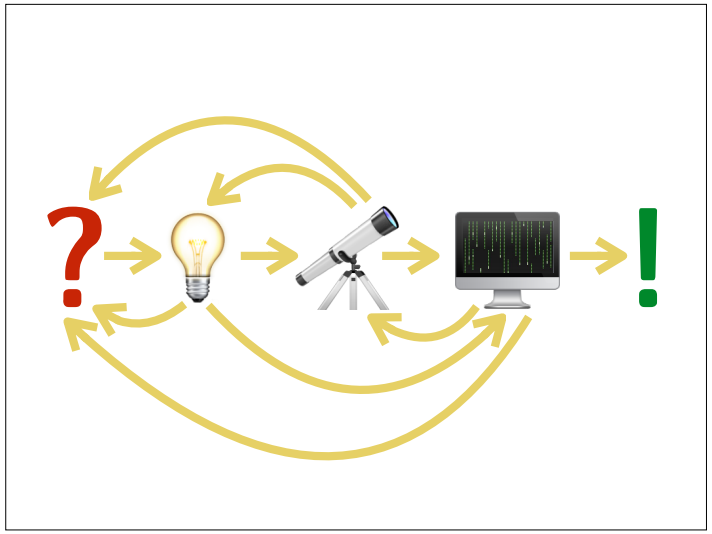
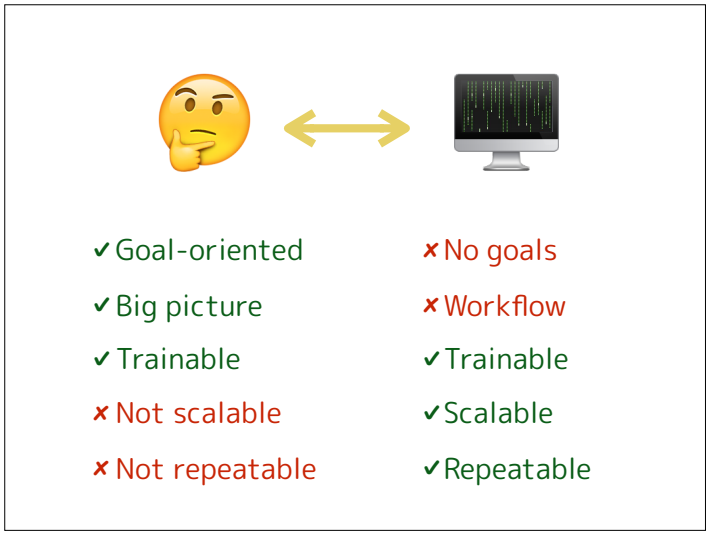
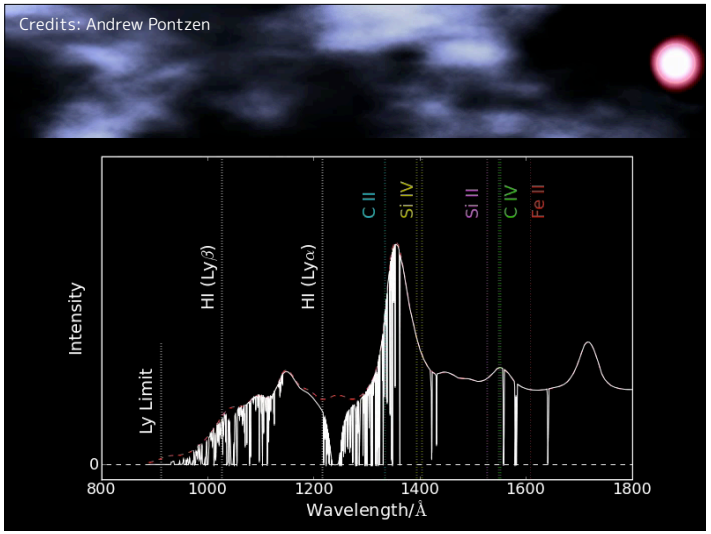
In this talk, I won't be speaking specifically about science and scientific results, but about how this results are achieved, i.e. about the treatment of the data we collect with our telescope and instruments.
As you will see, data treatment is becoming an integral part of the instrument design. This trend is going to be confirmed in the extremely-large telescope era, as the science cases become more sophisticated and increasingly relying on a proper analysis procedure.
I will focus in particular on spectroscopic data in the near-UV to near-IR band. I will use the example of VLT ESPRESSO to describe the state of the art and the near future prospects of the data analysis software.



