

Chassez la théorie...

Production, stockage et partage des
données de la recherche au LHC

Simon Dumas Primbault (CNRS, OpenEdition)

GdR ModMat, Banyuls-sur-Mer

22.08.2024

Introduction. Data: A Fourth Paradigm in Science?

CHRIS ANDERSON

SCIENCE JUN 23, 2008 12:00 PM

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Illustration: Marian Bantjes “All models are wrong, but some are useful.” So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies [...]



Introduction. Data: A Fourth Paradigm in Science?

CHRIS ANDERSON

SCIENCE JUN 23, 2008 12:00 PM

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

Illustration: Marian Bantjes “All models are wrong, but some are useful.” So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies [...]

“The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all.

There's no reason to cling to our old ways. It's time to ask: What can science learn from Google?”



Introduction. Data: A Fourth Paradigm in Science?

2009

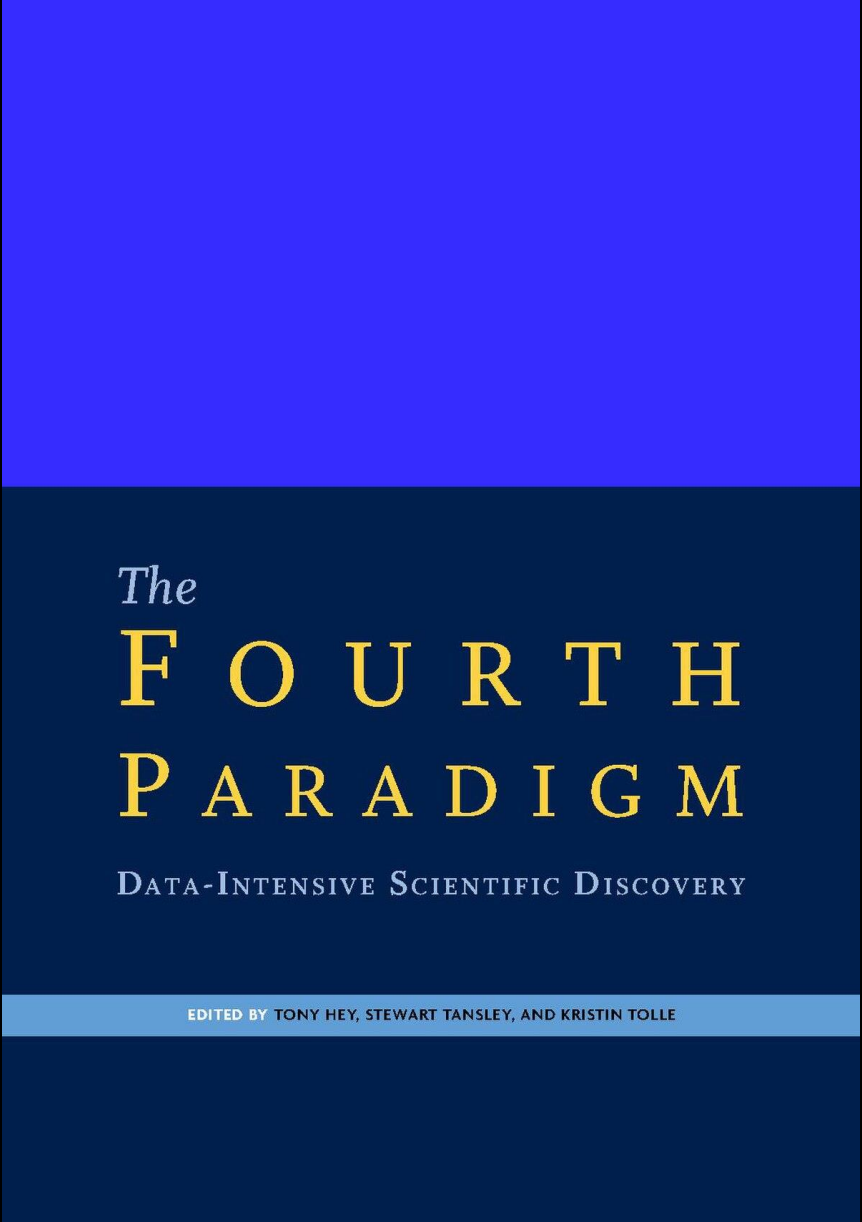
The
F O U R T H
P A R A D I G M
DATA-INTENSIVE SCIENTIFIC DISCOVERY

The book cover features a solid blue background. The title 'The Fourth Paradigm' is centered in a large, white, serif font, with 'The' in a smaller size above 'FOURTH'. Below the title, the subtitle 'DATA-INTENSIVE SCIENTIFIC DISCOVERY' is written in a smaller, white, sans-serif font. At the bottom of the cover, a thin white horizontal band contains the text 'EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE' in a small, white, sans-serif font.

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Introduction. Data: A Fourth Paradigm in Science?

2009



STEPHEN WOLFRAM

Writings



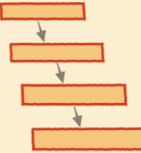

ABOUT WRITINGS

RECENT | CATEGORIES | Q

Contents

Multicomputation: A Fourth Paradigm for Theoretical Science

September 9, 2021

			
structural (antiquity)	mathematical (1600s)	computational (1980s)	multicomputational (2020s)
e.g. geometrical elements	e.g. differential equations	e.g. cellular automata	e.g. multiway systems
explicit time not considered	time as mathematical coordinate	time as progress of computation	many computational threads of time
static facts deduced by reasoning	find behavior at any time from formula	determine future only by running program	need model of observer to determine state

