFiles': u'$APPCONFIGOPTS/Moore/MooreSimProduction.py;$APPCONFIGOPTS/Conditions/TCK-0x002e002a.py;$APPCONFIGOPTS/Mo
ore/DataType-2010.py', u'CONDDB': u'sim-20101210-vc-mu100', u'isMulticore': u'N', u'DQTag': None}", "p2Step": "{u'Appl
icationName': u'Boole', u'Usable': u'Obsolete', u'StepId': 11338, u'ApplicationVersion': u'v21r9', u'ExtraPackages': u
'AppConfig.v3r89;SQLDDDB.v5r44', u'StepName': u'DIGI10', u'ProcessingPass': u'DIGI10', u'SystemConfig': None, u'mcTCK'
: None, u'DDDB': u'head-20101206', u'Visible': u'N', u'OptionsFormat': None, u'OptionFiles': u'$APPCONFIGOPTS/Boole/De
fault.py;$APPCONFIGOPTS/Boole/DataType-2010.py;$APPCONFIGOPTS/Boole/IgnoreFlatSpillover.py;$APPCONFIGOPTS/L0/L0TCK-0x0
02A.py', u'CONDDB': u'sim-20101210-vc-mu100', u'isMulticore': u'N', u'DQTag': None}", "SimCondition": "Beam3500GeV-Oct
2010-MagUp-Nu2,5", "crTime": "2011-03-30T13:56:11"}atrisovi@lxplus072: ~ >
atrisovi@lxplus072: ~ >
```


Backup

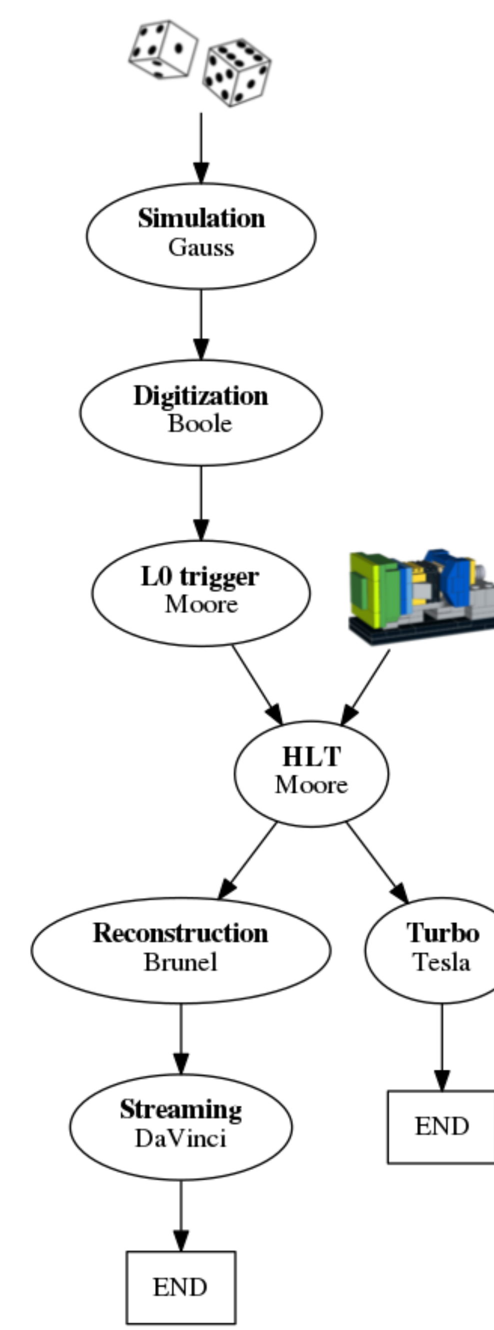
Chapter 4

Production on the Cloud

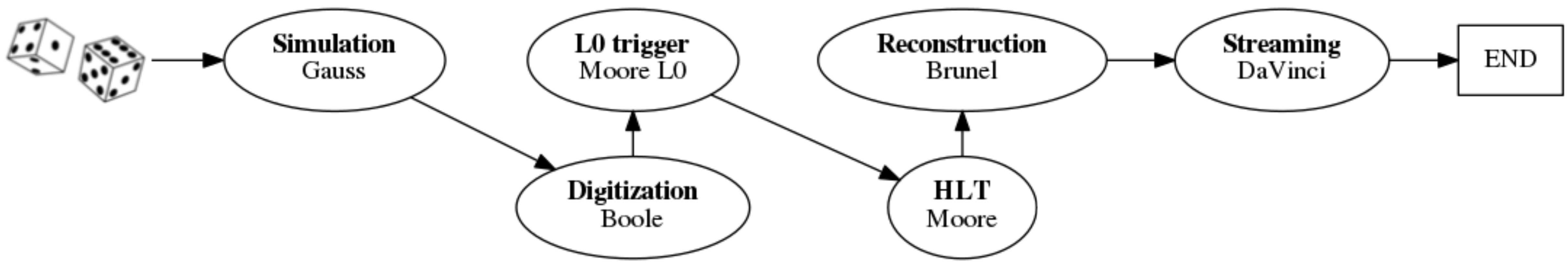
- Running data production outside of the LHCb infrastructure, on the CERN Cloud

- Procedure:

    1. Node is deployed with Docker
    2. Environment setup from CVMFS
    3. Production steps from preservation db
    4. Data & key tab iff Reco-Stripping
    5. Run production workflow



Workflow




```bash
#!/bin/bash

source  /cvmfs/lhcb.cern.ch/lib/LbLogin.sh --no-userarea
LbLogin -c x86_64-slc5-gcc46-opt

# simulation
lb-run --use ProdConf --use  AppConfig.v3r171 --use DecFiles.v27r6 Gauss/v45r3 gaudirun.py prodConf_Gauss1.py
# digitization
lb-run --use AppConfig.v3r171 --use ProdConf Boole/v26r3 gaudirun.py prodConf_Boole2.py
# trigger
LbLogin -c x86_64-slc5-gcc43-opt
lb-run --use AppConfig.v3r171 --use ProdConf Moore/v12r8g3 gaudirun.py prodConf_Moore3.py
# reconstruction
lb-run --use AppConfig.v3r171 --use ProdConf Brunel/v43r2p7 gaudirun.py prodConf_Brunel4.py
# stripping
lb-run --use AppConfig.v3r171 --use ProdConf DaVinci/v32r2p3 gaudirun.py prodConf_DaVinci5.py
```

Event comparison

Original events

p & pt

lhcbIDs

```
Run number Event number 2492005 102001
3900.72 219.011908904 47269003791
269007412
269012060
269015676
269020319
269023932
269028685
269032374
269037014
269040627
269658336
269662691
269666509
269670866
269674691
269680190
269683785
269688415
269691994
269696612
539298917
539561090
541395436
```

Reproduced events

```
Run number Event number 2492005 102001
3900.72 219.011908904 47269003791
269007412
269012060
269015676
269020319
269023932
269028685
269032374
269037014
269040627
269658336
269662691
269666509
269670866
269674691
269680190
269683785
269688415
269691994
269696612
539298917
539561090
541395436
```

Event comparison

Original events

p & pt

```
Run number Event number 2492005 102001
3900.72 219.011908904 47269003791
269007412
269012060
269015676
269020319
269023932
269028685
269032374
269037014
269040627
269658336
269662691
269666509
269670866
269674691
269680190
269683785
269688415
269691994
269696612
539298917
539561090
541395436
```

lhcbIDs

Reproduced events

```
Run number Event number 2492005 102001
3900.72 219.011908904 47269003791
269007412
269012060
269015676
269020319
269023932
269028685
269032374
269037014
269040627
269658336
269662691
66509
0866
691
0190
3785
3415
1994
5612
8917
539561090
541395436
```

Event comparison

Real - downloaded from the grid

```
atrisovi@lxplus004: /eos/lhcb/user/a/atrisovi/lhcb-compare-events/real > ll
total 187360163
-rw-r--r--. 1 atrisovi z5 3710499059 Aug 22 14:26 00024919_00000001_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3641542078 Aug 22 13:28 00024919_00000002_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3605349187 Aug 22 14:33 00024919_00000003_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3620057849 Aug 22 14:25 00024919_00000004_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3655382191 Aug 22 13:17 00024919_00000005_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3644897803 Aug 22 14:13 00024919_00000006_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3623223567 Aug 22 13:14 00024919_00000007_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3649216508 Aug 22 14:46 00024919_00000008_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3639759223 Aug 22 14:40 00024919_00000009_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3654475723 Aug 22 14:37 00024919_00000010_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3634906944 Aug 22 13:34 00024919_00000011_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3631858157 Aug 22 15:02 00024919_00000012_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3673533811 Aug 22 14:43 00024919_00000013_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3697831255 Aug 22 13:18 00024919_00000014_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3624013649 Aug 22 13:37 00024919_00000015_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3615846045 Aug 22 14:46 00024919_00000016_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3631818514 Aug 22 14:55 00024919_00000017_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3693872562 Aug 22 13:25 00024919_00000018_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3675550185 Aug 22 13:12 00024919_00000019_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3623378208 Aug 22 14:58 00024919_00000020_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3693936288 Aug 22 14:12 00024919_00000021_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3678138717 Aug 22 14:59 00024919_00000022_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 3787272537 Aug 22 14:21 00024919_00000023_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 3679624001 Aug 22 15:07 00024919_00000024_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 1203472485 Aug 22 14:31 00024919_00000026_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5   90738768 Aug 22 14:46 00024919_00000027_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5  185108844 Aug 22 14:34 00024919_00000028_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 1594924799 Aug 22 15:22 00024919_00000029_1.allstreams.dst
-rwxr-xr-x. 1 atrisovi z5 2575752716 Aug 22 14:51 00024919_00000030_1.allstreams.dst
-rw-r--r--. 1 atrisovi z5 2492407317 Aug 22 14:23 00024919_00000031_1.allstreams.dst
-rw-rw-r--. 1 atrisovi z5      14578 Aug 22 15:34 event_list.txt
-rw-rw-r--. 1 atrisovi z5        899 Aug 22 15:32 listEvents.py
-rw-rw-r--. 1 atrisovi z5       3168 Aug 22 13:08 MC2011Beam3500GeV-2011-MagDown-Nu2-Pythia8Sim08aDigi13Trig0x40760037Reco14aStripping20r1NoPrescalingFlagged2
1113001STREAMSDST.py
atrisovi@lxplus004: /eos/lhcb/user/a/atrisovi/lhcb-compare-events/real >
```

Event comparison

Reproduced - created on the Cloud

These events were recreated in my open stack machine ‘antris’ on docker

```
[atrisovic@antris lhcb-dataprod-demo]$ sudo docker build -t ana .