Let me take a step back and schematize for you how science is done. We typically start from a question and a possible idea to answer it. What we do as astronomers is collect observations to either confirm or reject the idea. But this observations come in the form of digital data that needs to be analyzed. Only after a careful analysis we can actually obtain an answer to our original question. Actually, reality is much less straightforward than this, but the point is clear: a computer-assisted treatment of the observational data is an unavoidable step of our scientific work. And pretty often a time-consuming one.
The same treatment must be applied, of course, to simulated data used to test our understanding of the actual phenomena. This is why I think it is worthy, on a conference about ELT science, to consider also the data treatment step.



Let me stress that a data treatment system is not merely a computer crunching data. I'd rather picture it as a combination of human mind and machine. The two can be roughly considered complementary in their capabilities. The basic idea is that the humans know what science they want to do, and that machines can help them to scale up their innate pattern-finding capabilities to scan large data samples. Of course, there is currently a great deal of discussion about how much machines can actually emulate (or reproduce) the best feats of the human mind. We all know that machines can learn (although probably not in the exact same way as we do). Anyway, for the time being, the cooperation between the two actors is unavoidable. Humans still matter.

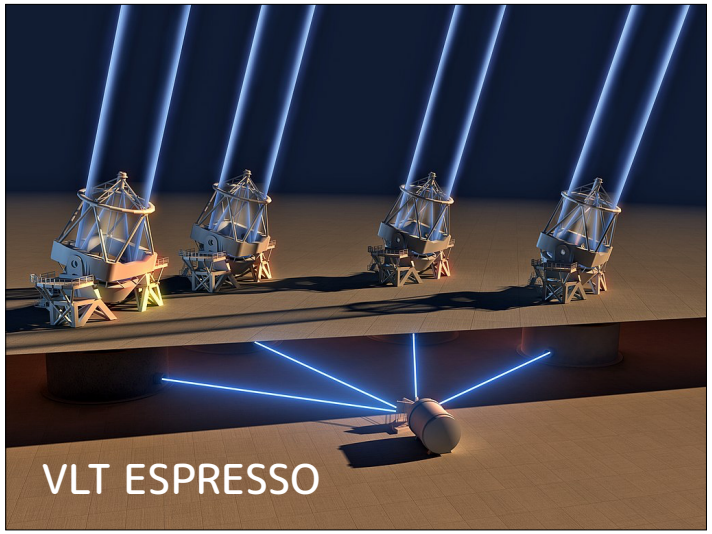


And this is true also in the specific field of the analysis of quasar spectra.
The picture display the typical features of a quasar spectrum. We have an original AGN emission spectrum, that is progressively redshifted and absorbed as the photons move through the intergalactic medium (and the occasional galaxy) to reach us. Both emission and absorption features contain a plethora of information to be interpreted, about the physics of the emitting source and the chemical-dynamical state of the intervening matter.
What I want to stress is that all this information is not immediately accessible. We need to disentangle the different features and properly interpret them according to our knowledge of how photon emission and absorption work. Here's where the human-machine cooperation comes into play: humans are very good at identifying and discriminating spectral features, but they are not equally good at repeating the task thousands of times, neither at assessing the uncertainty of physical values extracted from the data.

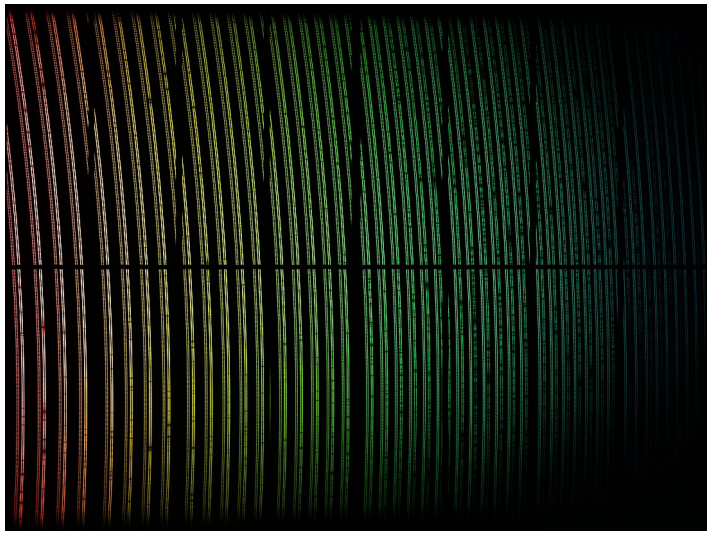
- Efficiency
- Resolution
- Stability
- Flexibility
- …