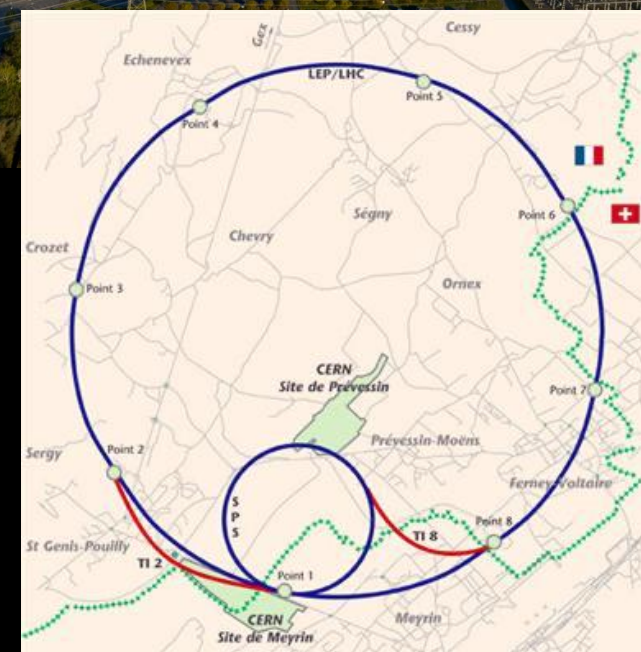
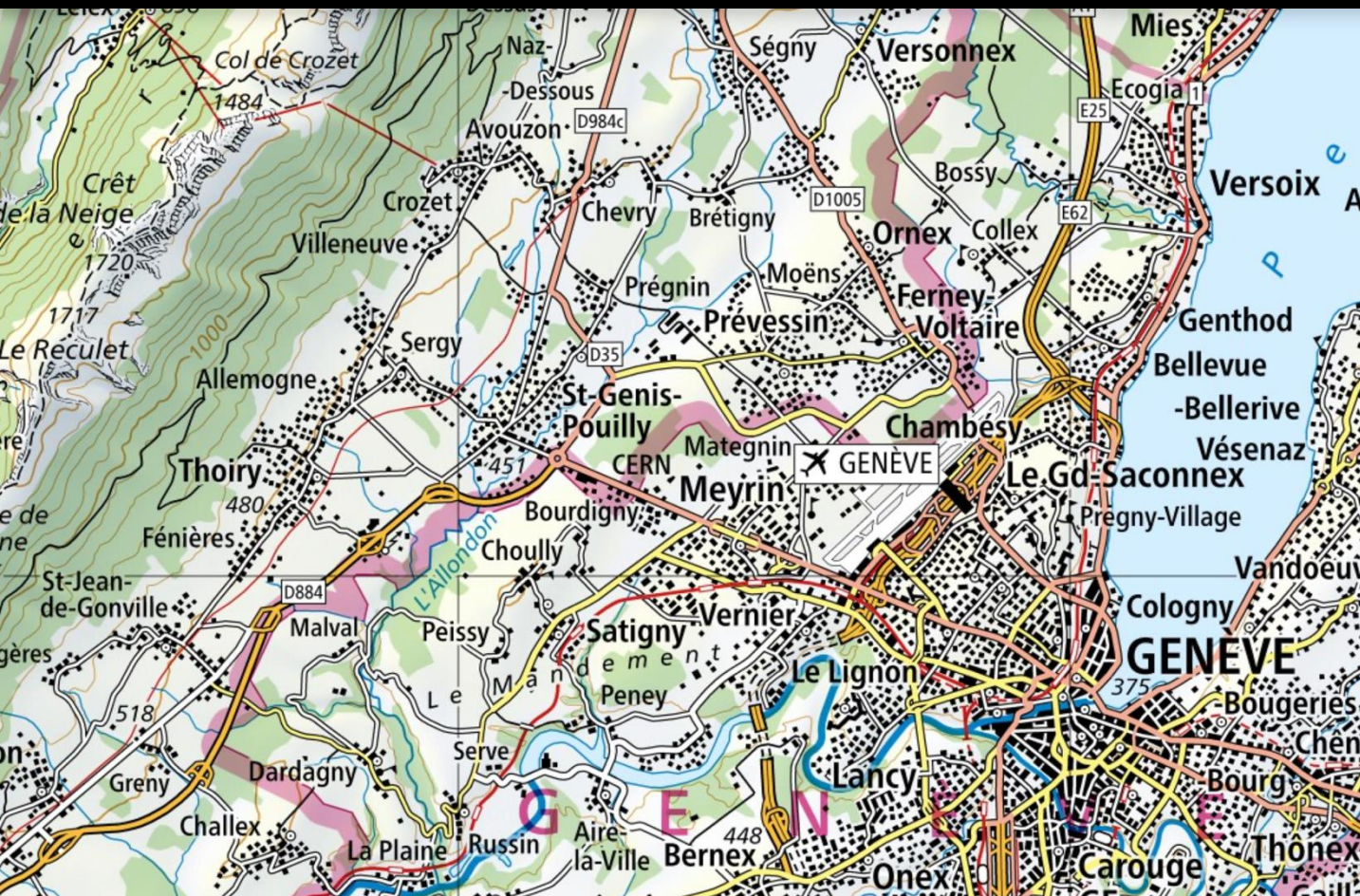
Introduction. Data: A Fourth Paradigm in Science?

Assumptions:

1. Radical empiricism
2. Radical inductivism
3. Neutrality of computational techniques
4. Correlation supersedes causality

Big Science: CERN as a data infrastructure

CERN (1954—today)



Big Science: CERN as a data infrastructure

CERN (1954—today)

1950: creation of CECA

1952: temporary council

1954: creation of CERN

1957 and 1960: first accelerators

1965: extension on French soil

1989: LEP

2008: LHC



Big Science: CERN as a data infrastructure

Big Science as a scientific regime

Domains: nuclear physics, space, biology

Applications: energies, transports, new materials...