[sudo] password for atrisovic:
Sending build context to Docker daemon  620.5kB
Step 1/5 : FROM lhcbdev/slc6-build
 ---> 851fe3f67b78
Step 2/5 : MAINTAINER Ana Trisovic "ana.trisovic@cern.ch"
 ---> Using cache
 ---> a4743d12984f
Step 3/5 : WORKDIR "/workspace"
 ---> Using cache
 ---> 263ddb8c1096
Step 4/5 : COPY scripts/* /workspace/
 ---> Using cache
 ---> 812723550bd2
Step 5/5 : COPY bashrun.sh /workspace/
 ---> Using cache
 ---> d5feed22014b
Successfully built d5feed22014b
Successfully tagged ana:latest
[atrisovic@antris lhcb-dataprod-demo]$ sudo docker run -v /cvmfs:/cvmfs -it --rm ana /bin/bash
[root@94263b0f75f5 workspace]#
```

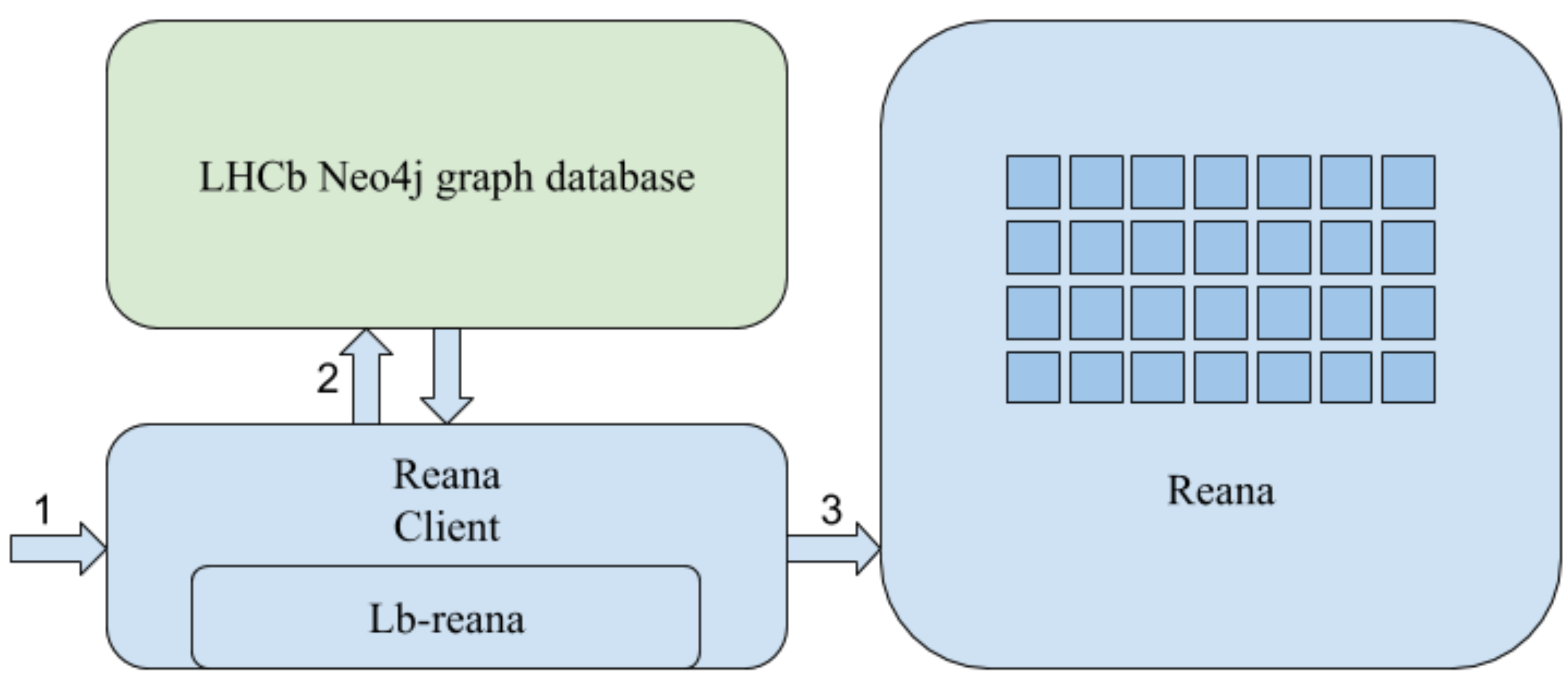
```
otal 8520
rw-r--r--. 1 root root  436010 Aug 22 16:10 00012345_00006789_1.sim
rw-r--r--. 1 root root  186477 Aug 22 16:11 00012345_00006789_2.digi
rw-r--r--. 1 root root  297953 Aug 22 16:12 00012345_00006789_3.digi
rw-r--r--. 1 root root  174035 Aug 22 16:13 00012345_00006789_4.dst
rw-r--r--. 1 root root     678 Jul 14 11:35 bashrun.sh
rw-r--r--. 1 root root  284521 Aug 22 16:11 Boole-1ev-histos.root
rw-r--r--. 1 root root 2612784 Aug 22 16:13 Brunel-1ev-histos.root
rw-r--r--. 1 root root 3149283 Aug 22 16:17 DVHistos.root
rw-r--r--. 1 root root  183213 Aug 22 16:17 finNew.dst
rw-r--r--. 1 root root  259952 Aug 22 16:10 Gauss-21113001-1ev-20170822-histos.root
rw-r--r--. 1 root root    6859 Aug 22 16:10 GeneratorLog.xml
rw-r--r--. 1 root root     331 Aug 22 16:11 pool_xml_catalogBoolev26r3.xml
rw-r--r--. 1 root root     330 Aug 22 16:13 pool_xml_catalogBrunelv43r2p7.xml
rw-r--r--. 1 root root     341 Aug 22 16:17 pool_xml_catalogDaVinciv32r2p3.xml
rw-r--r--. 1 root root     330 Aug 22 16:10 pool_xml_catalogGaussv45r3.xml
rw-r--r--. 1 root root     331 Aug 22 16:12 pool_xml_catalogMoorev12r8g3.xml
rw-r--r--. 2 root root     676 Aug 22 16:04 prodConf_Boole2.py
rw-r--r--. 2 root root     627 Aug 22 16:04 prodConf_Brunel4.py
rw-r--r--. 2 root root     718 Aug 22 16:05 prodConf_DaVinci5.py
rw-r--r--. 2 root root     777 Aug 22 16:04 prodConf_Gauss1.py
rw-r--r--. 2 root root     690 Aug 22 16:05 prodConf_Moore3.py
rw-r--r--. 1 root root    2016 Aug 22 16:11 summaryBoolev26r3.xml
rw-r--r--. 1 root root   29281 Aug 22 16:13 summaryBrunelv43r2p7.xml
rw-r--r--. 1 root root  798894 Aug 22 16:17 summaryDaVinciv32r2p3.xml
rw-r--r--. 1 root root   23505 Aug 22 16:10 summaryGaussv45r3.xml
rw-r--r--. 1 root root  205945 Aug 22 16:12 summaryMoorev12r8g3.xml
```

Further work

- Reana - **RE**producible **ANA**lysis - novel project with CERN IT to encourage automatisation in analysis

- But why not recreate-able data?

