The Big Data Challenge of WST in the 2040's

An (extragalactic) astronomer's perspective

Stefano Zibetti

INAF
ISTITUTO NAZIONALE DI ASTROFISICA
OSSERVATORIO ASTROFISICO DI ARCETRI

The long road to big-data science in astronomy

What is “big data”?

- “Big” is what we **cannot inspect and control step by step**
  - Most of astronomical research pre-2000’s and a lot of current research not big data in this sense (students and postdocs have been “used” to expand to large datasets, but same philosophy)
- The advent of the **SDSS** in 2001 marked a real change of paradigm:
  - ***Statistics*** instead of individual observations and measurements
  - Data reduction and analysis procedures working as a closed box ***machinery***
- Data volume / data flow not increased by much in the last 20 yrs for ***ground-based optical spectroscopy***:
  - VIMOS surveys well below the SDSS volume
  - 4MOST and WEAVE are within a factor of a few relative to the SDSS in data flow (multiplexing from 640 to ~1k), same for MOONS
  - IFS marked a leap in data complexity, but not much in data volume (see CALIFA, SDSS MaNGA etc).
    MUSE scales up the data volume per observation by a factor ~100, but it’s not a survey facility

SDSS DR7 (anno 2008)

Class	N(total)	N(main)	N(SEGUE)
All	1,640,960	1,374,080	266,880
Galaxies	929,555	928,567	988
Quasars (z <2.3)	104,740	103,121	1,619
Quasars (z ≥2.3)	16,633	15,411	1,222
M stars and later	84,047	76,125	7,922
Other stars	380,214	150,748	229,466
Sky spectra	97,398	75,209	22,189
Unknown	28,383	24,767	3,616