Unprecedented scales

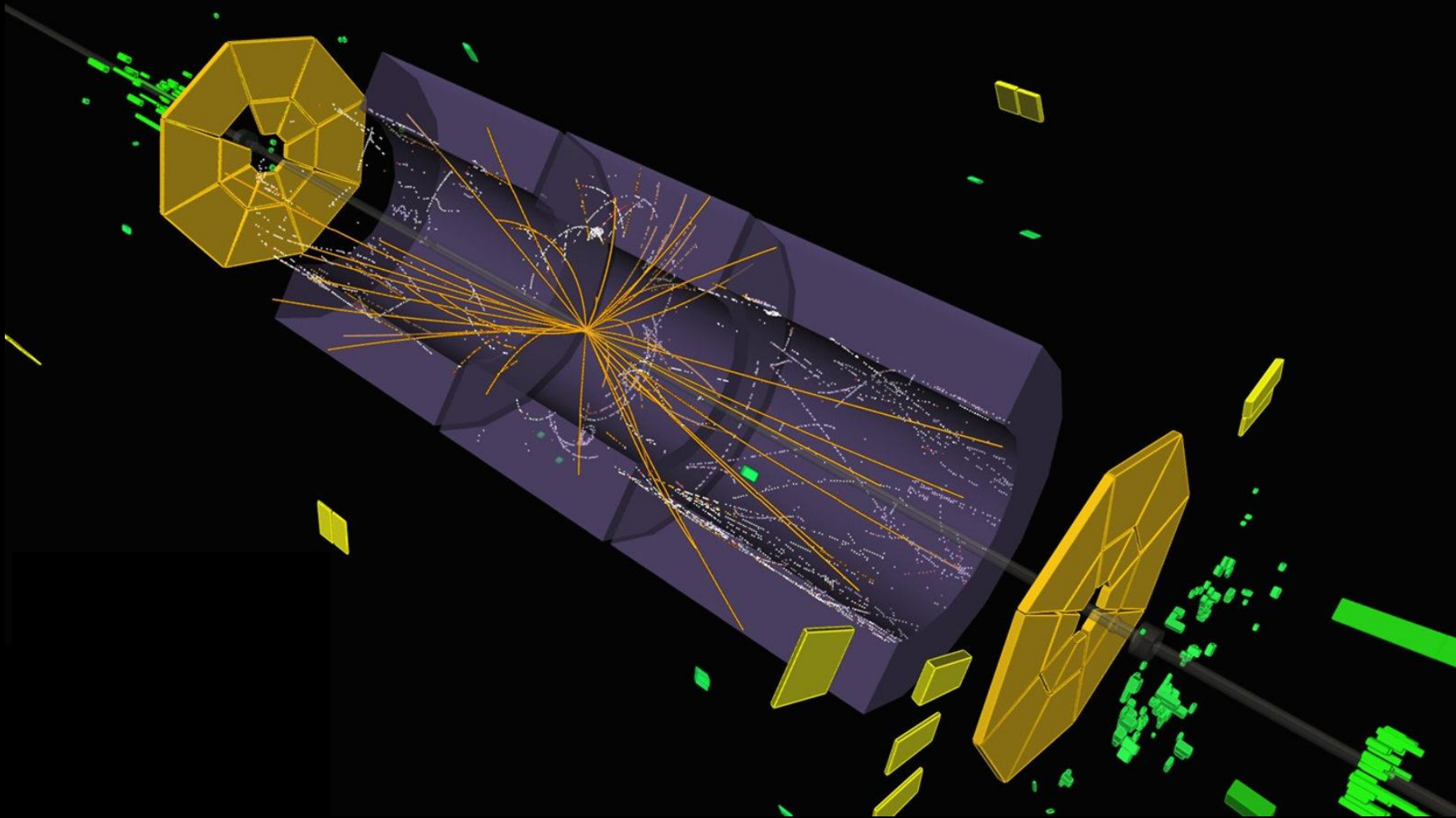
on all counts

Political construction

and, reciprocally, tool for political construction

Strong ties with technological industries

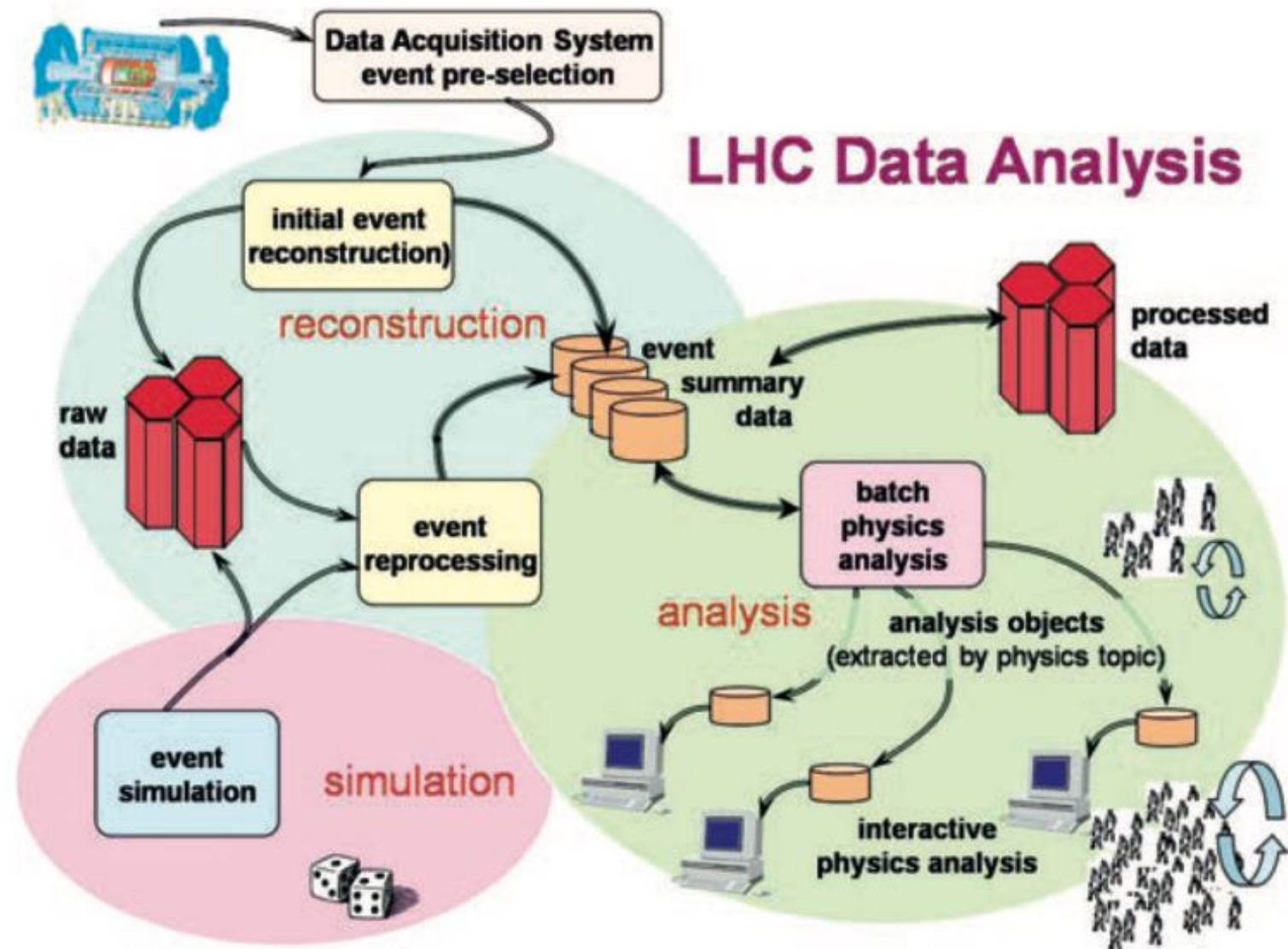
Data at LHC: Raw Data?



