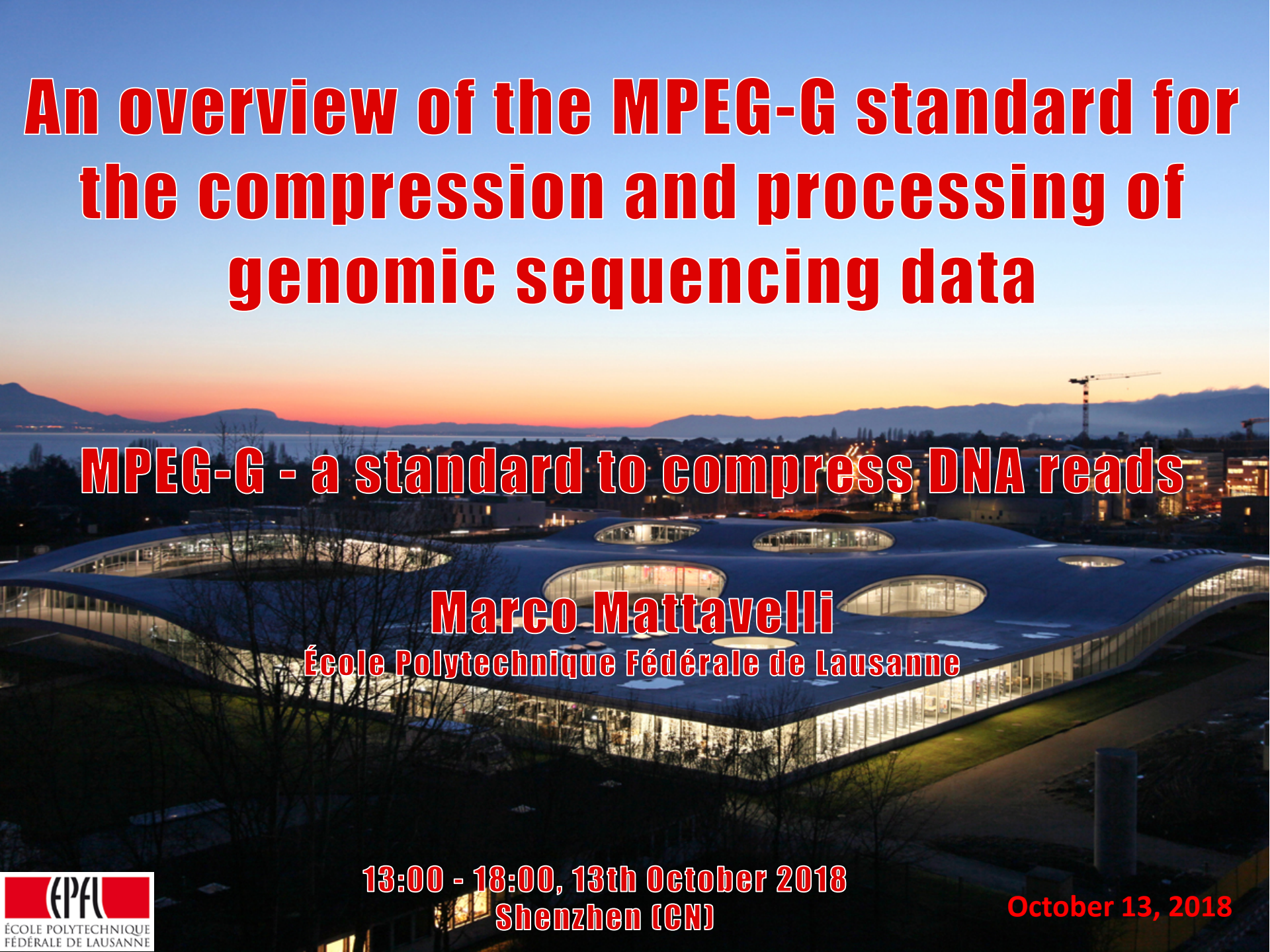
An overview of the MPEG-G standard for the compression and processing of genomic sequencing data

MPEG-G - a standard to compress DNA reads



Marco Mattavelli
École Polytechnique Fédérale de Lausanne

13:00 - 18:00, 13th October 2018
Shenzhen (CN)

October 13, 2018

EPFL
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Can we reduce the ICT costs of genomics?

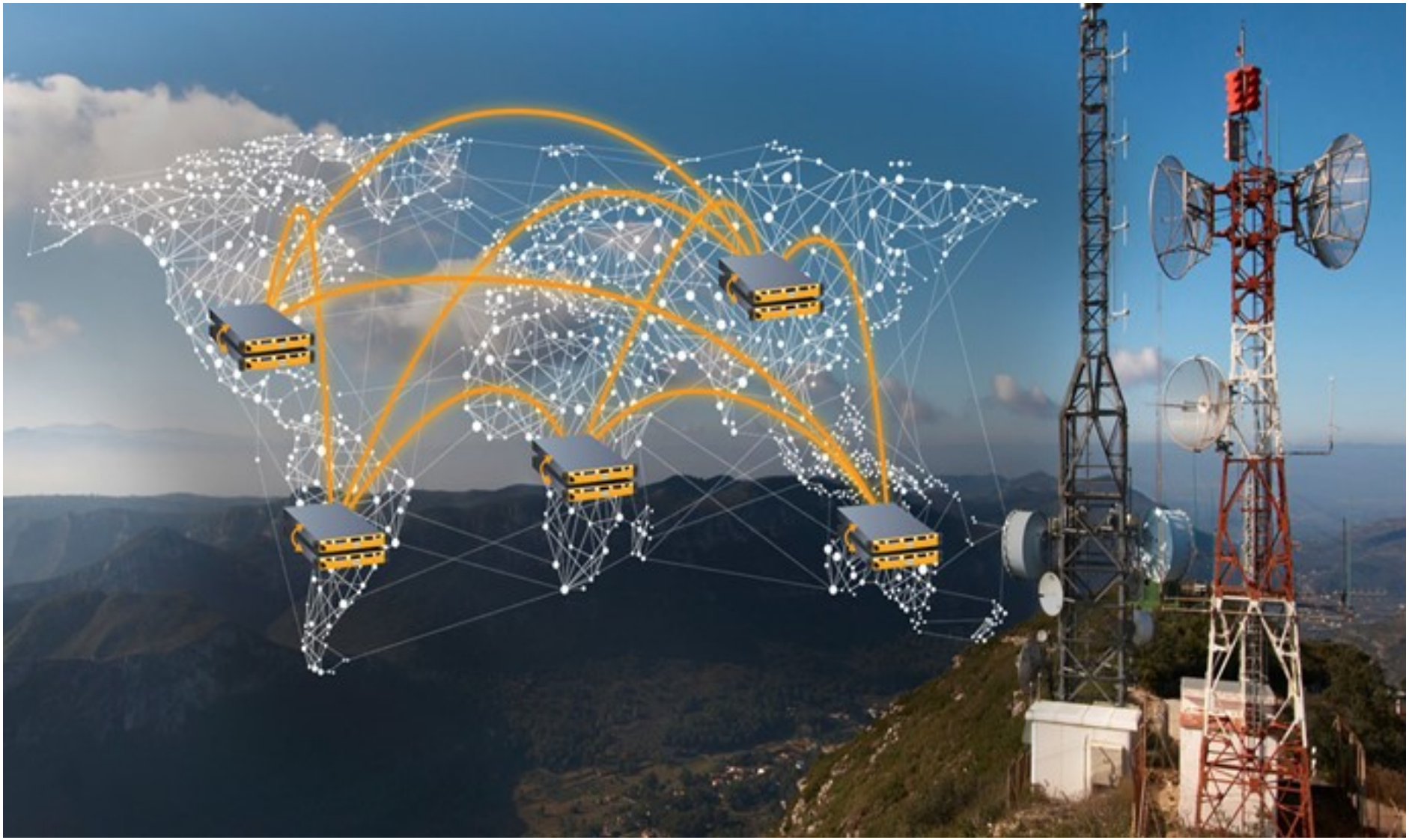
- A problem already seen in the recent past …..
- In the 1980s
  - Digital television handled ~200 Mbit/s (TV) and ~1 Gbit/s (HDTV)
  - Many proprietary compression format stifled the market
- In the early 1990s **MPEG developed MPEG-2**
  - Compression of ~100x and other functionalities (random access,...)
  - VLSI chips available from multiple sources
  - Developed, maintained and created a string of compression standards:
    - MPEG-2: ~50-100 x; MPEG-4: ~200 x; MPEG-H : ~400 x, MPEG-I: ~800 x
- On the path of what done in digital media **MPEG and ISO/TC276 are developing MPEG-G**,
  - Digital representation (including compression) of sequenced DNA
  - To be approved as Draft International Standard in January 2019

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

The lesson of 30 years of MPEG Digital Media

All digital media content is compressed …..