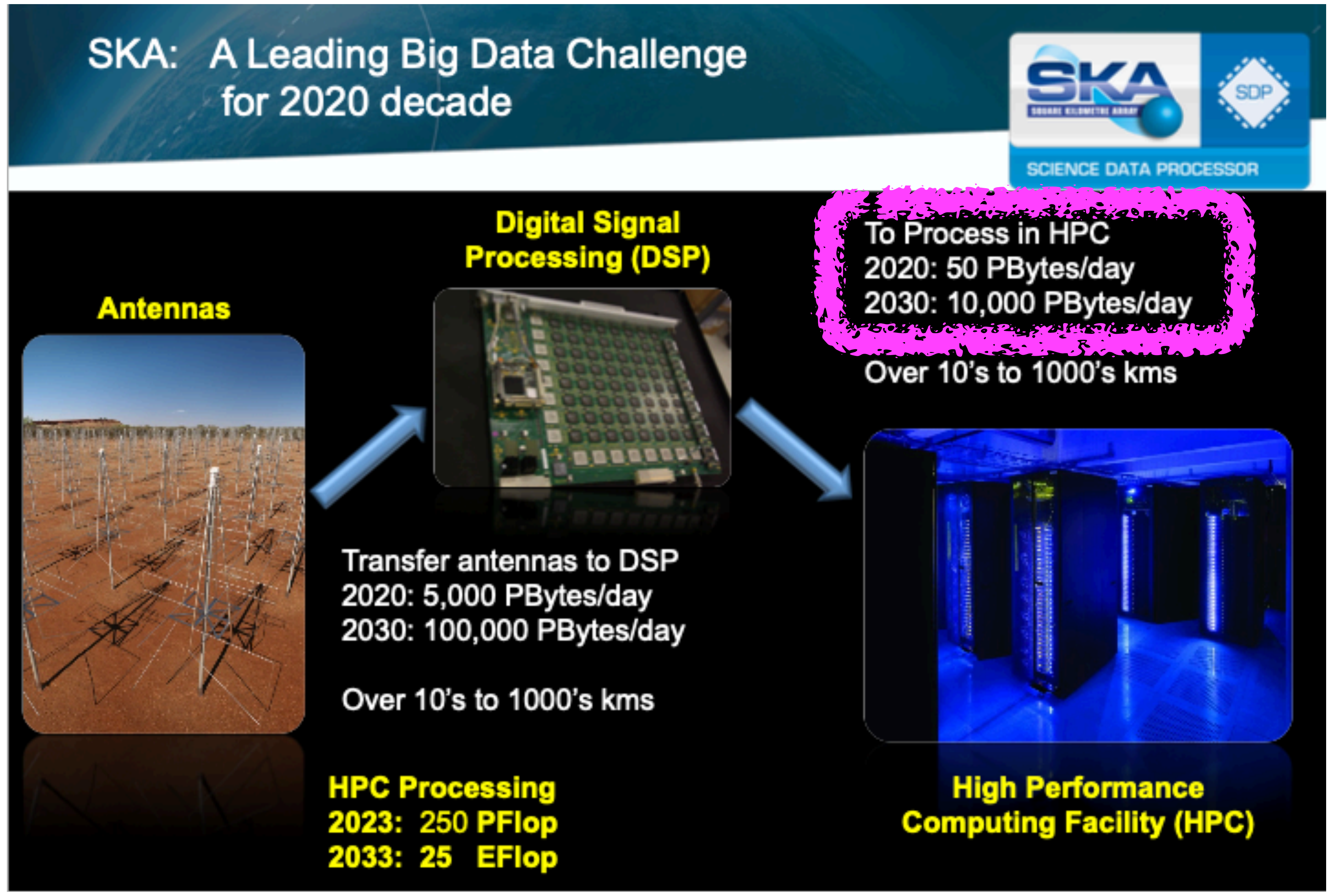
Next big step: ESA-GAIA, from space

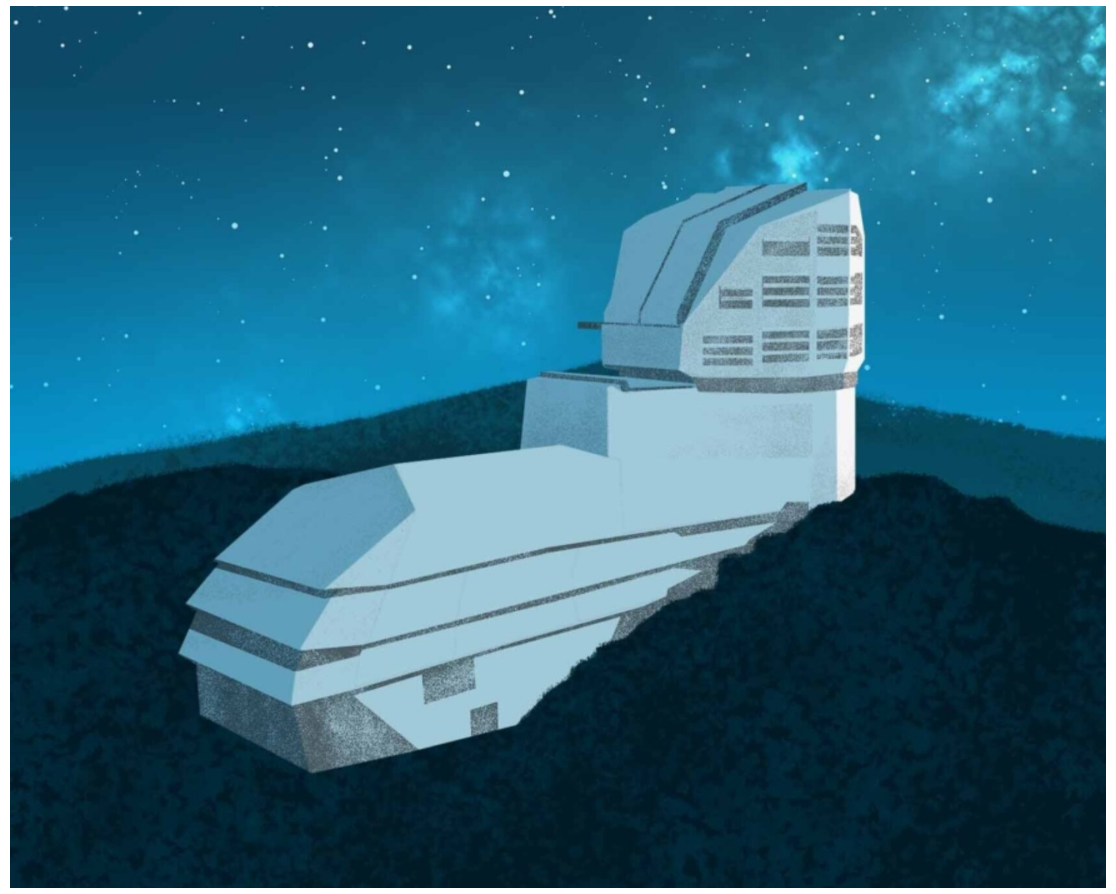
- Spectra for **470 million** objects, drawn from a catalog of 1.59 billion classified sources
- New concept of data center