Data at LHC: Raw Data?

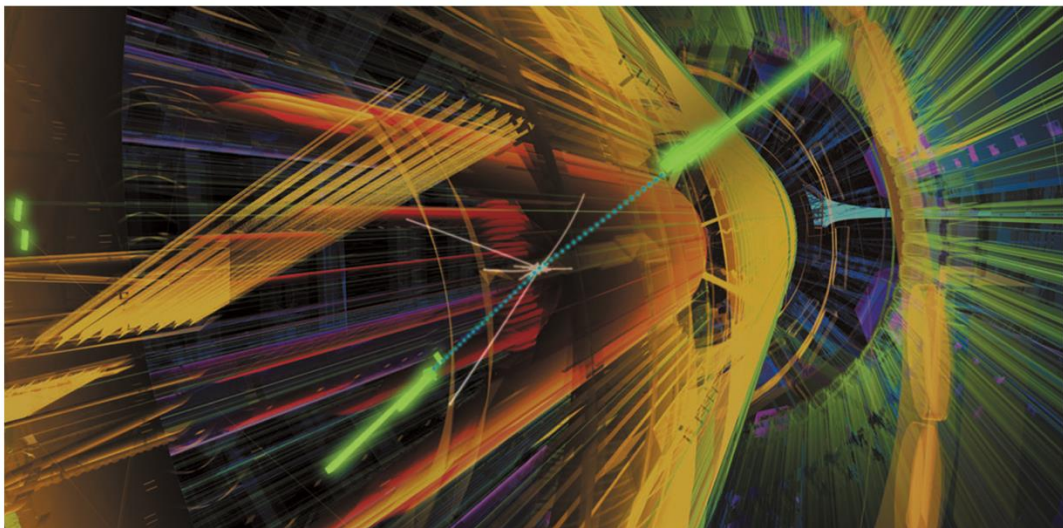
Data at LHC: Raw Data?

64. LHC Data Analysis – high level diagram of the data flow and the major processing stages.



Data at LHC: Raw Data?

NEWS IN FOCUS



Particle collisions at the Large Hadron Collider produce huge amounts of data, which algorithms are well placed to process.

PARTICLE PHYSICS

Artificial intelligence called in to tackle LHC data deluge

Algorithms could aid discovery at Large Hadron Collider, but raise transparency concerns.

BY DAVIDE CASTELVECCHI, GENEVA, SWITZERLAND

particle-physics lab that hosts the LHC. Computer scientists are responding in

the number of collisions will grow 20-fold, and that the detectors will have to use more

2015

scientific data

www.nature.com/scientificdata

Check for updates

OPEN

DATA DESCRIPTOR

LHC physics dataset for unsupervised New Physics detection at 40 MHz

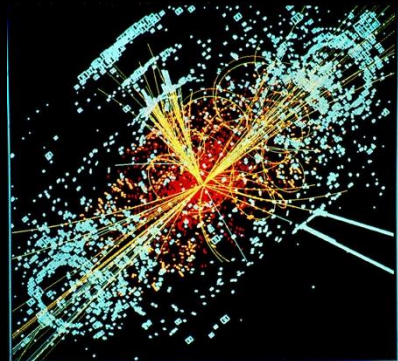
Ekaterina Govorkova¹, Ema Puljak^{1,2}, Thea Aarrestad¹, Maurizio Pierini¹, Kinga Anna Woźniak^{1,3} & Jennifer Ngadiuba^{2,4}

In the particle detectors at the Large Hadron Collider, hundreds of millions of proton-proton collisions are produced every second. If one could store the whole data stream produced in these collisions, tens of terabytes of data would be written to disk every second. The general-purpose experiments ATLAS and CMS reduce this overwhelming data volume to a sustainable level, by deciding in real-time whether each collision event should be kept for further analysis or be discarded. We introduce a dataset of proton collision events that emulates a typical data stream collected by such a real-time processing system, pre-filtered by requiring the presence of at least one electron or muon. This dataset could be used to develop novel event selection strategies and assess their sensitivity to new phenomena. In particular, we intend to stimulate a community-based effort towards the design of novel algorithms for performing unsupervised new physics detection, customized to fit the bandwidth, latency and computational resource constraints of the real-time event selection system of a typical particle detector.