- 1) users query to python client, e.g. production ID 2) REST query to preservation db which returns production python files 3) the client starts a production on the Cloud



Chapter 5

Reproducing physics analysis

- Goal: to identify obstacles and challenges in reproducing physics analysis
- Guided by an analysis note & a PhD thesis
- Looked into a rare decay of *D(s)*+ -> *Pi*+*Mu*+*Mu*-
  - Possibly an ineffective choice

Original analysis

- Data 2011
- Monte Carlo 2010

- Code: unknown
- Done by professionals

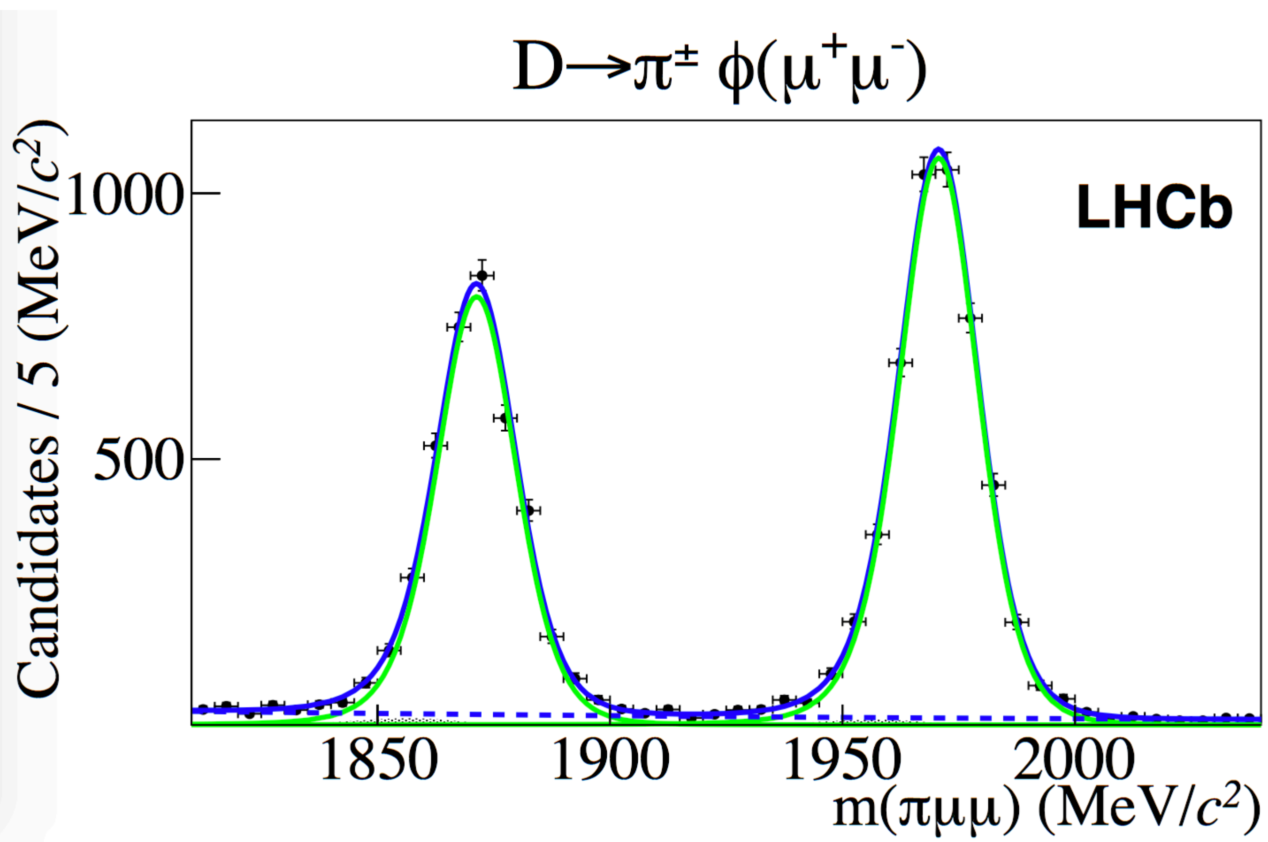
Reproduced analysis

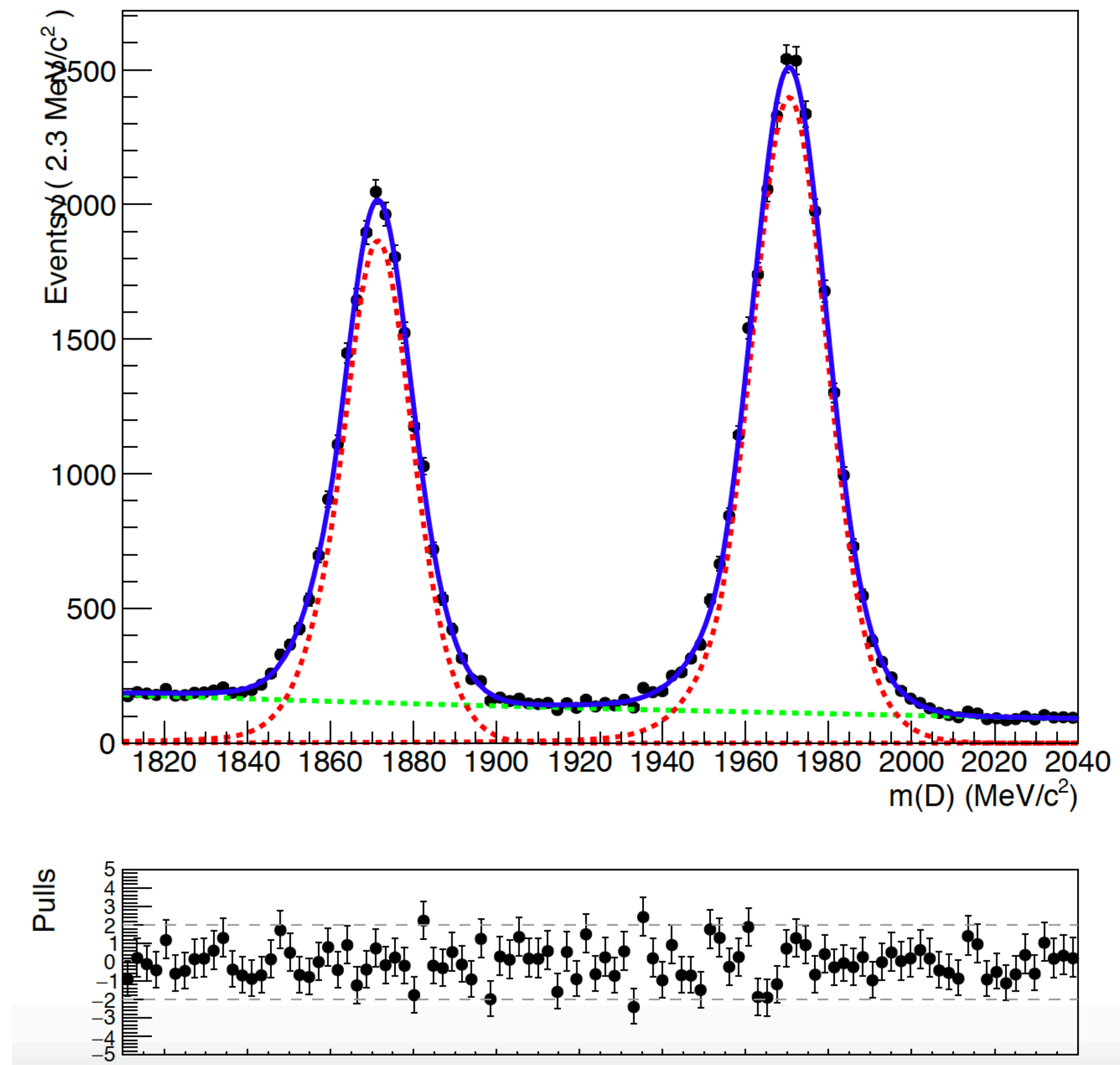
- Data 2011 and 2012
- Monte Carlo 2011 and 2012

- Code: https://github.com/atrisovic/analysis-case-study
- Used same methods as in the original analysis
- Unknown terminology and often ambiguous terms encountered
- Modest knowledge of physics analysis