…..today transmitted in compressed form everywhere …



Lesson from these 30 years:

- Compression is important, technology enabler, but it is not all: **MPEG «Systems APIs»** are even more important.
- Digital media applications are built «around» the MPEG standard «**Systems APIs**» :
  - All component are «synchronized» and linked
  - Access to data in the native compressed domain
- If a compression (standard) technology evolves the «**Systems**» standard **remains valid**!!

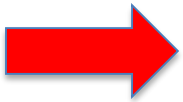
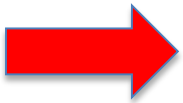
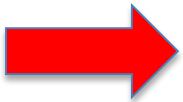
Can current genomic formats do the job?

- Current data storage and processing are based on an ASCII «file format» representation (a «huge» matrix of redundant information fields): SAM

- Genomic data compression is simply a SAM file zipped line by line: BAM

- Compression and selective data access using SAM and BAM are inefficient

Can current genomic formats do the job?

- **What is wrong with existing formats?**
  - Merging heterogeneous data into a (simplistic!) file format and then compress it
     **inefficient compression**
  - APIs not based on a native compressed format
     **inefficient selective access to data**
  - Missing transport format supporting APIs and selective data access in the compressed domain
     **inefficient access to remote data**

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

The MPEG Approach

The MPEG approach

- Decoder standards
- Normative text
- Reference software – decoder only
  - Reference implementation of standard, alternative implementations possible
  - Open source
- Conformance
  - Bitstreams and associated decoded output
- Encoder software – for information only
  - Commercial implementations improve over time
  - Speed, compression, complexity
- Amendments and Corrigenda

Everybody uses MPEG standards:
MP3, AAC, MPEG-2, MPEG-4, AVC, HEVC, DASH, OpenType
DVB, DVD, Blu-ray, FaceTime, ATSC, Netflix, Windows, macOS, iOS, Android, …

The MPEG approach

- Identification of a topic for standardization
  - Analyze industry needs
  - Seek industry participation
  - Select test data
- Open Call for Proposals – open evaluation
- Standards development
  - Core experiments
- System aspects (storage, transport, privacy) driven by industry
- Evaluation
- Final Standard (text and software)

Everybody uses MPEG standards:
MP3, AAC, MPEG-2, MPEG-4, AVC, HEVC, DASH, OpenType
DVB, DVD, Blu-ray, FaceTime, ATSC, Netflix, Windows, macOS, iOS, Android, …

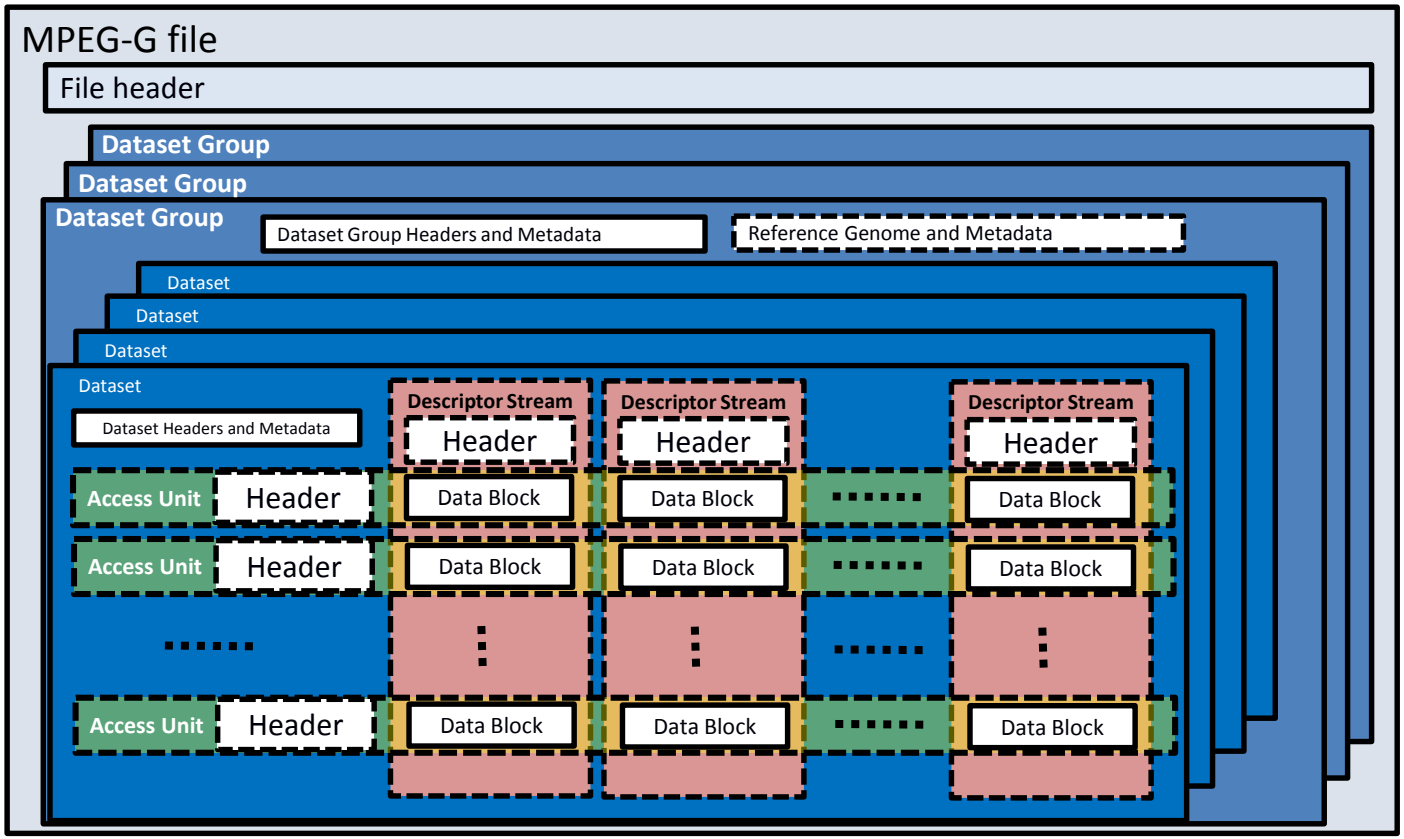
The MPEG-G Parts

- Part 1: **File and Transport Format (DIS Jan 2019)**
  - The technology to transport and access data
- Part 2: **Compression of genomic data (DIS Jan 2019)**
  - The compressed representation
- Part 3: **APIs (DIS Jan 2019)**
  - Standard interfaces with genomic data applications and legacy formats
- Part 4: **Reference Software (DIS July 2019)**
  - The standard support to the implementation of applications
- Part 5: **Conformance (DIS July 2019)**
  - The methodology to test compliance with the standard