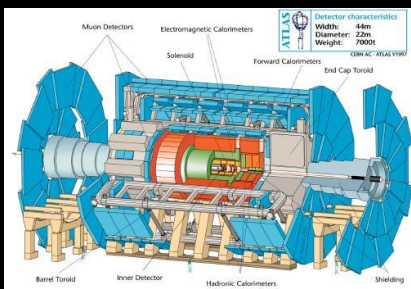
2022

Data at LHC: Raw Data?

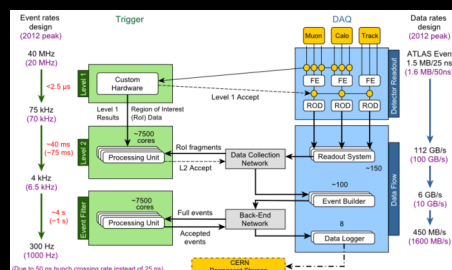
Simulations



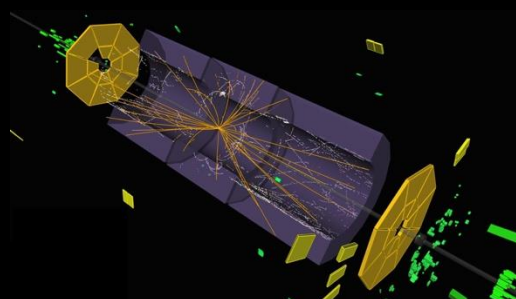
Production



Selection