- Reliability
- Repeatability
- Scalability
- Flexibility
- …

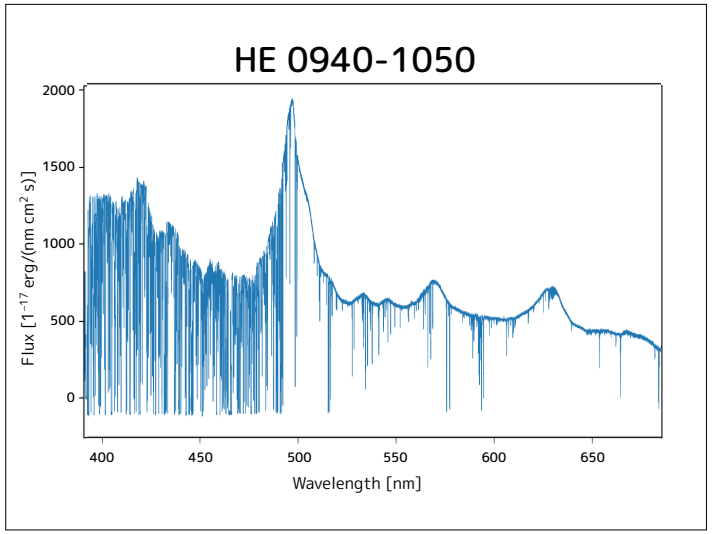
So, we have two sets of technological requirements, one for the instrument hardware, one for the instrument software.
The second one was hardly taken into account until recent times. Data treatment tools were created in-house by researchers, patching software from different sources without proper validation. Generalist software suites like IRAF or ESO-MIDAS were available, but they were targeted mostly to low level tasks. Nowadays they offer limited user-support. Instrument-oriented tools were limited to the reduction procedure, as I will describe shortly.



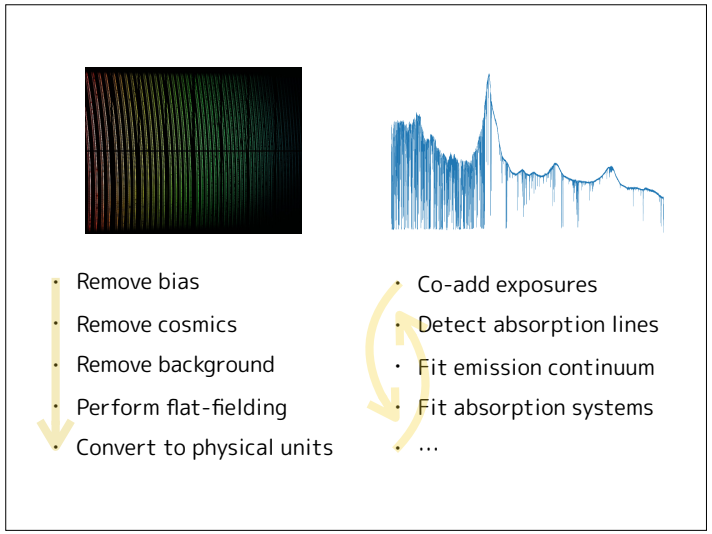
A new approach was implemented with VLT ESPRESSO, the high-resolution spectrograph currently under commissioning you have already heard of. Due to its restricted number of science cases, ESPRESSO was designed since its inception as a “science machine” to perform measurements in realtime. It is equipped with a dedicated Data Analysis Software to analyze both stellar and quasar spectra, in addition to the standard Data Reduction pipeline which is provided to all VLT instruments.