The specific MPEG-G approach

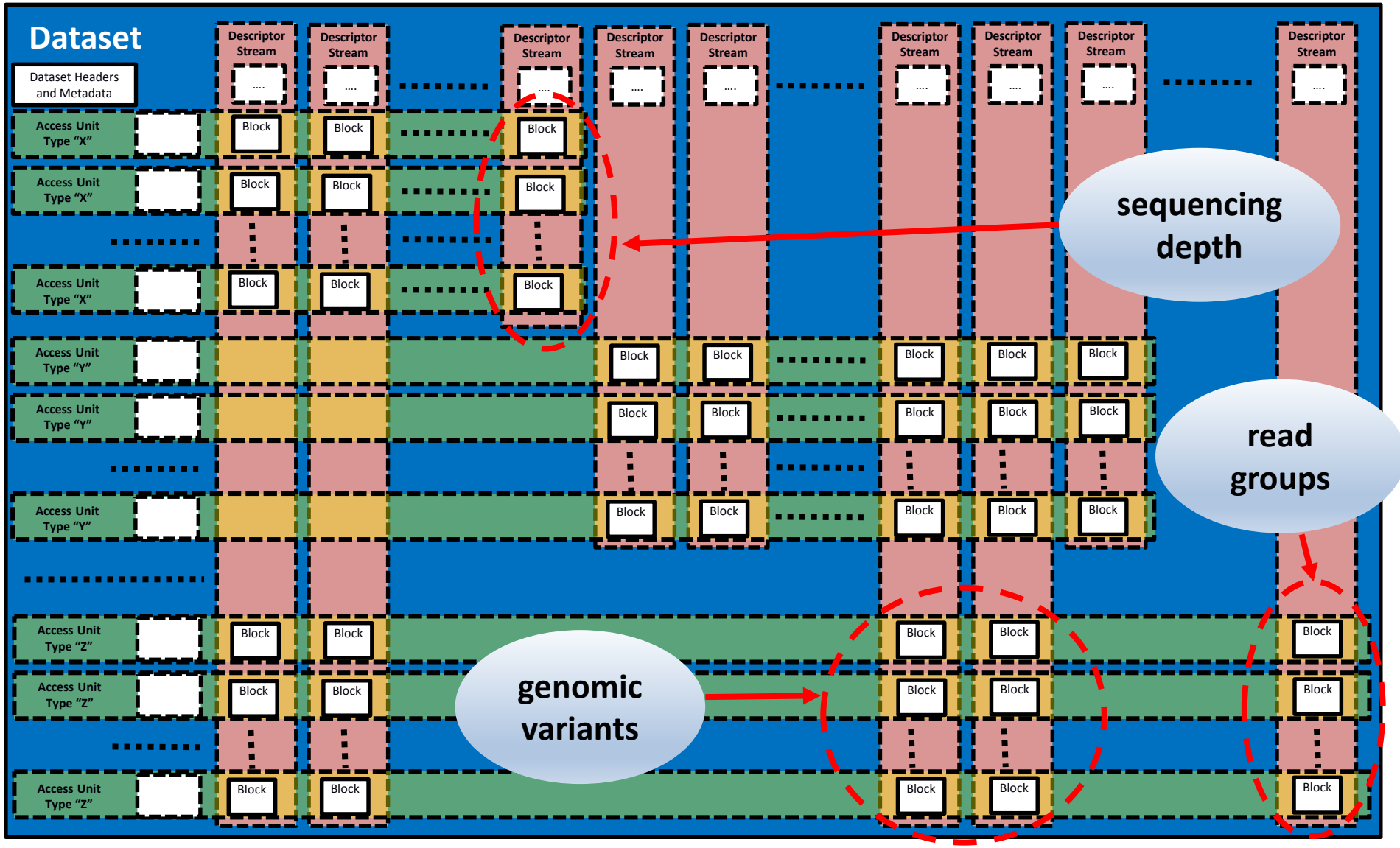
- Genomic data (i.e. sequence reads) are **classified into homogeneous sets** (classess of data) and represented by **minimal sub-sets** of **«descriptors»**
- Meta-data associated to classified reads is represented by **specific «descriptors»**
- Descriptors sub-sets are **compressed individually** and then are stored into structured **«Access Units»**
- **Access Units** are included into an indexed **«File Format»**

The MPEG-G File Format

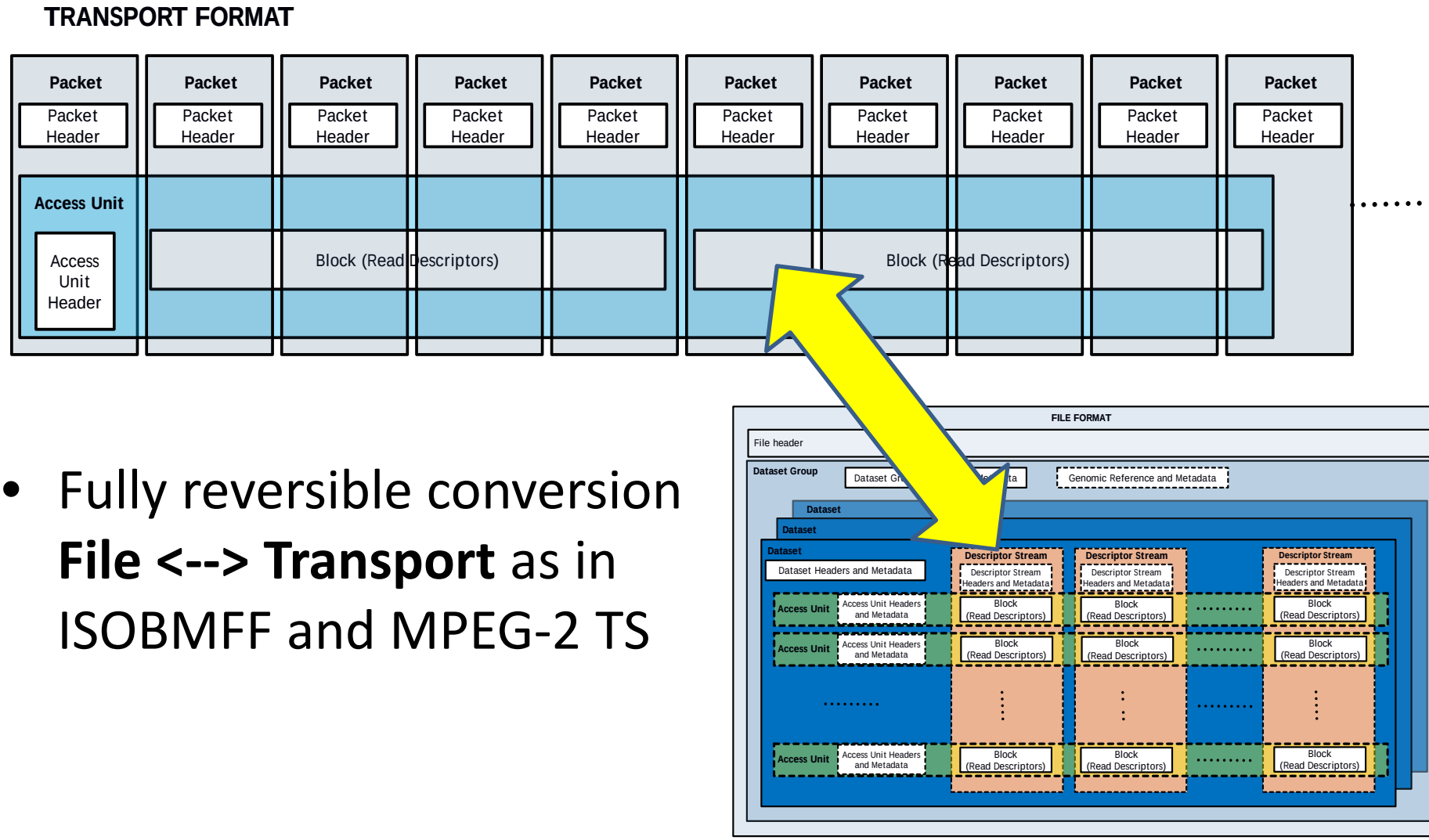


- File
  A trio (Dad, Mum, Child)
- Dataset group
  One individual's history
- Dataset
  One sequencing run