Big data in astronomy: today and in the next decades

TRULY BIG data

- ***SKA***: **50 PB/day** growing
- ***LSST***:
  - "LSST ten-year survey will make more than five million exposures, collecting over **50 petabytes (5 PB/yr)** of raw image data to produce a deep, time-dependent, movie of about 20,000 square degrees of sky.





- ***DESI***:
  - **10 TB/yr** raw data
  - "After running the data through the pipelines at NERSC (using millions of CPU hours), there will be about **100 TB year of data products** that will be made available as data releases approximately once per year throughout DESI's 5 years of operations.

WST data flow

Back-of-the-envelope calculations (based on white paper specs)

Estimated raw-data volume per single exposure

- **MOS-LR**
  - multiplexing 20,000
  - R~4,000
  - Wavelength range 3700-9700Å → 6000 Å
  - Resolution element
    $$< \Delta\lambda > \simeq \frac{< \lambda >}{R} = \frac{6{,}700}{4{,}000} \simeq 1.6\text{Å}$$
  - Sampling 3 pix/res element → 1.6/3 = 0.5 Å per pix
  - Pixels per spec : 6000/0.5 = 12,000
  - Total pixels: 12,000 x 20,000 = 240 Mpix
  - 32 bit/pix → **~1 GB**

- **MOS-HR**
  - multiplexing 2,000
  - R~40,000
  - Wavelength range 3700-9700Å → 6000 Å, but only 3-4 windows for an effective range of ~4,000Å
  - Resolution element
    $$< \Delta\lambda > \simeq \frac{< \lambda >}{R} = \frac{6{,}700}{40{,}000} \simeq 0.16\text{Å}$$
  - Sampling 3 pix/res element → 0.16/3 = 0.05 Å per pix
  - Pixels per spec : 4000/0.05 = 80,000
  - Total pixels: 80,000 x 2,000 = 160 Mpix
  - 32 bit/pix → **~650 MB**

- **IFS**
  - $\mathrm{FoV} \sim 3 \times 3\ \mathrm{arcmin}^2 \implies$ for a spatial sampling of 0.25″/pix: $N_{\mathrm{spaxels}} \sim \frac{30{,}000}{0.25^2} \sim 500\mathrm{k}$
  - R~3,500
  - Wavelength range 3700-9700Å → 6000 Å
  - Resolution element
    $$< \Delta\lambda > \simeq \frac{< \lambda >}{R} = \frac{6{,}700}{3{,}500} \simeq 1.9\text{Å}$$
  - Sampling 3 pix/res element → 1.9/3 ~ 0.6 Å per pix
  - Pixels per spaxel : 6000/0.6 = 10,000
  - Total pixels: 10,000 x 500,000 ~ 5 Gpix
  - 32 bit/pix → **~16 GB**

Considering 3 exposures per hour and 10h per night:

$$18\ \mathrm{GB} \times 3\ \mathrm{exp} \times 10\mathrm{h} \simeq 540\ \mathrm{GB/night}$$

or $\lesssim 200\ \mathrm{TB/year}$

A lot compared to optical spectroscopic facilities (eg. DESI is 10 TB/yr raw), but modest compared to LSST (20 TB/night) or even negligible compared to SKA (50 PB/day!)