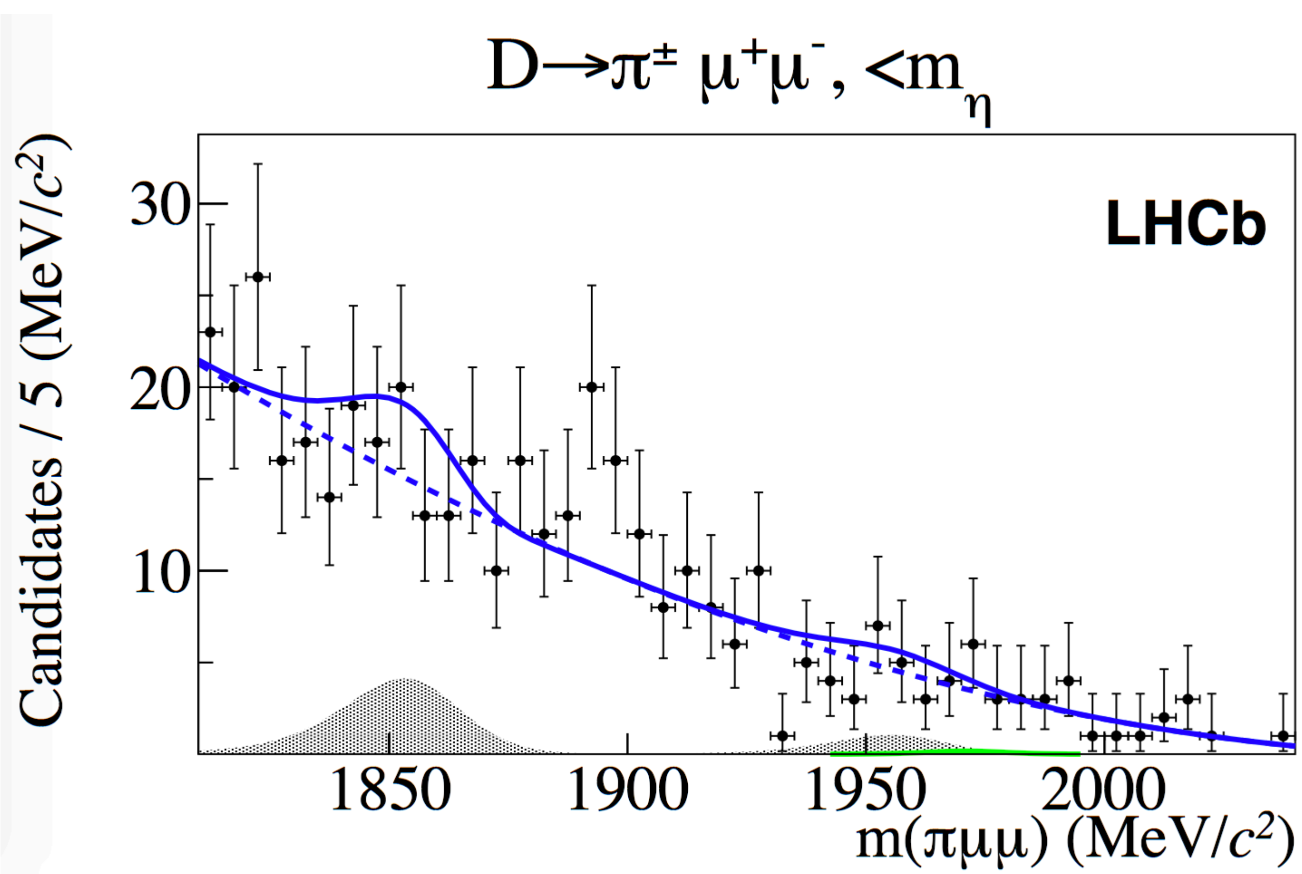
Original results

Reproduced results

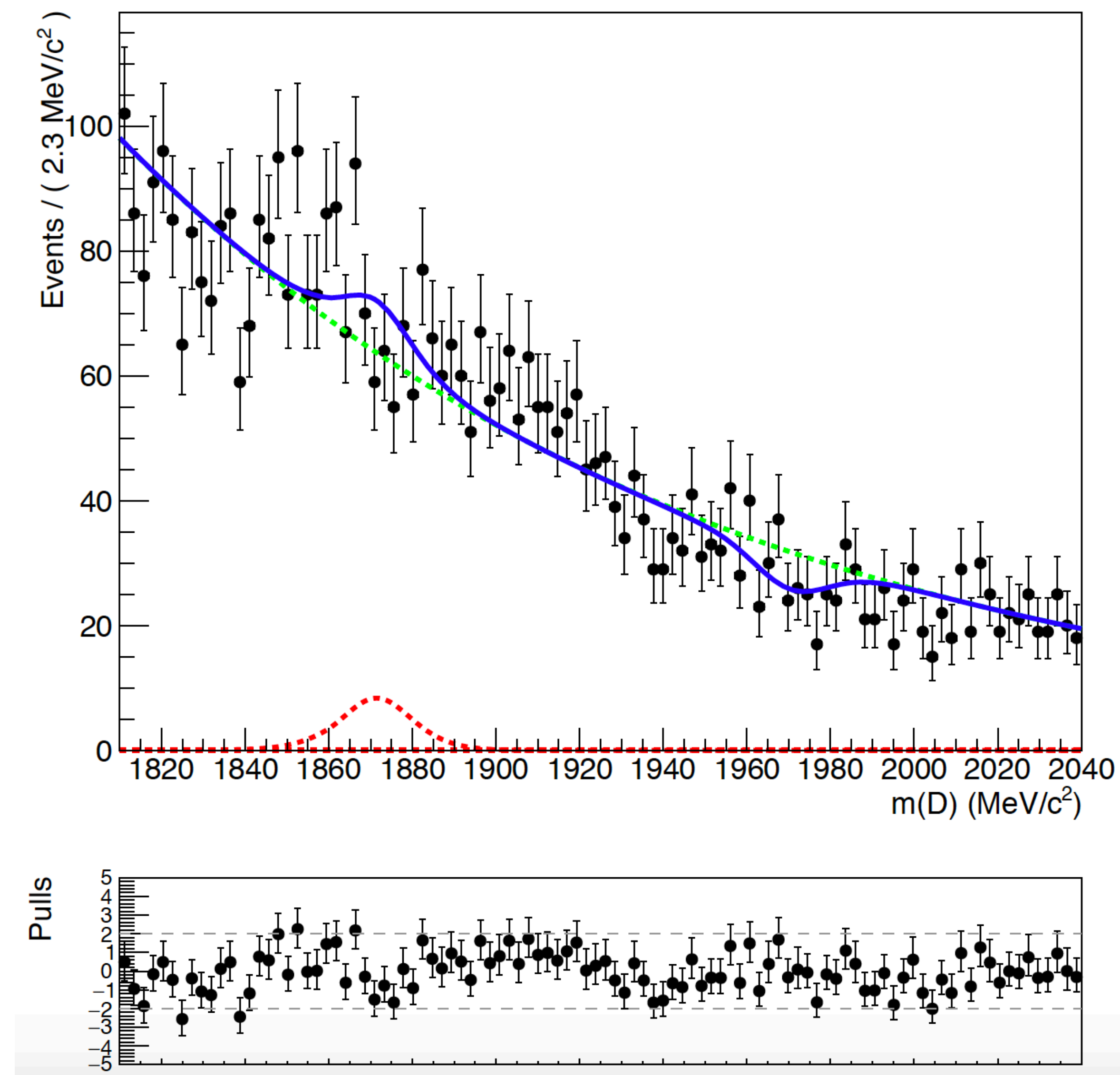




Original results

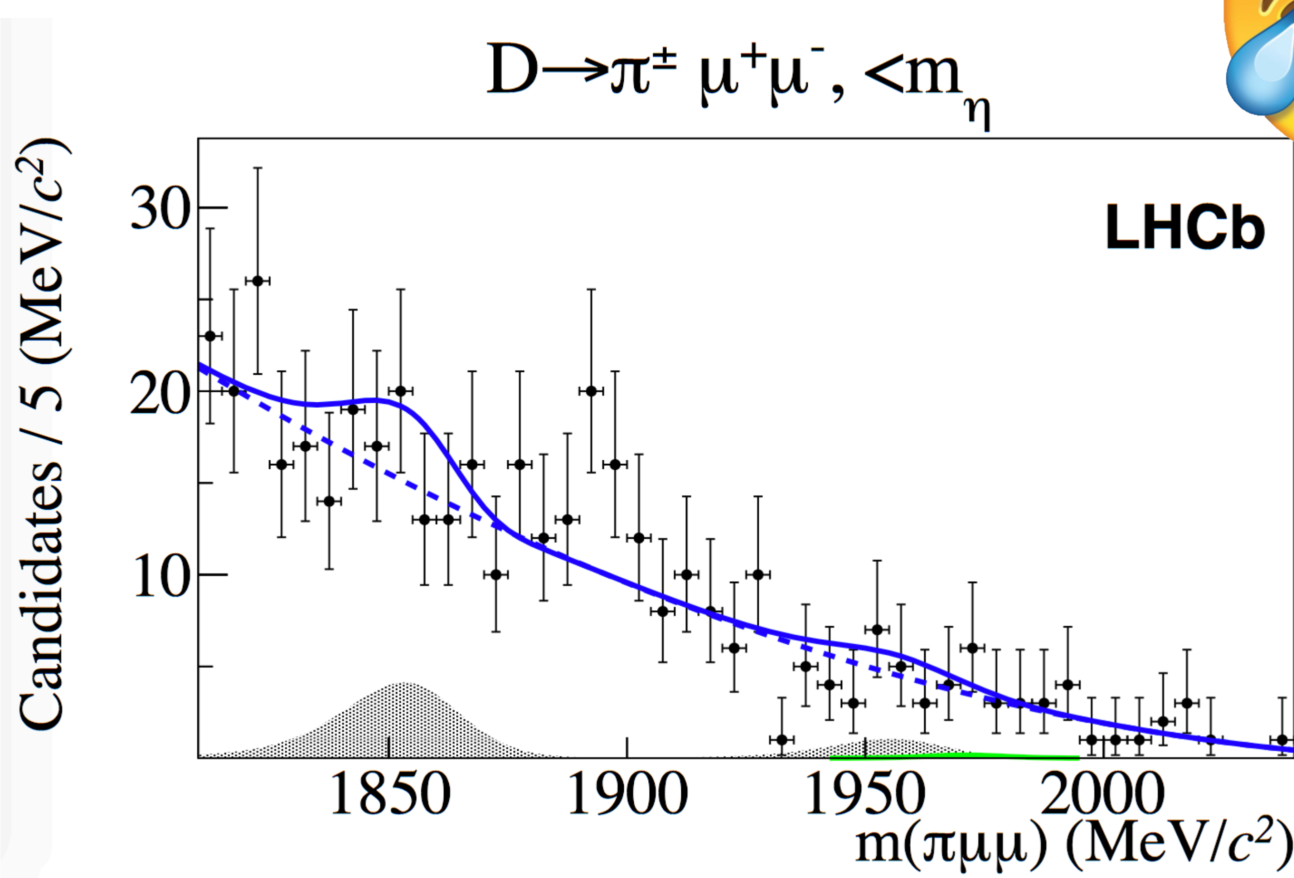


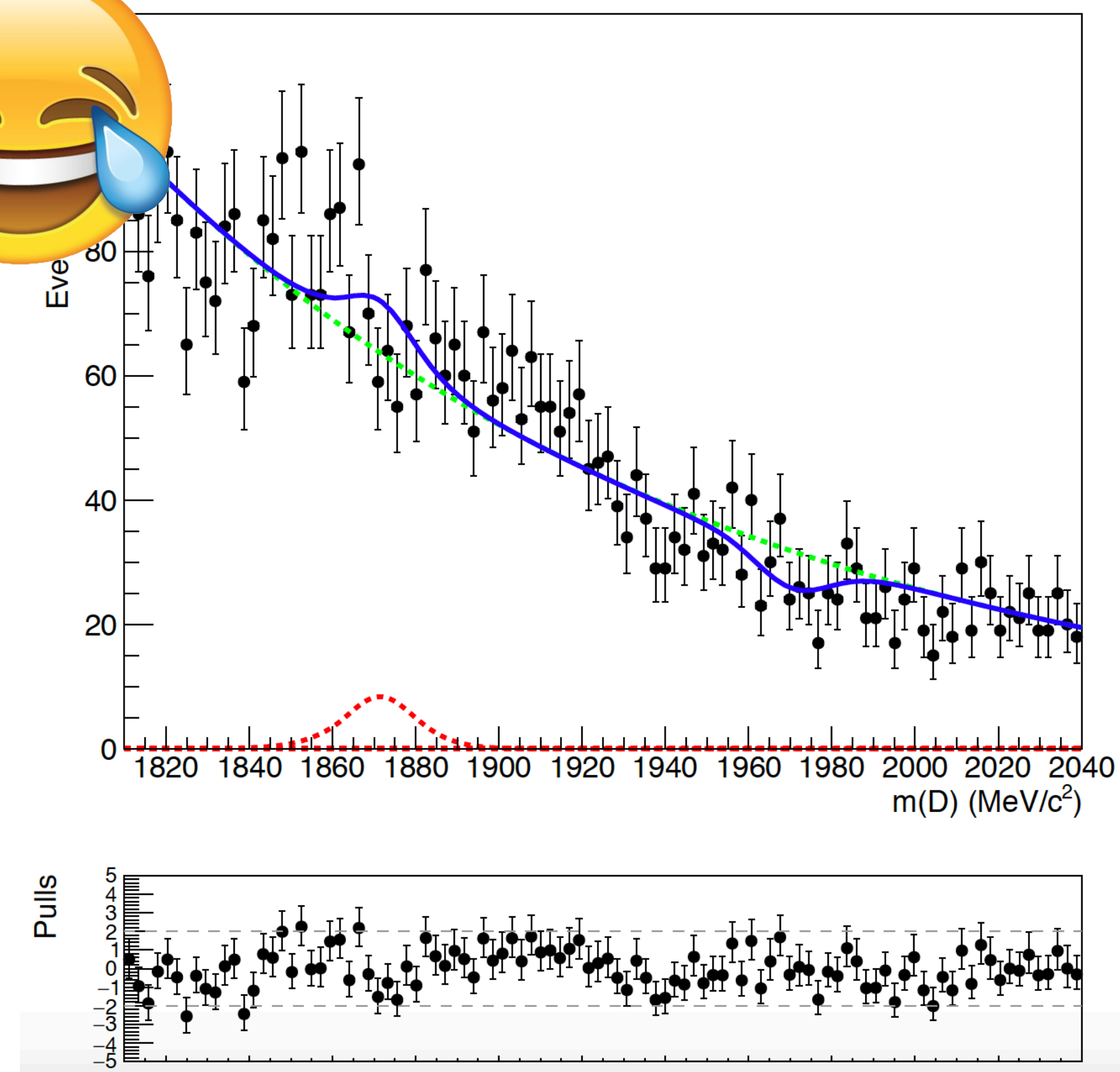
Reproduced results



Original results

Reproduced results





*this is the worst-looking plot

Chapter 6

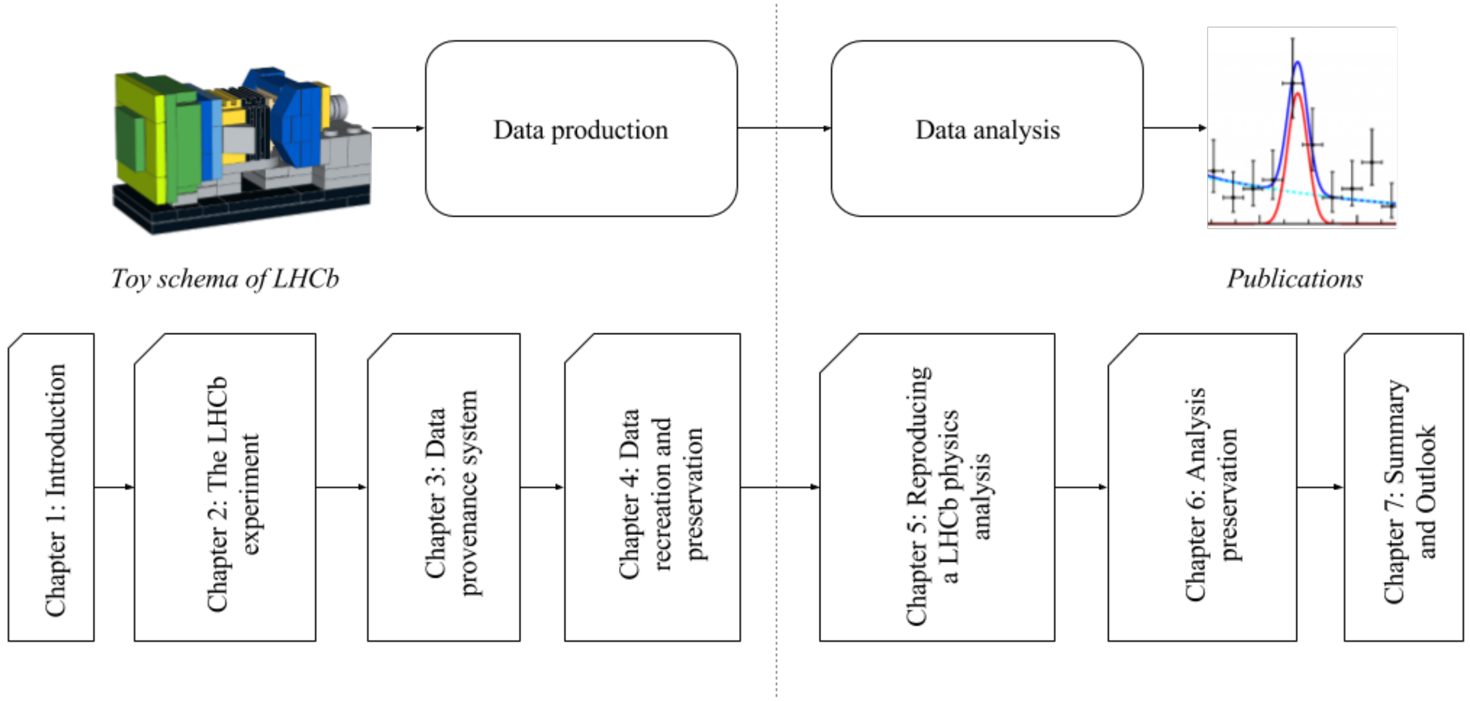
Analysis preservation

- An argument on why:
  - Version control system
  - Docker
  - Continuous integration and
  - Automated workflow

aid in analysis preservation and reproducibility and how analysts benefit using them in the long term

- In the *D* analysis this is done in make and bash

Git



- **Chapter 3**
  - To populate the preservation database: https://github.com/atrisovic/lb-preserve-scripts
  - Web server: https://github.com/atrisovic/lbc-preserve
  - Provenance tracking service in Gaudi: https://github.com/atrisovic/lhcb-metadatasvc
- **Chapter 4**
  - Data production on the CERN Cloud: https://github.com/atrisovic/lhcb-montecarlo-demo
  - Compare events: https://github.com/atrisovic/lhcb-compare-events
  - LHCb REANA client: https://github.com/atrisovic/lbreana
- **Chapter 5 & 6**
  - Reproduce and preserve physics analysis: https://github.com/atrisovic/analysis-case-study


i ii iii iv 1 2 3 4 5
6 7 8 9 10 11 12 13 14
15 16 17 18 19 20 21 22 23
24 32
The thesis is basically here
51 52 53 54 55 56 57 58 59
60 61 62 63 64 65 66 67 68
69 70 71 72 73 74 75 76 77

Publications

Recording the LHCb data and software dependencies

Ana Trisovic[1,2], Ben Couturier[1], Val Gibson[2] and Chris Jones[2]
[1] CERN, Geneva, Switzerland
[2] University of Cambridge, Cambridge, the United Kingdom
E-mail: ana.trisovic@cern.ch

Abstract.