The MPEG-G File Format

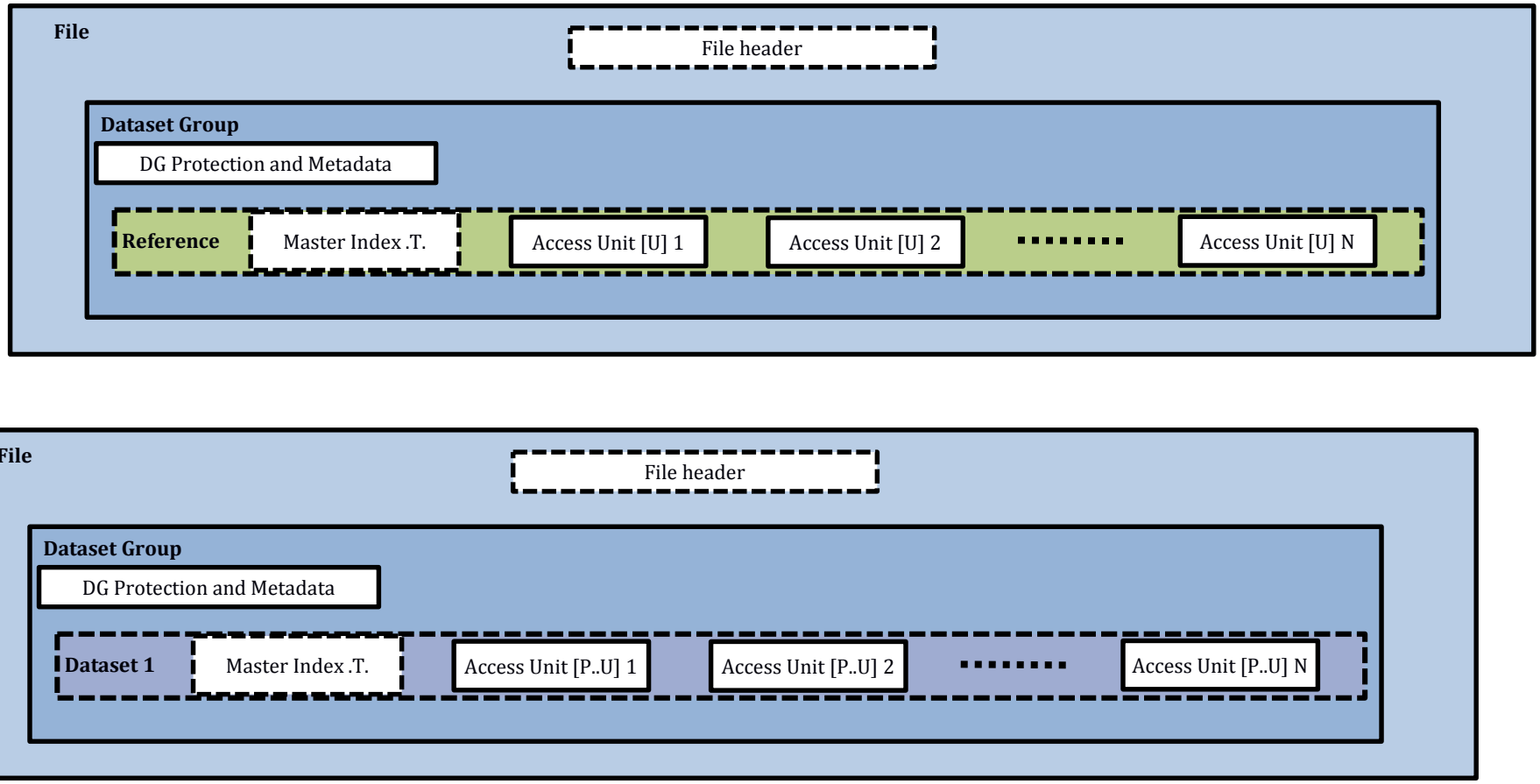


MPEG-G File Format and transport format

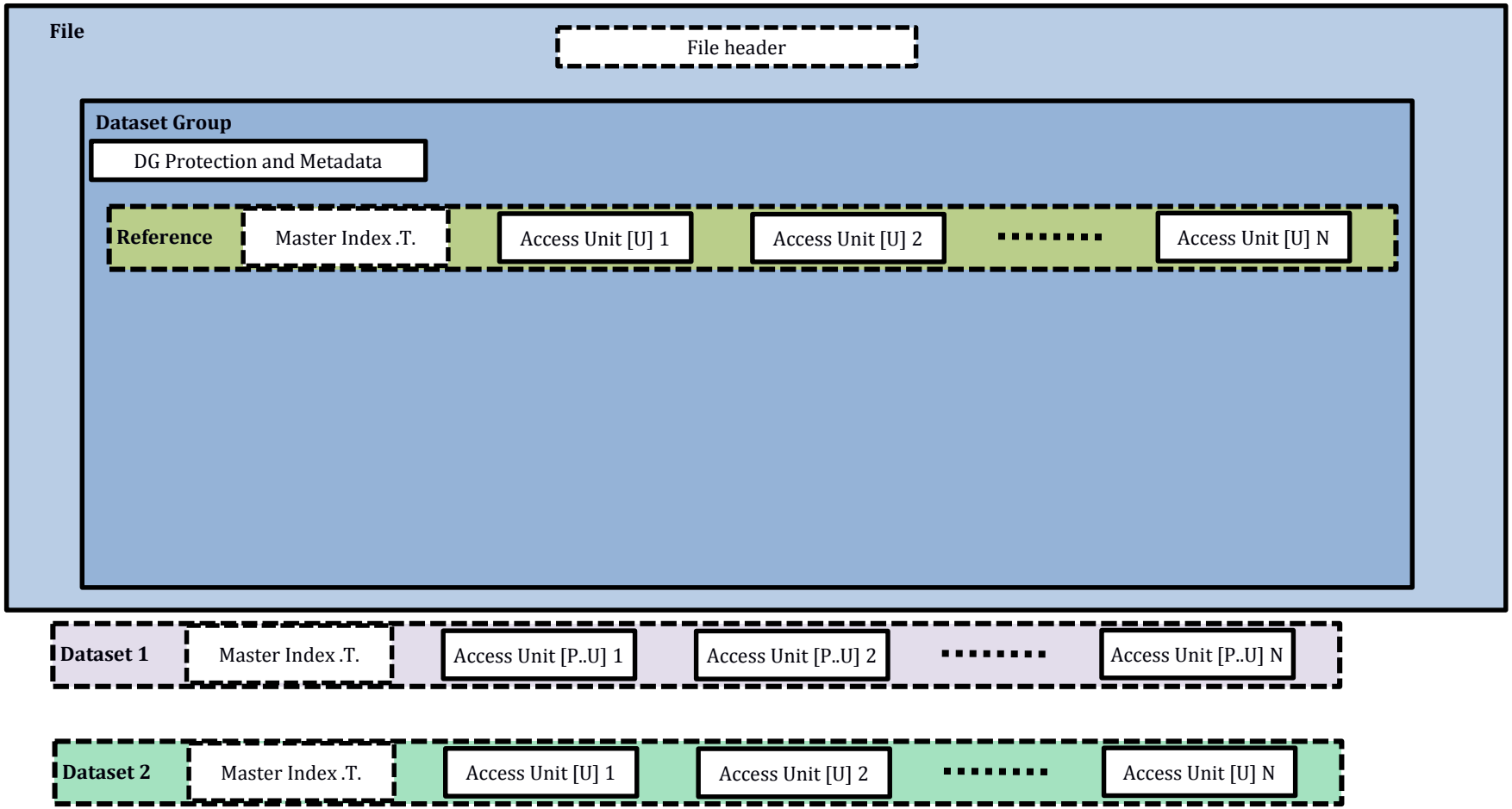


- Fully reversible conversion **File <--> Transport** as in ISOBMFF and MPEG-2 TS

MPEG-G Data Aggregation



MPEG-G Data Aggregation



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

The MPEG-G Compression

The MPEG-G standard compression concepts

- Read Classification
- CABAC Context definition
- QV Quantization Strategies
- Reference Based Read Coding
- Read Names Tokenization
- Access Units Formatting
- Referenceless Read Coding
- Reference Transformation
- Descriptors Binarization
- ……