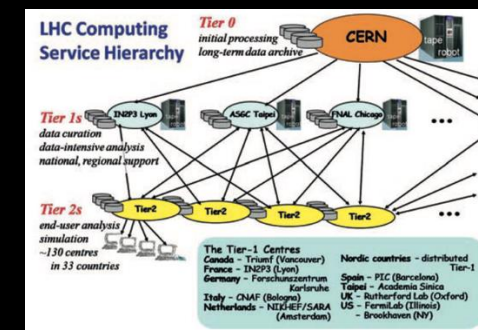
Reconstruction



Storage

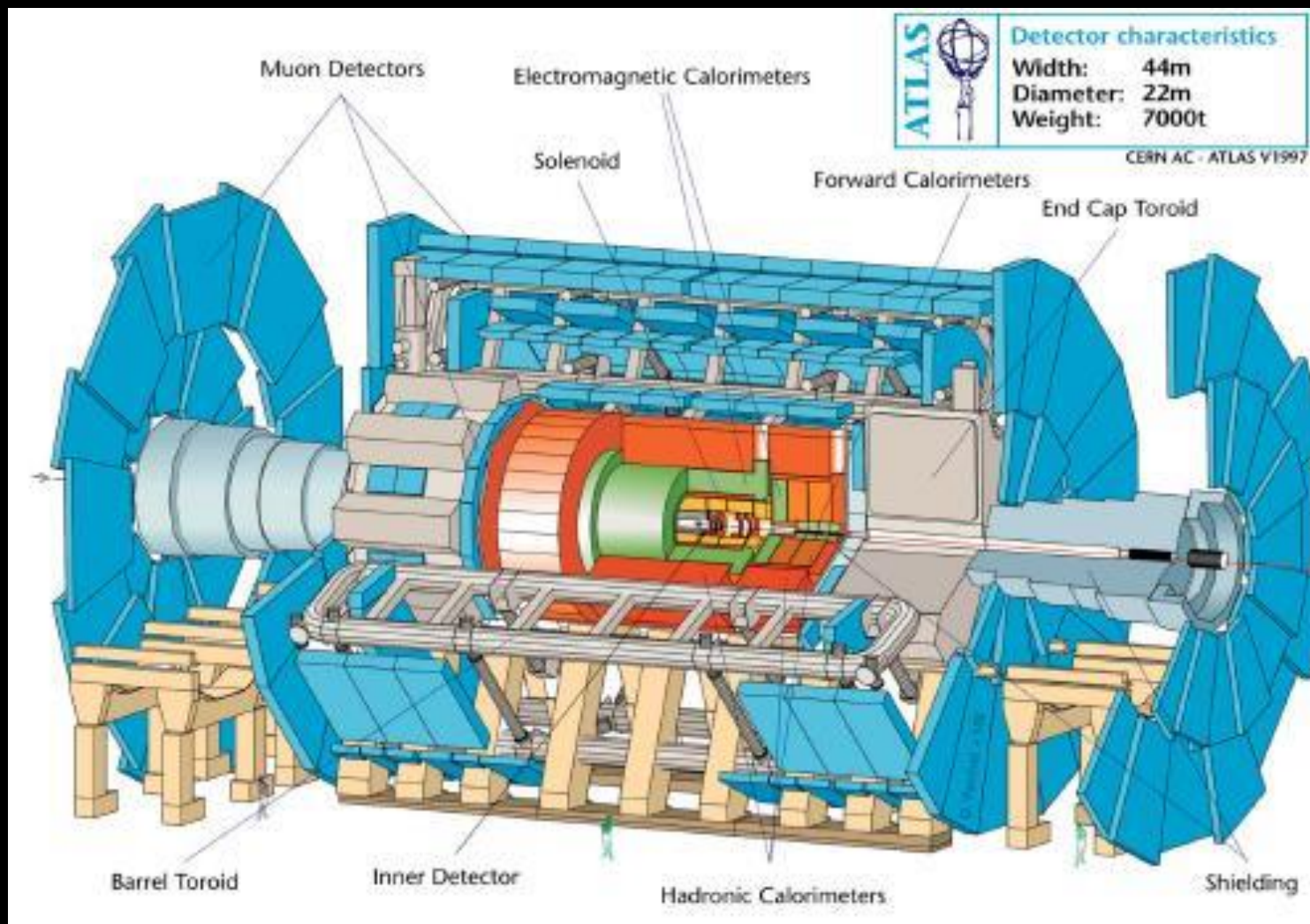


Distributed Analysis



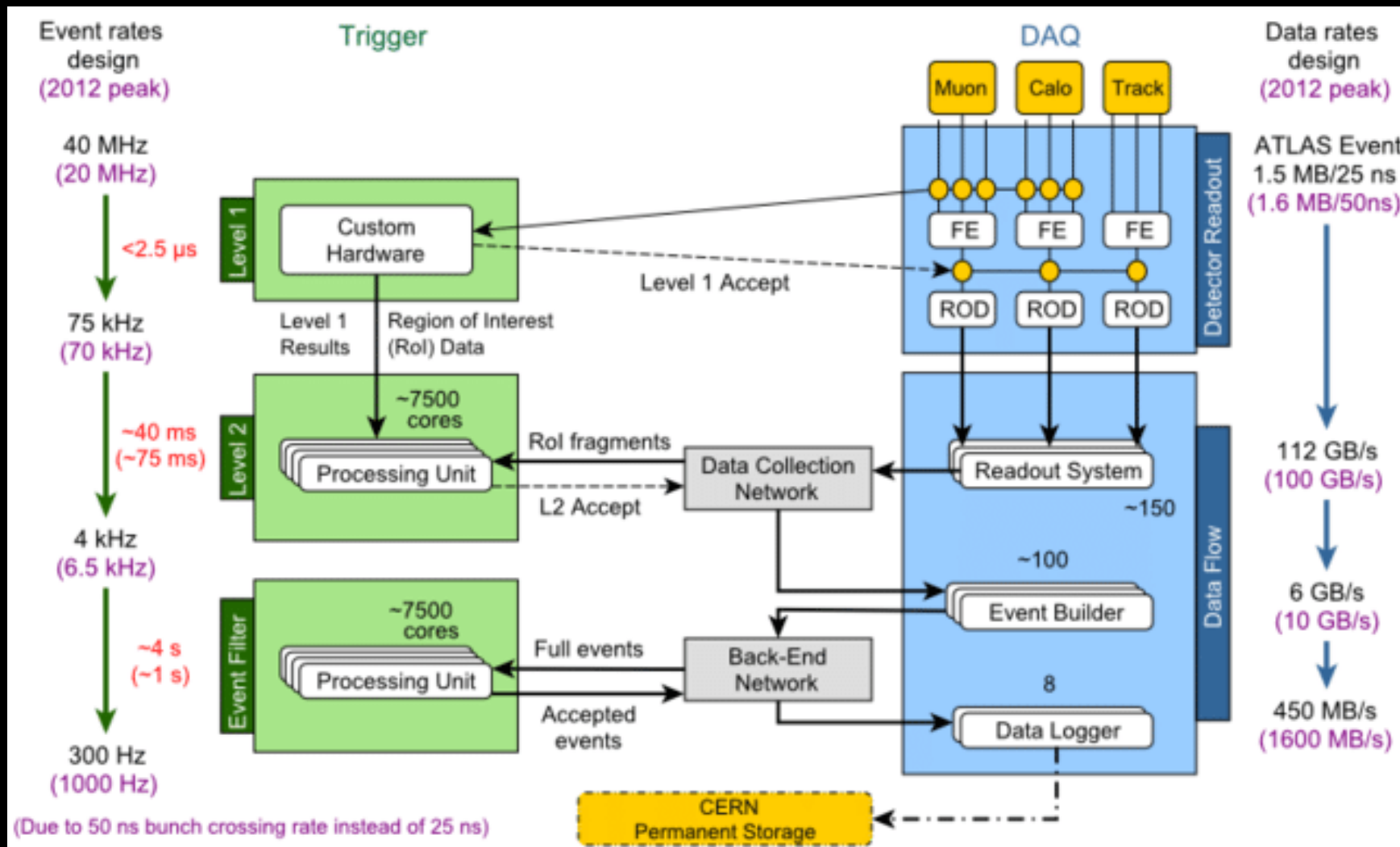
Data at LHC: Raw Data?