In recent years awareness of the importance of preserving the experimental data and scientific software at CERN has been rising. To support this effort, we are presenting a novel approach to structure dependencies of the LHCb data and software to make it more accessible in the long-term future. In this paper, we detail the implementation of a graph database of these dependencies. We list the implications that can be deduced from the graph mining (such as a search for the legacy software), with emphasis on data preservation. Furthermore, we introduce a methodology of recreating the LHCb data, thus supporting reproducible research and data stewardship. Finally, we describe how this information is made available to the users on a web portal that promotes data and analysis preservation and good practise with analysis documentation.

Keywords— data preservation, data provenance, graph database

1. Introduction

The Large Hadron Collider beauty (LHCb) experiment at CERN is a general purpose detector in the forward region, which focuses on investigating the differences between matter and antimatter by studying the decays of beauty (B) and charm (D) mesons. The detector has been recording data from proton-proton collisions since 2010 and is expected to record data throughout the 2020s. Due to the rapid development of both the hardware and software used to process the data, many questions have been raised about data compatibility and preservation. In response to this, we are creating a database to record the metadata of our software and the data provenance. The recorded dependencies are expected to ease the process of running the software and analysing the data in the long-term future.

1.1. Data preservation initiative

The data preservation project in High Energy Physics (HEP) aims to ensure the preservation of experimental and simulated data, as well as the scientific software and documentation. The main objective is to assist analysing HEP data in the future. The major use cases include looking for signals predicted by new theories and improving current measurements, in addition to physics outreach and educational purposes.

The LHCb data is processed with software and hardware that are changing over time. The information about the data, software and the changes have been logged in the internal databases

www.nature.com/scientificdata

SCIENTIFIC DATA

OPEN

Comment: If these data could talk

Thomas Pasquier[1], Matthew K. Lau[2], Ana Trisovic[3,4], Emery Boose[2], Ben Couturier[3], Mercè Crosas[5], Aaron M. Ellison[2], Valerie Gibson[4], Chris Jones[4] & Margo Seltzer[1]

Received: 12 April 2017
Accepted: 24 July 2017
Published: xx xxx 2017

In the last few decades, data-driven methods have come to dominate many fields of scientific inquiry. Open data and open-source software have enabled the rapid implementation of novel methods to manage and analyze the growing flood of data. However, it has become apparent that many scientific fields exhibit distressingly low rates of repeatability and reproducibility. Although there are many dimensions to this issue, we believe that there is a lack of formalism used when describing end-to-end published results, from the data source to the analysis to the final published results. Even when authors do their best to make their research and data accessible, this lack of formalism reduces the clarity and efficiency of reporting, which contributes to issues of reproducibility. Data provenance aids both repeatability and reproducibility through *systematic* and *formal* records of the relationships among data sources, processes, datasets, publications and researchers.

In the last few decades, data-driven methods have come to dominate many fields of scientific inquiry. Open data and open-source software have enabled the rapid implementation of novel methods to manage and analyze the growing flood of data. However, it has become apparent that many scientific fields exhibit distressingly low rates of repeatability and reproducibility. Although there are many dimensions to this issue, we believe that there is a lack of formalism used when describing end-to-end published results, from the data source to the analysis to the final published results. Even when authors do their best to make their research and data accessible, this lack of formalism reduces the clarity and efficiency of reporting, which contributes to issues of reproducibility. Data provenance aids both repeatability and reproducibility through *systematic* and *formal* records of the relationships among data sources, processes, datasets, publications and researchers.

Reproducibility & Repeatability

The success and power of science depends on the transparency and validation of its findings. However, issues with reproducibility have surfaced across a broad swath of scientific disciplines. Reports of such issues have emanated from fields ranging from the social sciences to physics and the life-sciences, including medicine[1]. Although the lack of reproducibility does not necessarily imply incorrect results[2], it remains a worrisome issue. This comes at a time when the rate of scientific publication is increasing exponentially[3]. At the same time, the data and the processes that produce results are becoming more computationally demanding.

Reproducibility is the cornerstone of science, so it is imperative that we improve the quality and reliability of publications by going beyond the publication of results and data to making analytical processes, not only available, but more importantly, intelligible[4]. Too often, despite the best efforts of authors, transparency, adequate for the replication of computational processes, is elusive. We advocate open-data, open-source and *open-process*, which we define as the formal record of the workflow that produced a result. Changes to the pipeline that transforms raw data to results can lead to non-trivial differences in results, which are impossible to explain without sufficient reporting. For example, a re-examination of studies of carbon flux in forested ecosystems in the Amazon detected differences in estimates up to 140%, which could mean as much as 7 metric tons of carbon per year in an area roughly the size of a football field, resulting from small differences in analytical pipelines[5]. Also, seemingly simple details, such as the version of the initial (raw) data or versions of the analytical software programs, are

[1]School of Engineering and Applied Sciences, Harvard University. [2]Harvard Forest, Harvard University. [3]European Organization for Nuclear Research (CERN), Geneva, Switzerland. [4]Cavendish Laboratory, University of Cambridge. [5]Institute for Quantitative Social Science, Harvard University. Correspondence and requests for materials should be addressed to T.P. (email: tfjmp@seas.harvard.edu). Q1 Q2

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH (CERN)

CERN-EP-20XX-YYY
LHCb-PAPER-20XX-YYY
June 8, 2017

LHCb Analysis Preservation Roadmap

The LHCb collaboration†

Abstract

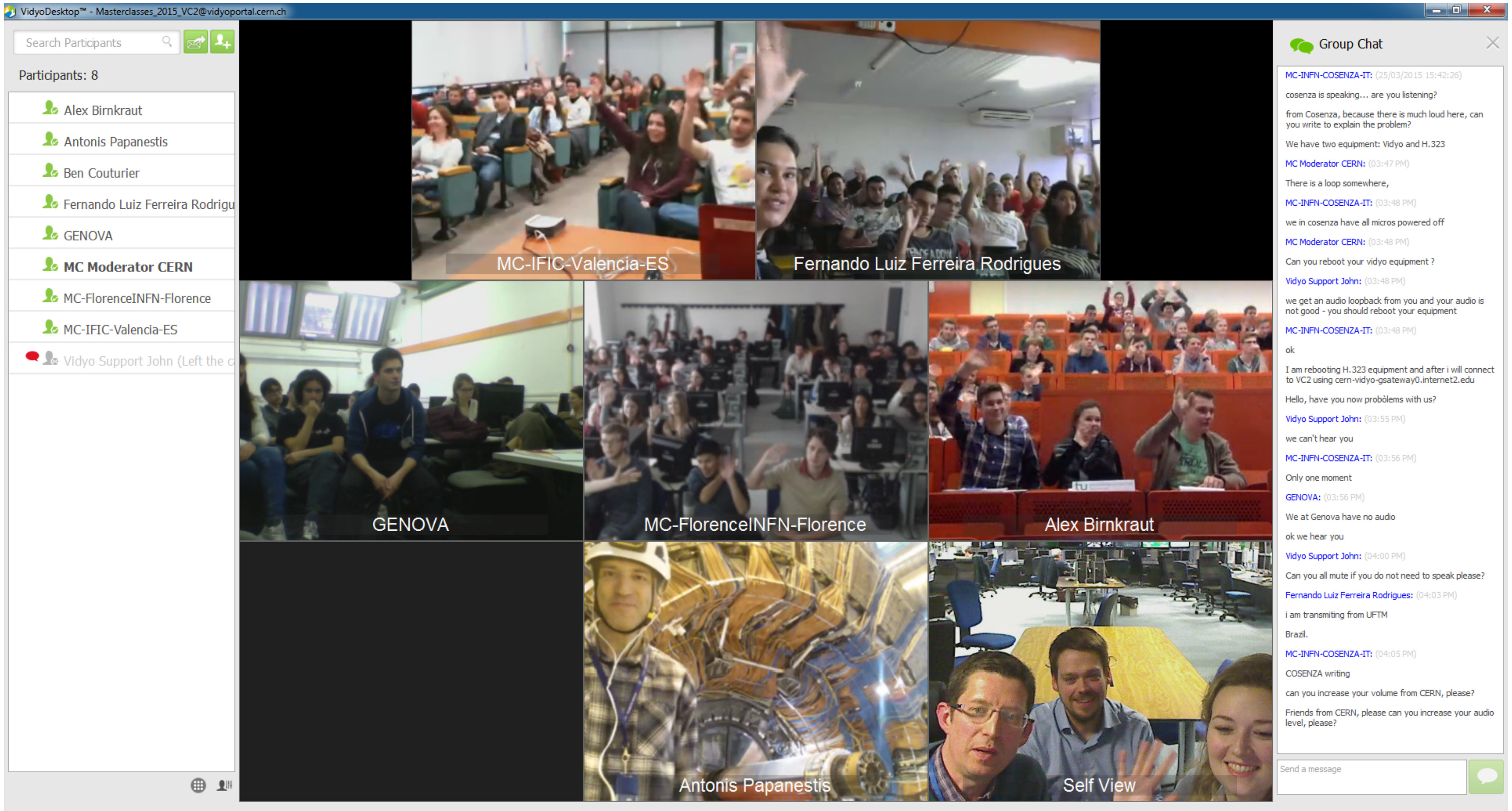
This document presents recommendations from the LHCb analysis preservation and collaborative working groups. The document is divided in two parts. Part I outlines a set of best practices for analysis preservation. Part II presents recommendations on which tools and technologies should be used in order to implement these practices.