Coding Optimization Problem

Standard Decoding Syntax

MPEG-G File

Conformant Decoders

MPEG-G Compression: some results for raw data

		MPEG-G mode 1	MPEG-G mode 2	7Zip	Uncompressed
ERR174310_1 (Illumina Platinum Genome)	**Sequences**	5’129’906’462 **(24.22%)**	1'190'561'152 **(5.62%)**	5’149’876’546 **(24.32%)**	21’173’105’634
	QV	6’986’205’872 **(33%)**		7’512’355’308 **(35.48%)**	21’173’105’634
Ultra high coverage. Short reference	**Sequences**	14’203’191’419 **(23.78%)**	336'755'605 **(0.55%)**	2’310’893’902 **(3.79%)**	61’014’790’288
	QV	6’643’303’736 **(10.89%)**		7’948’243’190 **(13.03%)**	61’014’790’288

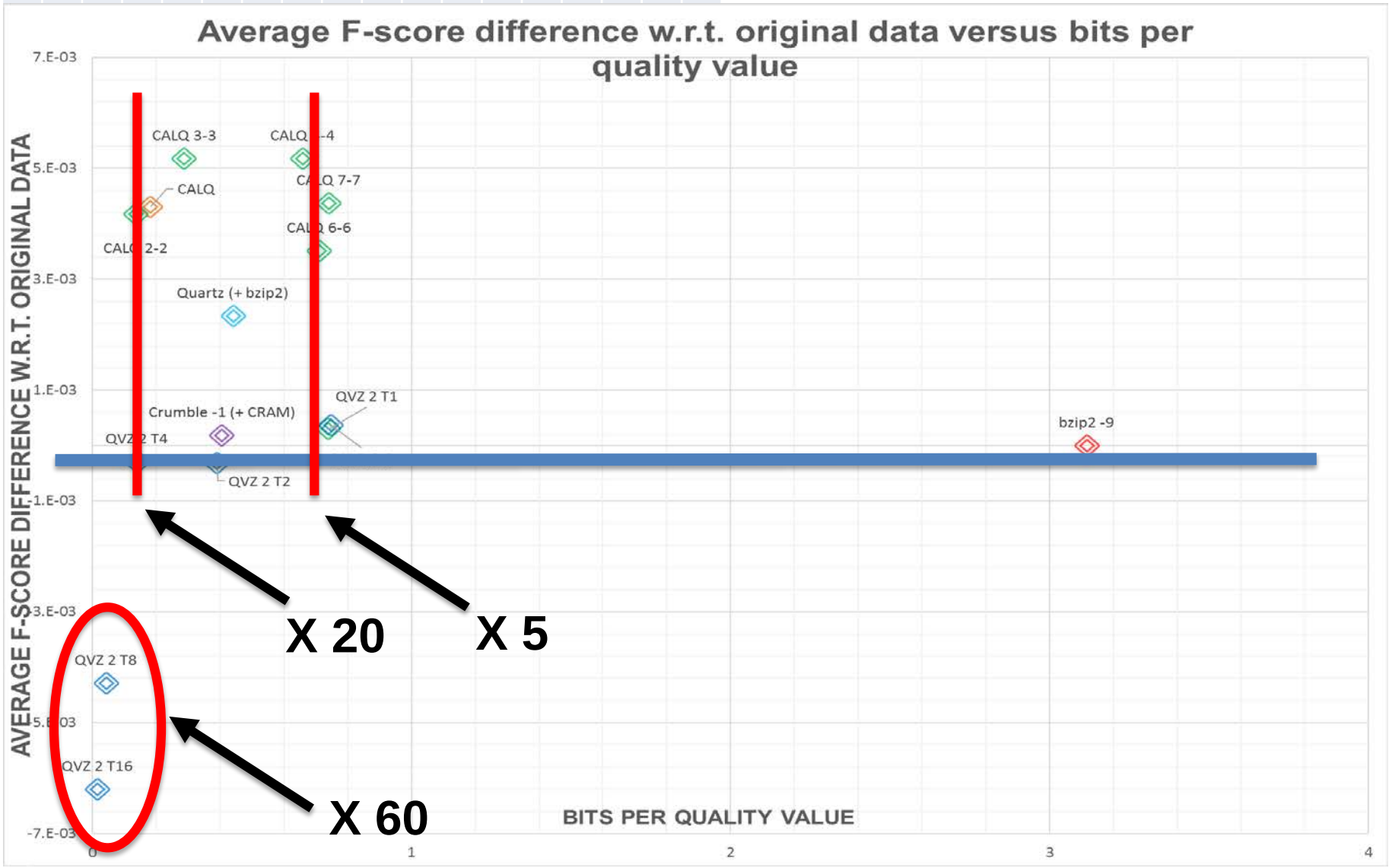
Fast, low memory

Slower, high memory

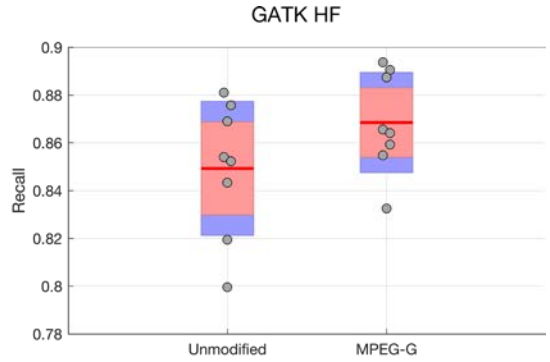
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

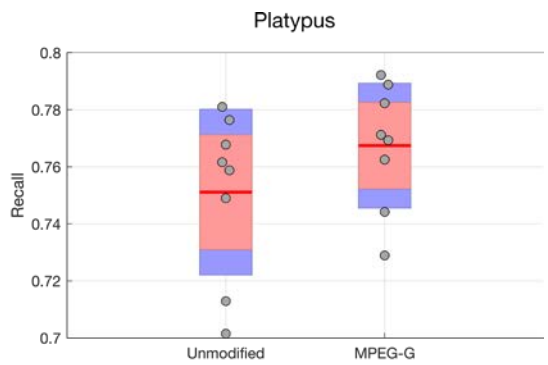
Compressing samples of a noisy process

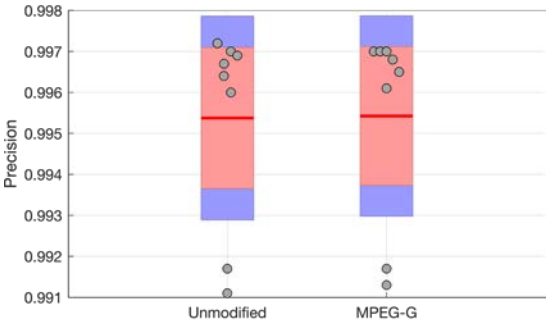
Rate-Distortion for Quantized Quality Values

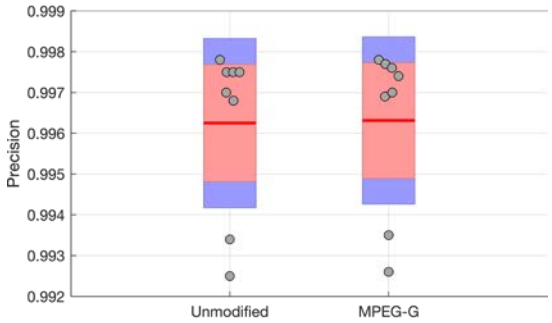


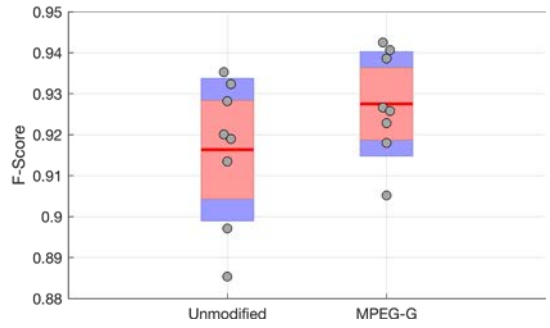
«Functional Equivalence» and QV quantization