Data Production



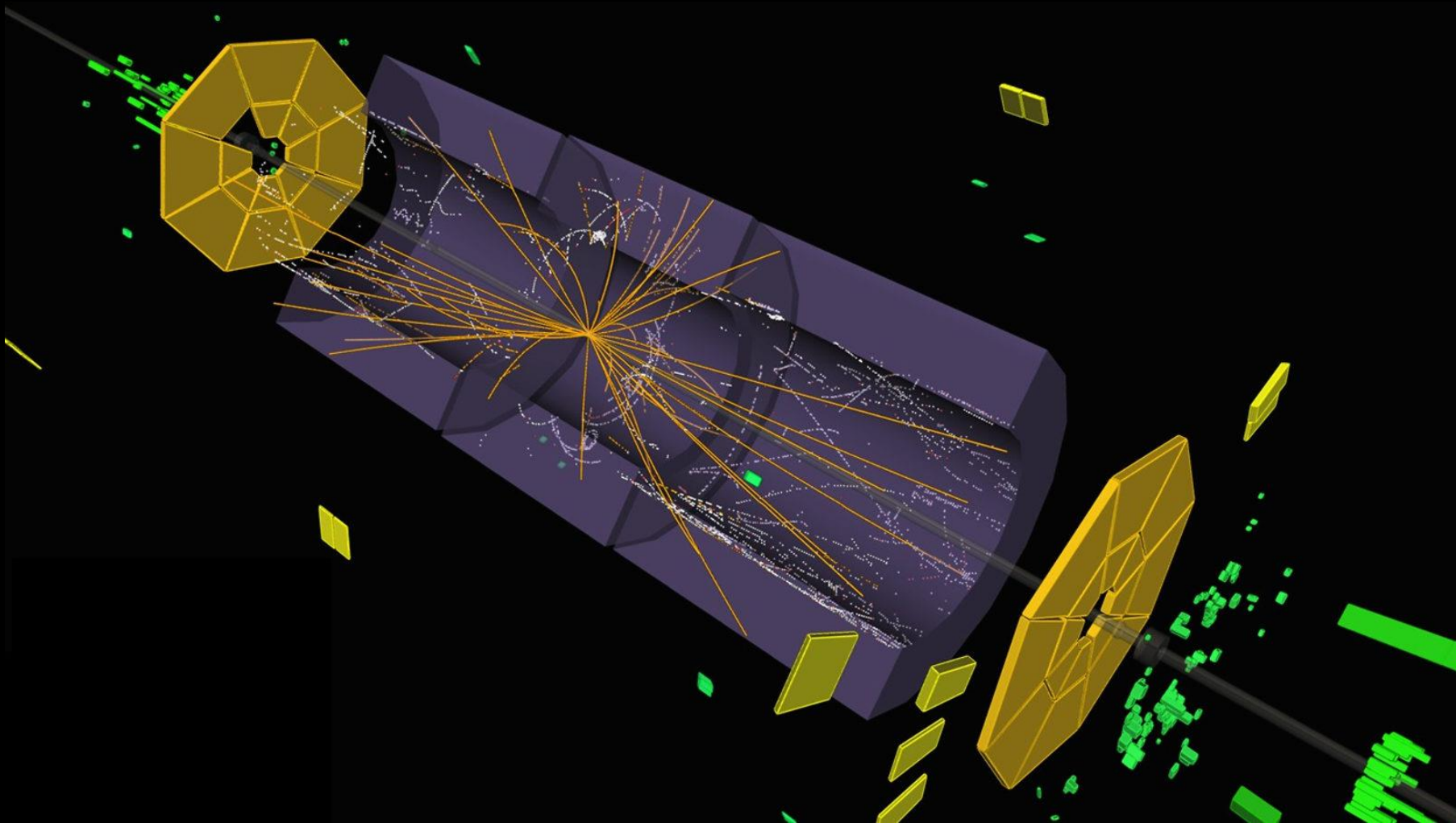
Data at LHC: Raw Data?

Data Selection



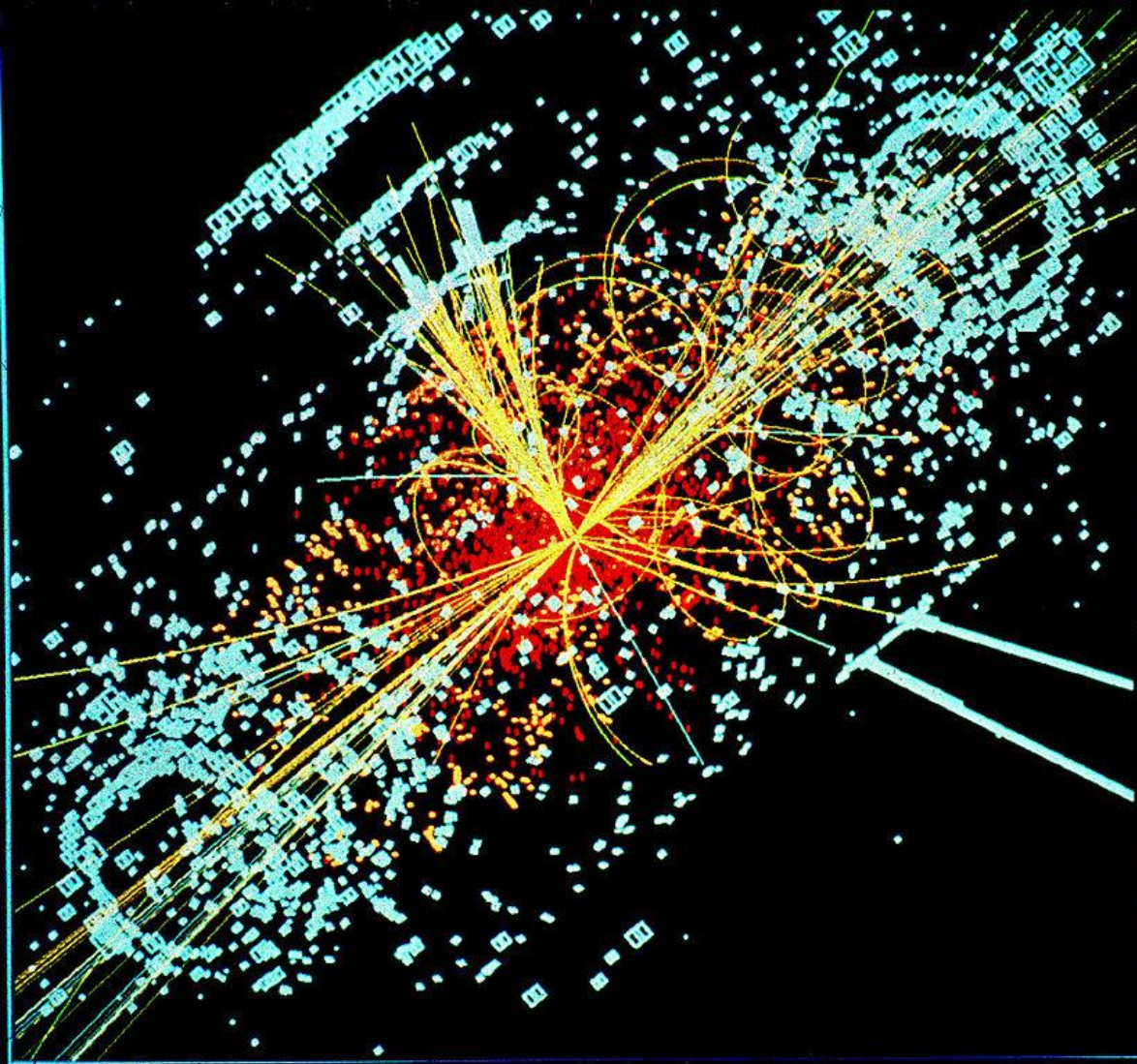
Data at LHC: Raw Data?