The practices are categorized into four domains: analysis repository, analysis pipelines, run-time environment and data preservation. In each of those four domains a minimal set of practices is identified, which forms the basis of all analysis preservation efforts. In addition we recommend an expanded set of best practices, which are needed to be able to host an analysis on the central CERN analysis preservation infrastructure.

†Authors are listed at the end of this paper.

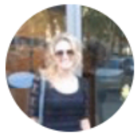
Other activities

- CERN analysis preservation group CAP
- CERN Collaboration spotting
- CERN open days, TEDxCERN, Django girls
- CERN guide
- LHCb masterclass



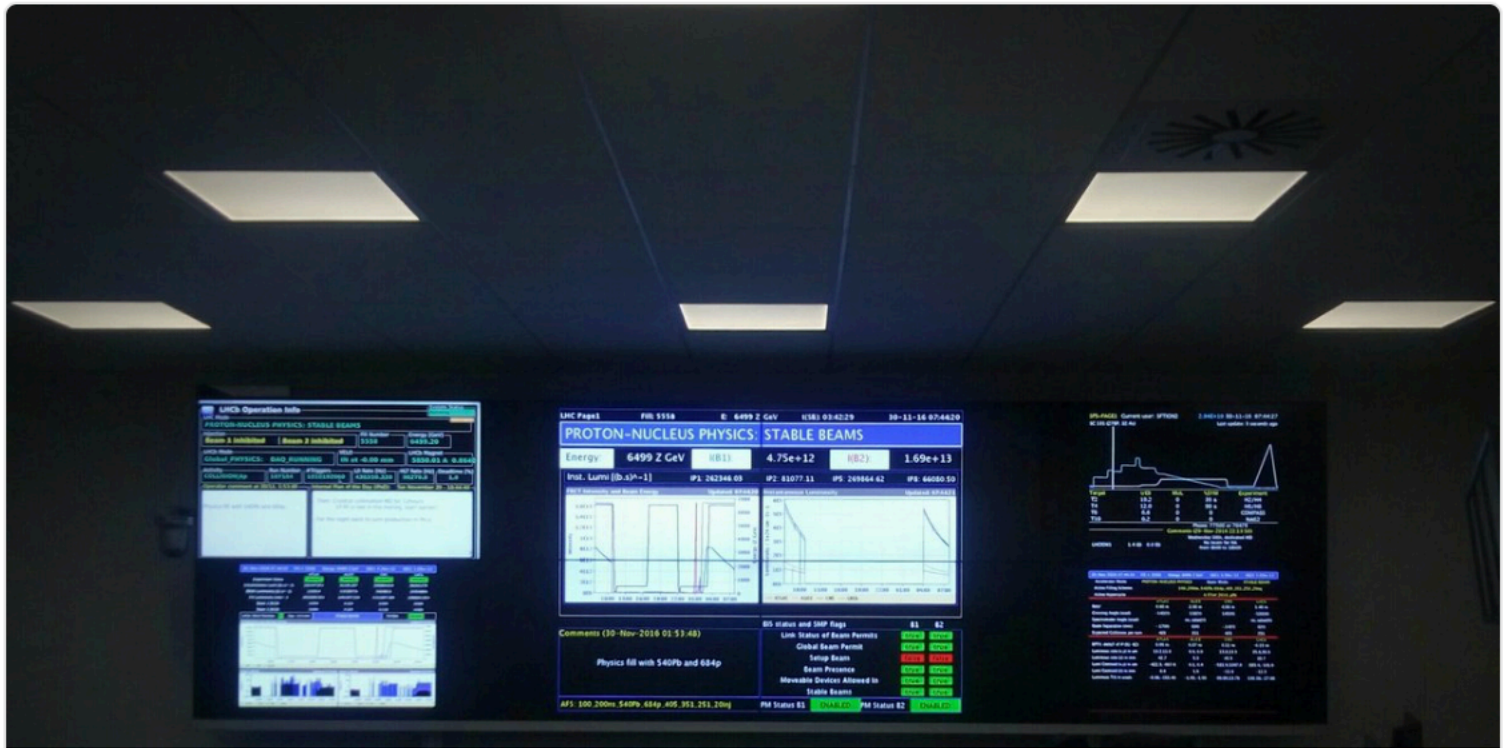
and of course:

Shifts



Ana Trisovic
@anatrisovic

Things are looking good at the #LHCb experiment this morning! 💫💫💫 #stablebeams



What I've learned

LHCb

git

python

docker

neo4j

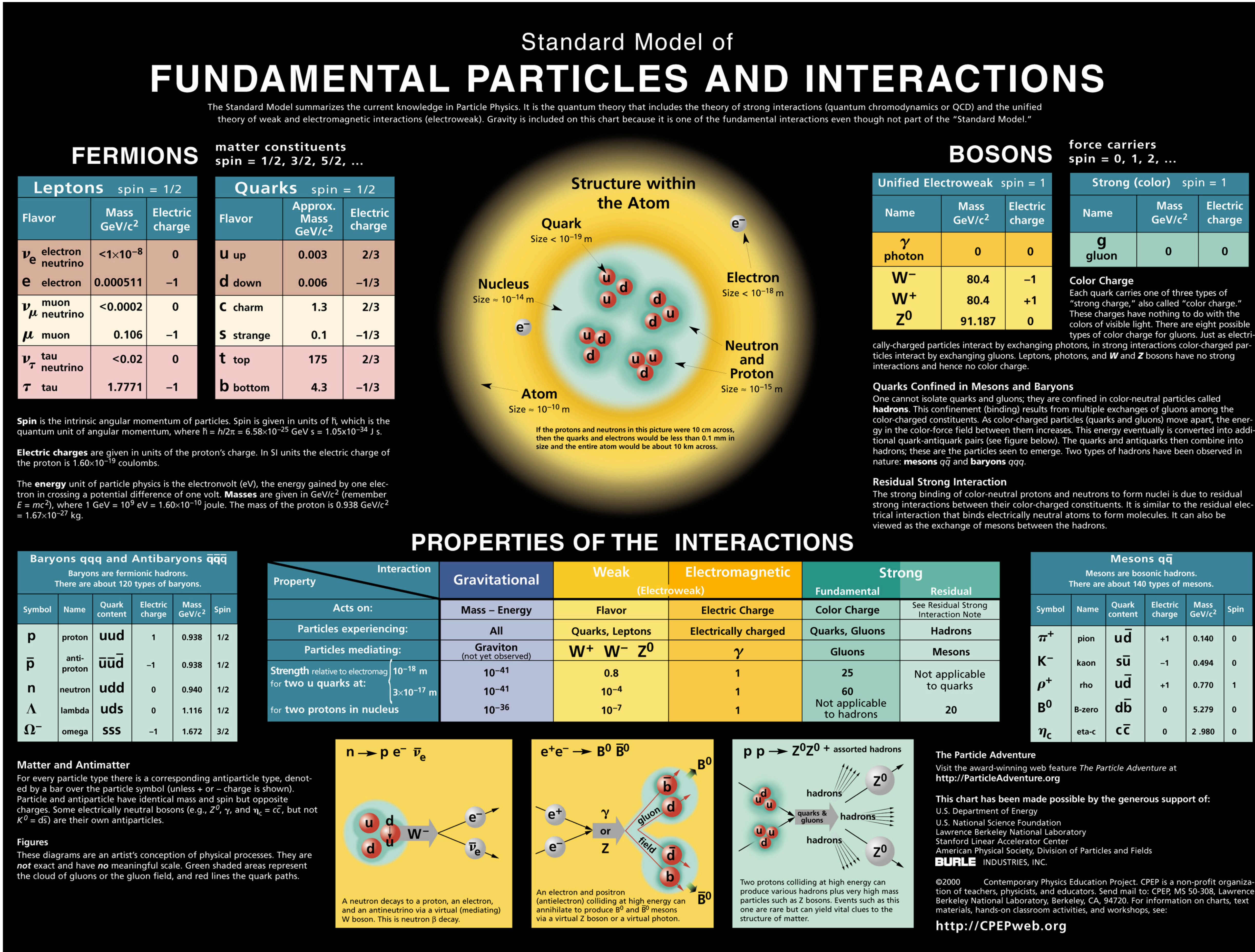
Vim

Flask

ROOT
Data Analysis Framework

and to cook

and of course:



Standard Model of

FUNDAMENTAL PARTICLES AND INTERACTIONS

The Standard Model summarizes the current knowledge in Particle Physics. It is the quantum theory that includes the theory of strong interactions (quantum chromodynamics or QCD) and the unified theory of weak and electromagnetic interactions (electroweak). Gravity is included on this chart because it is one of the fundamental interactions even though not part of the "Standard Model."

FERMIONS

matter constituents
spin = 1/2, 3/2, 5/2, ...

Leptons spin = 1/2

Flavor	Mass GeV/c²	Electric charge
ν_e electron neutrino	$<1\times10^{-8}$	0
e electron	0.000511	–1
ν_μ muon neutrino	<0.0002	0
μ muon	0.106	–1
ν_τ tau neutrino	<0.02	0
τ tau	1.7771	–1

Quarks spin = 1/2

Flavor	Approx. Mass GeV/c²	Electric charge
u up	0.003	2/3
d down	0.006	–1/3
c charm	1.3	2/3
s strange	0.1	–1/3
t top	175	2/3
b bottom	4.3	–1/3

Spin is the intrinsic angular momentum of particles. Spin is given in units of ħ, which is the quantum unit of angular momentum, where $\hbar = h/2\pi = 6.58\times10^{-25}$ GeV s $= 1.05\times10^{-34}$ J s.