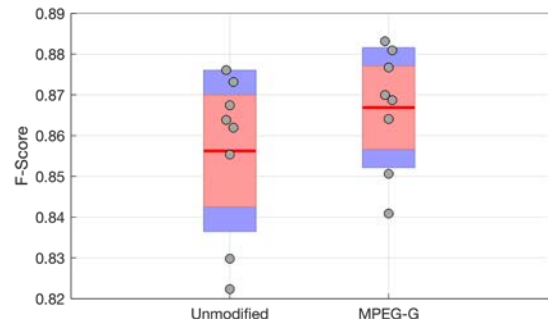




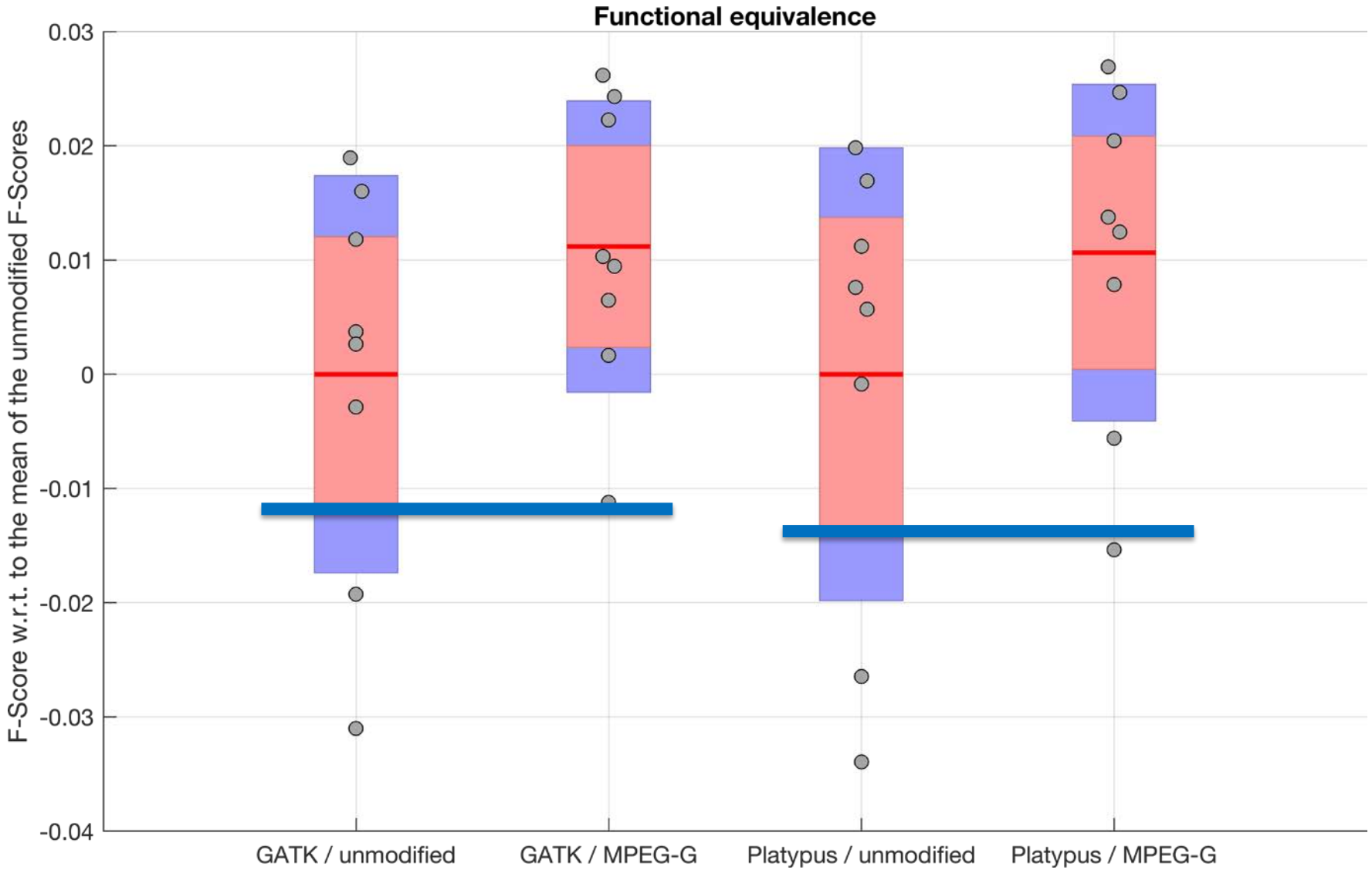








«Functional Equivalence» and QV quantization

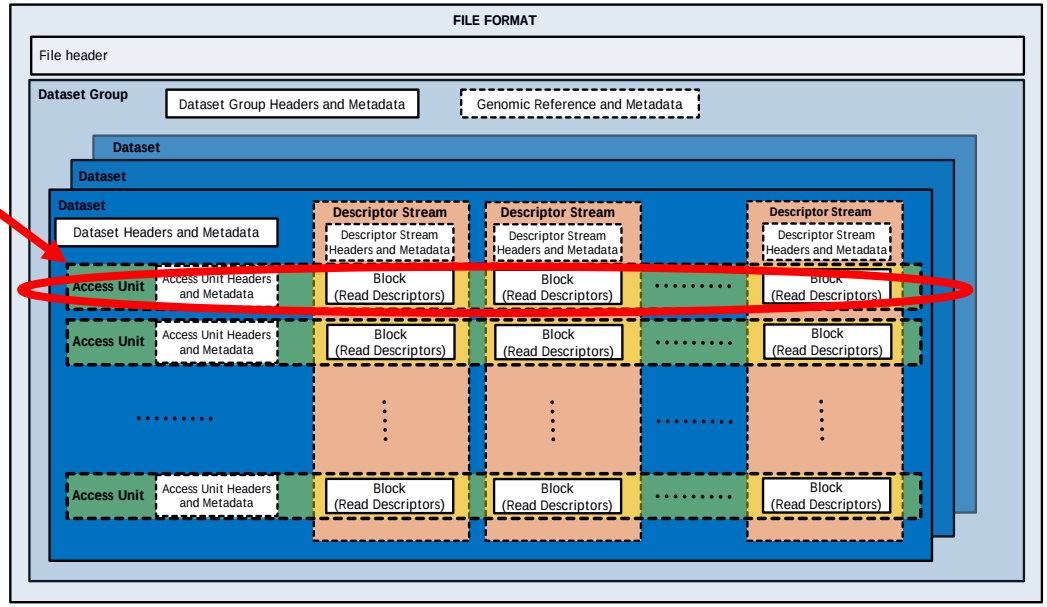


ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

MPEG-G Metatadata APIs and Data Protection

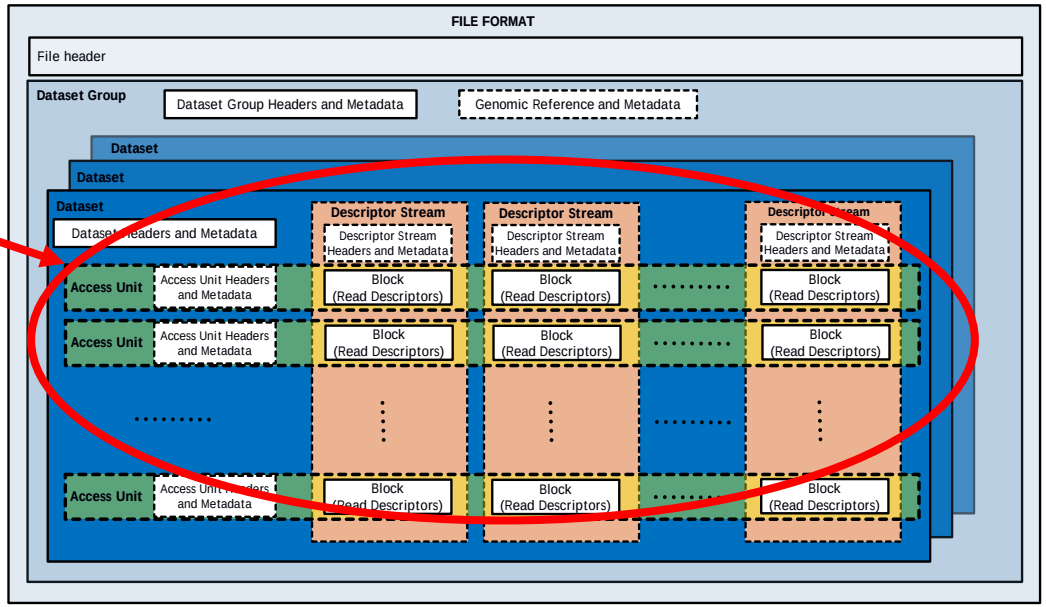
MPEG-G Metatadata APIs and Data Protection