Data Reconstruction



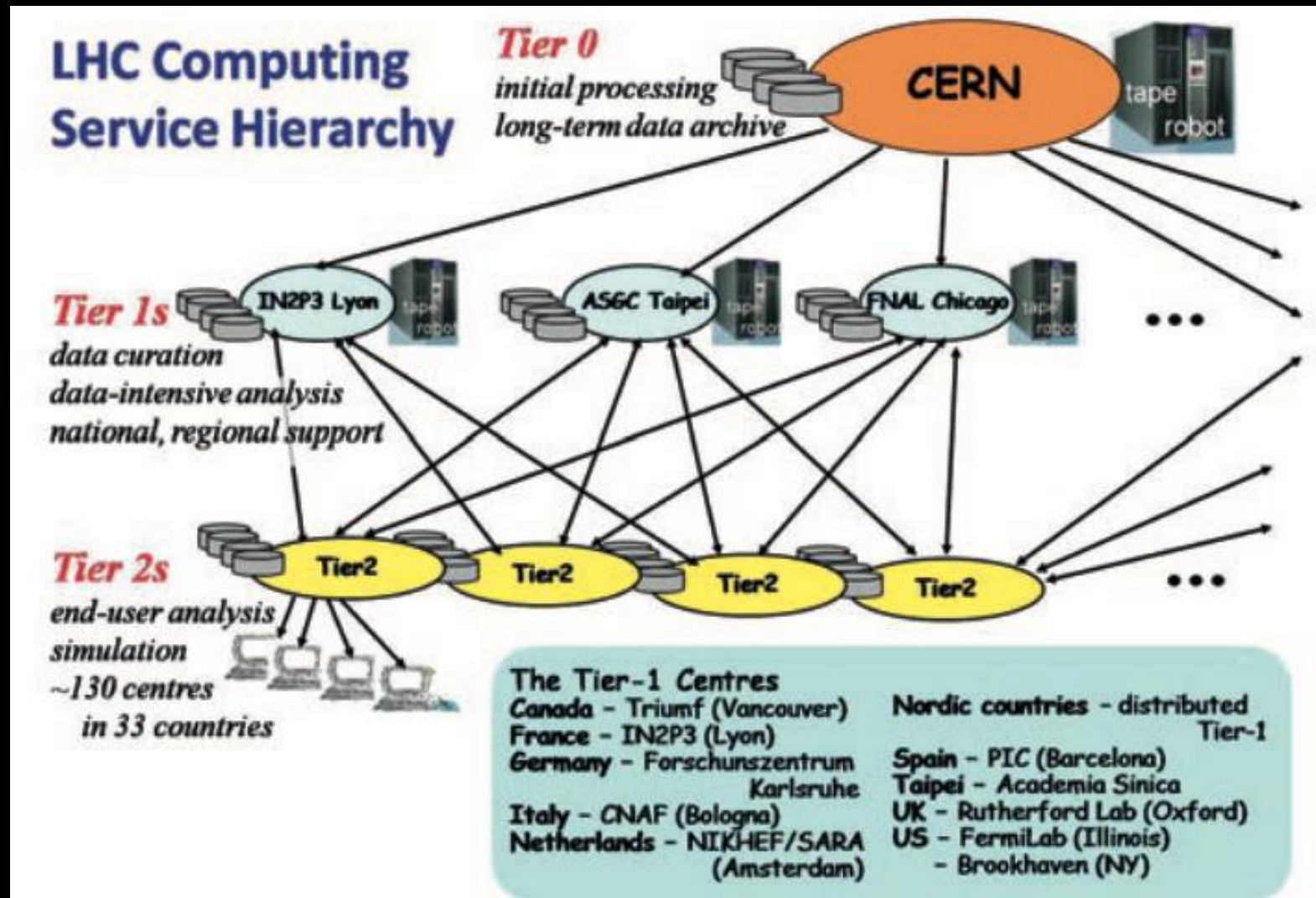
Data at LHC: Raw Data?

Data Simulation



Data at LHC: Raw Data?

Distributed Data Analysis



Data at LHC: Raw Data?

Distributed Data Analysis

open**data**
CERN

Help ▾ About ▾

LHCb 2012 Beam4000GeV MagUp CHARMCOMPLETEEVENT Stream Stripping21r0p2

LHCb collaboration

Cite as: LHCb collaboration (2023). LHCb 2012 Beam4000GeV MagUp CHARMCOMPLETEEVENT Stream Stripping21r0p2. CERN Open Data Portal.
DOI:[10.7483/OPENDATA.LHCb.GLGS.QPET](https://doi.org/10.7483/OPENDATA.LHCb.GLGS.QPET)

Dataset

Collision

LHCb

8TeV

pp

CERN-LHC

Description

Data from proton-proton (pp) collisions collected by the LHCb experiment in the year 2012 of Run1 of the LHC.

Dataset characteristics

53573595 events. **4389** files. **5.5 TiB** in total.

How were these data selected?

This dataset was created in several production steps. These steps, software used and the configuration is provided below.

Prod ID:

94006

Prod type:

Merge

Parent Prod ID:

94005

Parent Prod type:

DataStripping

Conditions:

Data at LHC: Raw Data?

Data Storage



Conclusion

Data Are Not New

Data Are Constructed

Data Are Shaped

Data Are Partial



Thank you!