WST:
a *big-data* challenge

WST:
a *big* *data*-challenge

WST: a complex facility

- Different (possibly/partly simultaneous) **observing modes**:
  - Low-res MOS
  - Hi-res MOS
  - IFS

- Dramatically different **targets and requirements**:
  - Short to long integrations
  - Low to high SNR
  - Possible multiple co-adding
  - Requirements on sky subtraction and calibration

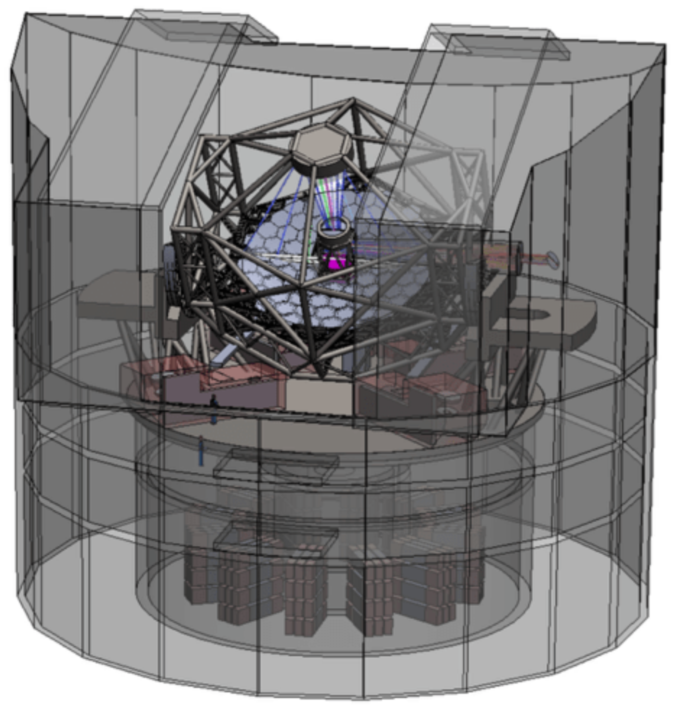
- Different **data products**, e.g.
  - absorption vs emission
  - Redshift/doppler kinematics
  - Spectral shape vs detailed features
  - ...

How to treat all this complexity in a unified data reduction and analysis system, maintaining that it has to live in a big data context?

A (very simplified) vision of the dataflow

Scheme of reduction-data analysis-science without all “bells and whistles”

Prototypical scheme adopted by WEAVE, 4MOST…



Raw data
+
Calibrations

→ **Data reduction** →

“L1” data products:

- Reduced spectra (individual)
- Reduced datacubes

→ **Data analysis** →

“L2” data products:

- Aggregated spectra (stacks, super stacks)
- Processed datacubes (spatial processing)
- “Objective” parameters: spectral indices, line EW, line fluxes, z, kinematics… (for individual spectra and in 2D)

→ **Science analysis** →

“Science” data products:

- Physical parameters (maps)
- Extracted catalogs

Data reduction challenges

Data Reduction System requirements

- **Supervise the data reduction flow**: how can we maintain full control of the dataflow for a large data-volume like WST?
- Implement **quality checks**
  - **Automation**: overall statistical quality assessment, anomaly detection, warning… ⇒ ***AI***?
  - **Visualization and interactive** analysis **tools**
  - Possibility of **intervention** along the data reduction flow
- Enable **flexibility and adaptation** in the reduction pipelines taking into account the diversity of observed datasets
  - Allowing for pipeline optimization and development is essential in the facility planning
  - Sustainability and reprocessing must be balanced
- Enable **different levels of "reduction accuracy"**, from "first look" to "optimized" (in different aspects related to specific surveys/goals)
  - *Time critical* reductions
  - *Interface with survey planning* to implement flexible observing strategies (e.g. flexible integration/visits based on achieved data quality)

Data reduction challenges

Data Reduction algorithms

- Open to new methodological approaches
  - Fast
  - Effective & computationally efficient
  - Matching the data complexity and flexible
- Critical challenges for WST surveys
  - Combining observations (stacks) in non-fully homogeneous conditions
  - IFS processing — challenges related to spatial reconstruction on very large FoV

- Machine Learning / AI
  - e.g. sky subtraction, telluric corrections, flux calibrations
- Parallel computing (GPU)
  - e.g. ML/AI, data compression, resampling, convolution
- HPC needed??