- Any indexed object in a MPEG-G file can be protected
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



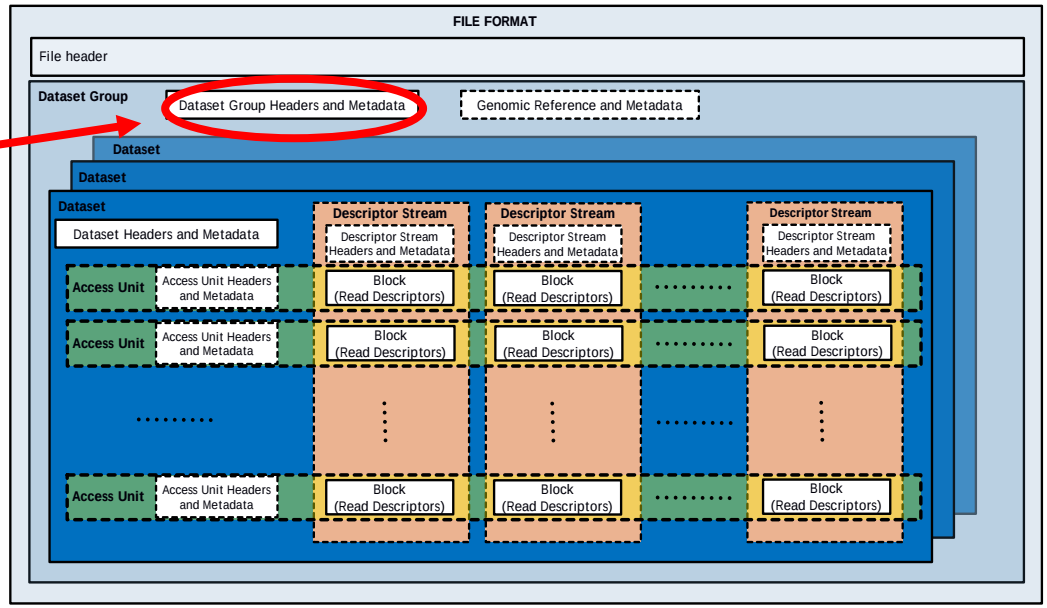
MPEG-G Metatadata APIs and Data Protection

- Any indexed object in a MPEG-G file can be protected
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



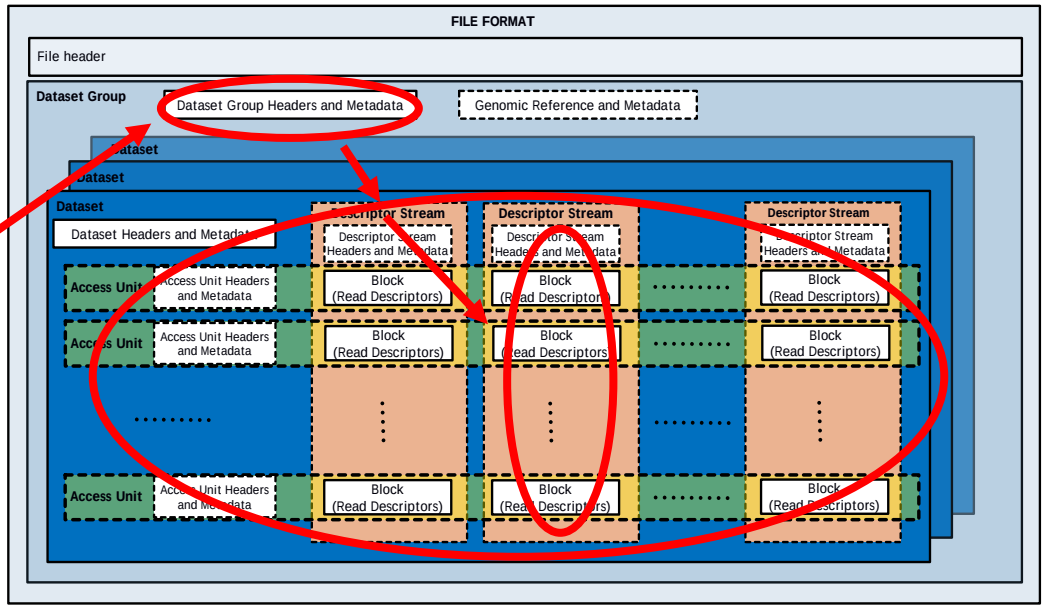
MPEG-G Metatadata APIs and Data Protection

- Any indexed object in a MPEG-G file can be protected
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



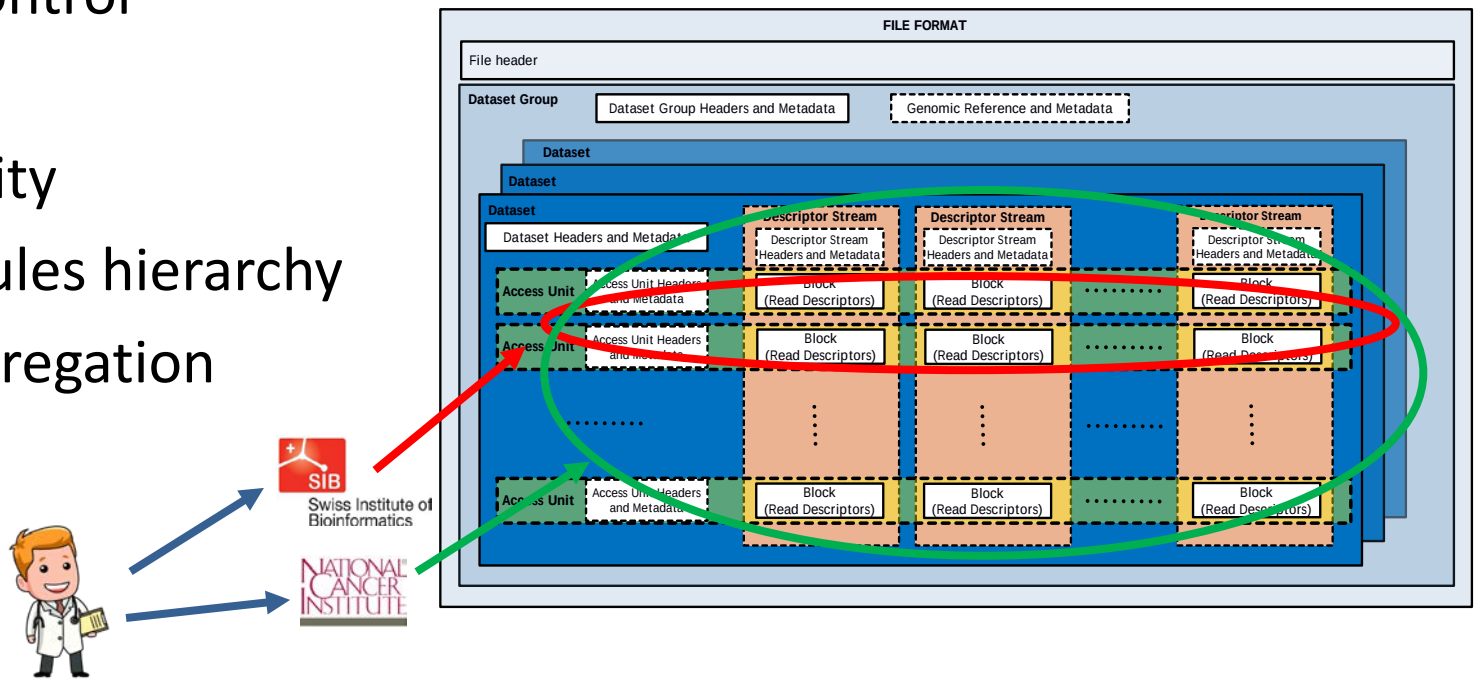
MPEG-G Metatadata APIs and Data Protection