Electric charges are given in units of the proton's charge. In SI units the electric charge of the proton is 1.60×10^{-19} coulombs.

The **energy** unit of particle physics is the electronvolt (eV), the energy gained by one electron in crossing a potential difference of one volt. **Masses** are given in GeV/c^2 (remember $E = mc^2$), where 1 GeV = 10^9 eV = 1.60×10^{-10} joule. The mass of the proton is 0.938 GeV/c^2 = 1.67×10^{-27} kg.

Structure within the Atom

If the protons and neutrons in this picture were 10 cm across, then the quarks and electrons would be less than 0.1 mm in size and the entire atom would be about 10 km across.

BOSONS

force carriers
spin = 0, 1, 2, ...

Unified Electroweak spin = 1

Name	Mass GeV/c²	Electric charge
γ photon	0	0
W^-	80.4	–1
W^+	80.4	+1
Z^0	91.187	0

Strong (color) spin = 1

Name	Mass GeV/c²	Electric charge
g gluon	0	0

Color Charge
Each quark carries one of three types of "strong charge," also called "color charge." These charges have nothing to do with the colors of visible light. There are eight possible types of color charge for gluons. Just as electrically-charged particles interact by exchanging photons, in strong interactions color-charged particles interact by exchanging gluons. Leptons, photons, and ***W*** and ***Z*** bosons have no strong interactions and hence no color charge.

Quarks Confined in Mesons and Baryons
One cannot isolate quarks and gluons; they are confined in color-neutral particles called **hadrons**. This confinement (binding) results from multiple exchanges of gluons among the color-charged constituents. As color-charged particles (quarks and gluons) move apart, the energy in the color-force field between them increases. This energy eventually is converted into additional quark-antiquark pairs (see figure below). The quarks and antiquarks then combine into hadrons; these are the particles seen to emerge. Two types of hadrons have been observed in nature: **mesons** $q\bar{q}$ and **baryons** qqq.

Residual Strong Interaction
The strong binding of color-neutral protons and neutrons to form nuclei is due to residual strong interactions between their color-charged constituents. It is similar to the residual electrical interaction that binds electrically neutral atoms to form molecules. It can also be viewed as the exchange of mesons between the hadrons.

Baryons qqq and Antibaryons $\bar{q}\bar{q}\bar{q}$
Baryons are fermionic hadrons.
There are about 120 types of baryons.

Symbol	Name	Quark content	Electric charge	Mass GeV/c²	Spin
p	proton	uud	1	0.938	1/2
$\bar{p}$	anti-proton	$\bar{u}\bar{u}\bar{d}$	–1	0.938	1/2
n	neutron	udd	0	0.940	1/2
Λ	lambda	uds	0	1.116	1/2
Ω^-	omega	sss	–1	1.672	3/2

PROPERTIES OF THE INTERACTIONS

Property \ Interaction	Gravitational	Weak (Electroweak)	Electromagnetic (Electroweak)	Strong: Fundamental	Strong: Residual
Acts on:	Mass – Energy	Flavor	Electric Charge	Color Charge	See Residual Strong Interaction Note
Particles experiencing:	All	Quarks, Leptons	Electrically charged	Quarks, Gluons	Hadrons
Particles mediating:	Graviton (not yet observed)	W^+ W^- Z^0	γ	Gluons	Mesons
Strength relative to electromag for two u quarks at: 10^{-18} m	10^{-41}	0.8	1	25	Not applicable to quarks
Strength relative to electromag for two u quarks at: 3×10^{-17} m	10^{-41}	10^{-4}	1	60	
for two protons in nucleus	10^{-36}	10^{-7}	1	Not applicable to hadrons	20

Mesons $q\bar{q}$
Mesons are bosonic hadrons.
There are about 140 types of mesons.

Symbol	Name	Quark content	Electric charge	Mass GeV/c²	Spin
π^+	pion	$u\bar{d}$	+1	0.140	0
K^-	kaon	$s\bar{u}$	–1	0.494	0
ρ^+	rho	$u\bar{d}$	+1	0.770	1
B^0	B-zero	$d\bar{b}$	0	5.279	0
η_c	eta-c	$c\bar{c}$	0	2 .980	0

Matter and Antimatter
For every particle type there is a corresponding antiparticle type, denoted by a bar over the particle symbol (unless + or – charge is shown). Particle and antiparticle have identical mass and spin but opposite charges. Some electrically neutral bosons (e.g., Z^0, γ, and $\eta_c = c\bar{c}$, but not $K^0 = d\bar{s}$) are their own antiparticles.

Figures
These diagrams are an artist's conception of physical processes. They are *not* exact and have ***no*** meaningful scale. Green shaded areas represent the cloud of gluons or the gluon field, and red lines the quark paths.

A neutron decays to a proton, an electron, and an antineutrino via a virtual (mediating) W boson. This is neutron β decay.

An electron and positron (antielectron) colliding at high energy can annihilate to produce B^0 and $\bar{B}^0$ mesons via a virtual Z boson or a virtual photon.

Two protons colliding at high energy can produce various hadrons plus very high mass particles such as Z bosons. Events such as this one are rare but can yield vital clues to the structure of matter.

The Particle Adventure
Visit the award-winning web feature *The Particle Adventure* at
http://ParticleAdventure.org

This chart has been made possible by the generous support of:
U.S. Department of Energy
U.S. National Science Foundation
Lawrence Berkeley National Laboratory
Stanford Linear Accelerator Center
American Physical Society, Division of Particles and Fields
BURLE INDUSTRIES, INC.

http://CPEPweb.org

Achievements*

CERN Bulletin

Issue No. 9-10/2017-Wednesday 01 March 2017
More articles at *http://home.cern/cern-people*

COLLIDING IDEAS OVER LUNCH

The LunchCollider project sees strangers brought together in Restaurant 1 to build new friendships. (Image: Piotr Nikiel; Mietek Dabrowski; Roberto Campesato; Esther Zanon/CERN)

It's a typical, cloudy February morning on CERN's Meyrin site. After hours spent hunched over your computer, a quick glance at the clock tells you it's 12 p.m. –

"We would be extremely happy if new CERN projects or new physics theories came to life thanks to LunchCollider," says Mietek Dabrowski, one of the founders of LunchCollider.

"Recently I had lunch with a person I've never met before. I had no idea who this person was, no idea where they came from, and no idea what they did at CERN," says one participant after their first experience of using LunchCollider. "But one hour and fifteen minutes passed very quickly. The list of topics we wanted to talk about, from our home countries to how we ended up at CERN, was long! Now we have to go back to work. That was an amazing experience; I'll do it again. Hopefully at 12 p.m. on Wednesday next week."

The lunch sessions are organised twice a week – either on Tuesdays and Thursdays or Wednesdays and Fridays. Between 8.30

Nikiel, Roberto Campesato and Esther Zanon. "We were bound together by a common vision," explains Piotr. "What if, in such a diverse and heterogeneous place like CERN, everyone could seamlessly exchange knowledge, ideas and interests with everybody else?"

The creators' hope to build not only friendships but also professional networks – places to seek and receive information.

"We understand that meeting strangers might be a bit frightening, but we want to fight that fear. Some people have admitted that they are slightly nervous before the meeting – they worry that the conversation won't flow smoothly. This never turns out to be the case and, in the end, everybody enjoys it," says Roberto.

The concept of meeting kind, smiling, albeit

*not really

Fun times

Occasional quick chess with Stavros

Very lucky to have shared the office with Mariana Rihl, Victor Coco, Stavros Moiras and Sebastian Ponce

The most fun with Agnieszka

Mountains are also kind of cool

Apprentice to Kazu & Rainer

Ana Trisovic 3.0

Fun times

Occasional quick chess with Stavros

Very lucky to have shared the office with Mariana Rihl, Victor Coco, Stavros Moiras and Sebastian Ponce

The most fun with Agnieszka

Mountains are also kind of cool

Apprentice to Kazu & Rainer

Ana Trisovic 3.0

Feedback & ideas

- Applaud to Marco for organising our LBC meetings
- Struggles in the beginning of PhD
  - but happy to have worked on this novel and important project
- It’s difficult to learn about the experiment and the research lifecycle as a newcomer
  - I’ve been (occasionally) a mediator between the groups
  - Early training to git & jira
  - TDR computing is an excellent literature
  - Maybe a newcomers guide?

What next?

- My plans after CERN are … undecided
- I’ll go back to Cambridge to graduate
- But then: working in industry is really cool, working at CERN is amazing
- Currently I am inclined to apply for postdoc positions in the US
- …but I will also apply for some other jobs… probably

Acknowledgements

- Thank you to the group for being always ready to help
- Particularly thank you to: Ben Couturier, Zoltan Mathe, Marco Cattaneo, Gloria Corti, Marco Clemencic, Stefan Roiser, Philippe Charpentier

Thank you for your attention!

My email when I leave CERN: ana.trisovic@gmail.com
I am on twitter as @atrisovic
Also on GitHub as @atrisovic