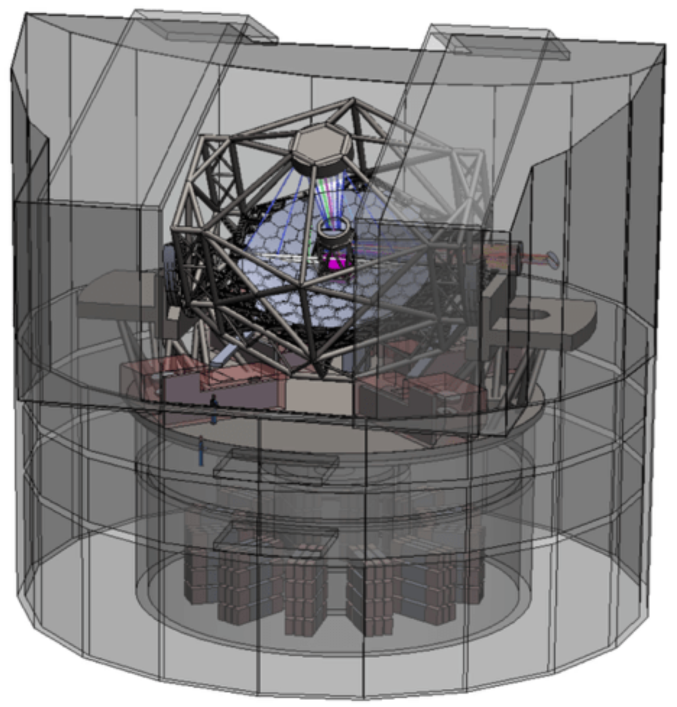
Data Analysis Challenges

A philosophical foreword

- What we call "scientific analysis" today, might become "standard data analysis" in 20 years
- However, a change of attitude is required in devising our "experiments", as **playing with the BIG data has a BIG cost**
  - Do we understand which ***information*** can be ***reliably extracted*** from spectroscopic data?
- **Flexibility**
  - Do we really want to *hardcode the production* of shaky parameters in a big data analysis pipeline? Or, vice versa, to be forced to *offline analysis* of parameters that were excluded years before the data are taken?
  - Data analysis **depends on the scientific question, no "standard"** (e.g. spatial processing of IFS cubes):
    allow science users to "play" with data analysis to get the most

A (very simplified) vision of the dataflow

Scheme of reduction-data analysis-science without all “bells and whistles”



Raw data + Calibrations

Data reduction

“L1” data products:
- Reduced spectra (individual)
- Reduced datacubes

Data analysis

Science analysis

“L2” data products:
- Aggregated spectra (stacks, super stacks)
- Processed datacubes (spatial processing)
- “Objective” parameters: spectral indices, line EW, line fluxes, z, kinematics… (for individual spectra and in 2D)

“Science” data products:
- Physical parameters (maps)
- Extracted catalogs

Data analysis and scientific analysis are tightly interlaced

Data Analysis Challenges

How and where to analyse the data?

- Who should (be allowed to) put her/his hands on the data analysis?
  - Move from the concept of a rigid and closed pipeline to a **new concept of distributed data-analysis environment**
    - cf. SDSS sky server at database level
    - Data-analysis plugins for a common pipeline skeleton?
    - Notebooks on cloud (LSST approach)?
    - Containers deployed to nodes?
    - Data analysis time/power granted in the same way as the observing time (BIG data analysis is a cost!)
- What we should **NOT** do with big data from a survey facility:
  - Have a (minimal?) set of **standard data products** from a standard pipeline
  - For anything else/more, download the data and **do it "at home"**