- Any indexed object in a MPEG-G file can be protected
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



MPEG-G Metatadata APIs and Data Protection

- Any indexed object in a MPEG-G file can be protected
- Access control
- Integrity
- Traceability
- Privacy rules hierarchy
- Roles segregation



MPEG-G Metatadata APIs

Dataset group

Element name	Element type	Mandatory
Title	String	Yes
Type	Controlled vocabulary	Yes
Abstract	String	No
Project centre name	ProjectCentre type	No
Description	String	No
Samples	ListOfSamples type	Yes
Extensions	ListOfExtensions type	No

Dataset

Element name	Element type	Description	Mandatory
Title	String		No
Type	Controlled vocabulary		No
Abstract	String		No
Project centres	ListOfProjectCentres type	Contact information of centres participating in the generation of the described study's data.	No
Description	String		No
Samples	ListOfSamples type	Identification of the samples, based on taxonomy/scientific name, common name or anonymized name and further attributes defined in a controlled library.	No
Extensions	ListOfExtensions type		No

MPEG-G Metatadata APIs

- Profiles are defined to support well-known sets
  - e.g. the SRA (Sequence Read Archive) schema for one Run
  - a profile is identified by a URI urn:mpeg:mpeg-g:metadata:profile:ega:run
- A normative mechanism for extensions is specified
- Not bound to one specific ontology

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Conclusions

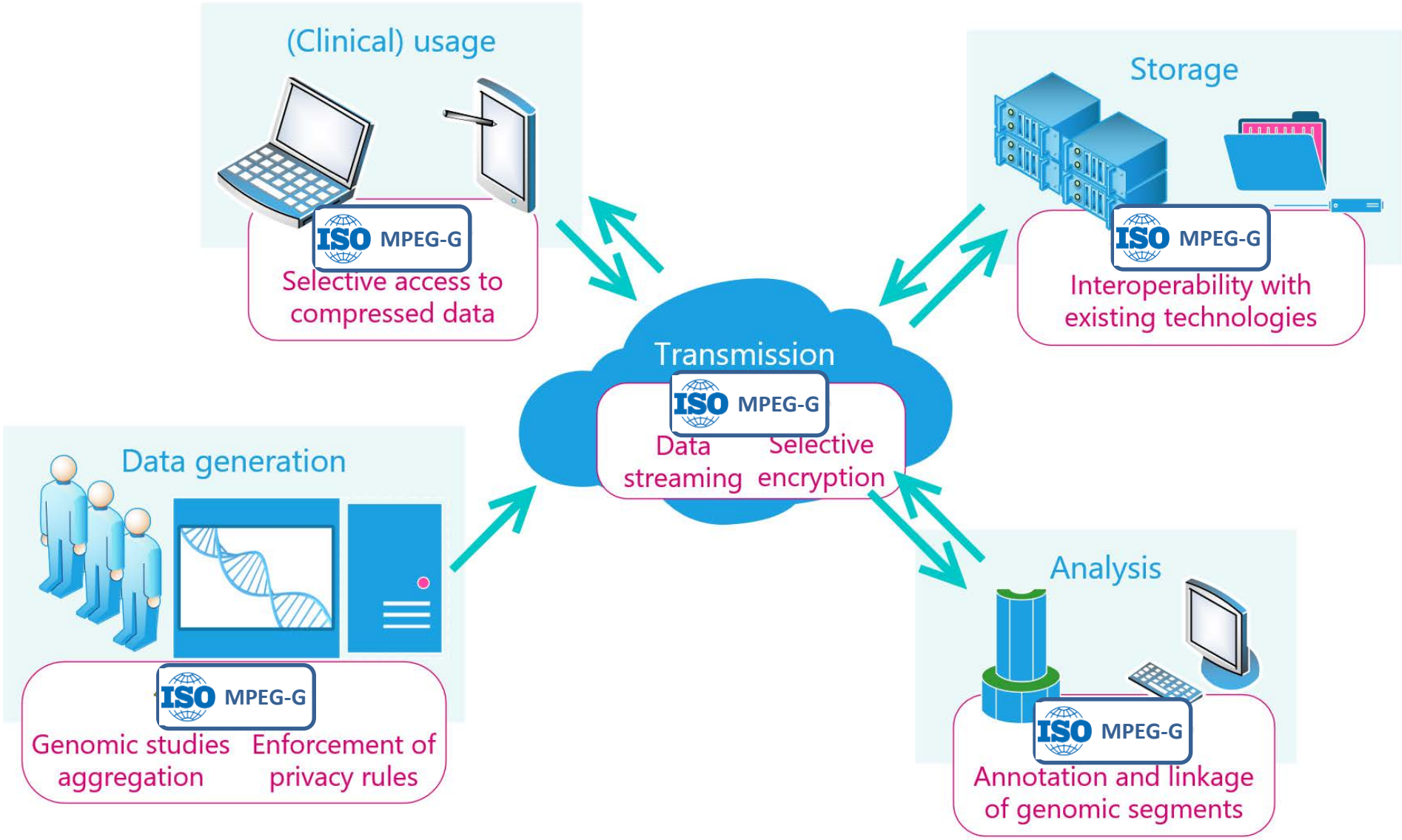
Summary of MPEG-G Features and Functionality

Selective access to compressed data	**Data streaming**	**Genomic studies aggregation**
Flexible access control and privacy protection	**Efficient compression**	**Labelling and association of genomic data in the compressed domain**
Interoperability with main existing technologies and legacy formats	**Incremental update of sequencing data and metadata**	**Custom selected subsets of sequencing data/metadata**

MPEG-G Technology and Ecosystem Evolution

- The MPEG-G compression will improve implementing more advanced encoding strategies
- Competition on MPEG-G encoder implementations
- Ecosystem of applications exchanging compliant bitstreams through standard interfaces

MPEG-G a standard for a complete ecosystem support



Conclusions

- MPEG-G is based on the experience of more than 30 years of Digital Media
- Expectations for genomic analysis applications
  - Sequencing data compression, transport and APIs will **improve and evolve in time**
  - Main APIs and transport **functionality will remain valid**

Many thanks to ….!

Collaborative and competitive efforts of many companies and individuals!!

- **Barcelona Supercomputing Centre (ES), Centre Nacional de Analisis Genómica (ES), Centre for Genomic Regulation (ES), DAPCOM (ES), EPFL (CH), GenomSys (CH), Hannover University (DE), Heidelberg Institute for Theoretical Studies (DE), IMEC (BE), Made of Genes (ES), Pirbright Institute (UK), Swiss Institute for Bioinformatics (CH), Silesian University of Technology (PL), Simon Fraser University (CA), Massachussets Institute of Technology (US), Stanford University (US), Univ. Politecnica de Catalunya (ES), Wellcome Trust Sanger Institute (UK), GenomSoft (CH), Istituto Europeo di Oncologia (IT), CEDEO (IT), AGINOME Scientific (CN) .....**

- **Martin Golebiewski, Yong Zhang, Jan Voges, Ioannis Xenarios, Tom Paridaens, Claudio Alberti, Filippo Medri, Joern Ostermann, Leonardo Chiariglione, Daniel Naro, Jaime Delgado, Giorgio Zoia, Daniele Renzi, Mikel Hernaez, Junaid Ahmad, Paolo Ribeca, Ibrahim Numancig, James Bonfield, Nicolas Guex, Christian Iseli, Thierry Schuepbach, Silvia Llorente, Josep Lluís Gelpí, Dmitry Repchevsky, Romina Royo, Leonor Frías, Oscar Flores, Glenn Van Wallendael, Wesley De Neve, Peter Lambert, Lukasz Roguski, Jordi Portell, Idoia Ochoa, Reggy Long, Noah Daniels, Cenk Sahinalp, Massimo Ravasi, Wenxiang Yang, Rongshan Yu, ... and many others**

Thanks for your attention!!