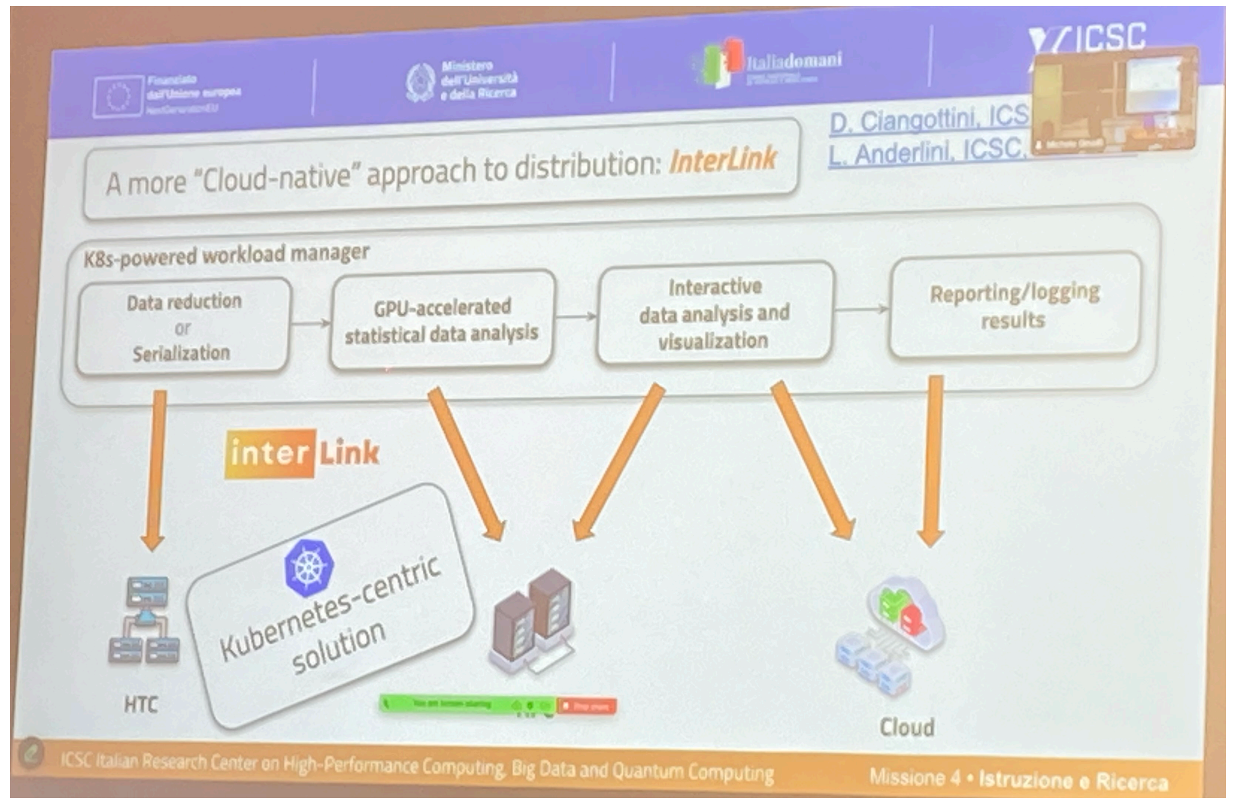
Re-thinking our data analysis tools

- We all aim at scientific results, not at **computing efficiency**:
  most (all?) of our astronomical spectral analysis software is not efficient, but just "good enough" for our current science goals
  - With an upscale of 100x of the data volume, rethinking our tools with the help of IT engineers is mandatory:
    a problem of TIME and ENERGY consumption, i.e. ***sustainability***
    - Improve codes
    - Consider radical **parallelization** (CPU —> GPU)
    - Consider if **alternative** approaches (e.g. Machine Learning/AI) can help
  - *Example*: decoupling the stellar continuum from the nebular emission in galaxies is currently done with a code (pPXF, M. Cappellari) that is great, but definitely not fast and not optimized to the specific goal
    - Could be re-written to run faster and parallel on GPUs?
    - Could it be done faster/better with ML?
    - Can we take advantage of some sort of data compression?

Which infrastructure?

Need to discuss with experts and engineers… but:

- We have to provide storage and processing capabilities as well
  - Data center(s)?
  - Distributed storage and computing?
- Let us get inspiration and guidance from other big-data experiments (e.g. LSST, space missions… SKAO), but also from other fields, e.g. particle and high-energy physics
- Dedicated brain-power in a well structured, stable environment can make the difference —> Bianca's legacy



InterLink: credits to L. Anderlini, D. Spiga, D. Ciangottini @INFN

Time for thinking
and
for discussion

Thanks!