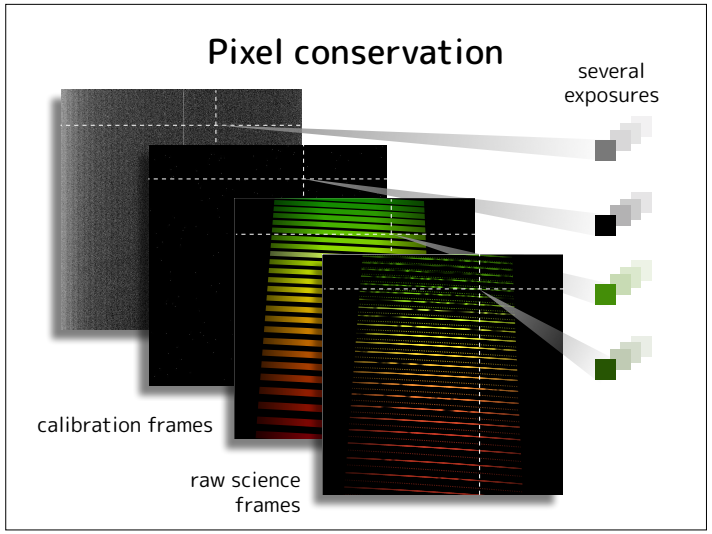
To better understand what I am talking about, I show you what an ESPRESSO spectrum looks like. In this picture you can easily see the cross-dispersed echelle orders, each one split into two fiber traces, dedicated respectively the spectrum calibration source (or the sky) and the target. Simultaneous calibration adds up to the instrument thermo-mechanical stability as a way to maintain a steady calibration for long periods of time.
I call data reduction the set of operations needed to transform this raw spectrum…



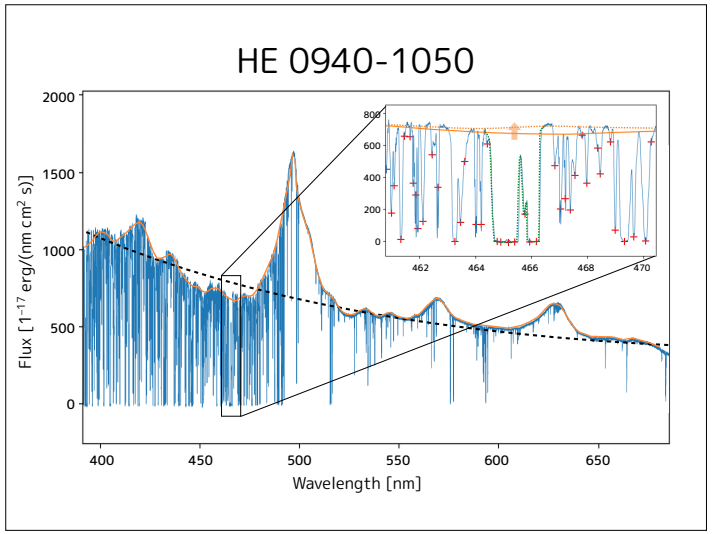
…into this. Here you can see a nicely calibrated flux-vs-wavelength spectrum of this quasar, with multiple emission features (Lyman alpha, CIV) and absorption lines, due to both neutral hydrogen and metals. All the operations needed to extract information from the reduced spectrum are referred to as data analysis.



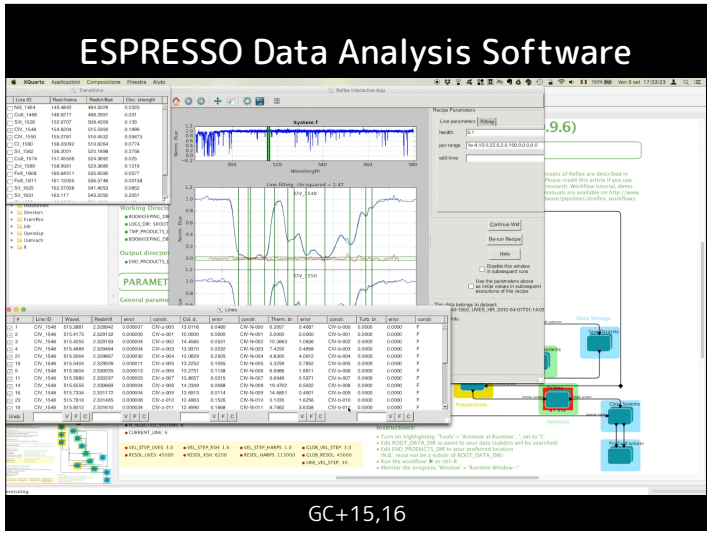
This is a list of the operations to be performed on each input. You can easily see that the reduction procedure is somehow linear and one-way. Once the spectrum is reduced, there's in principle no need to go back to the raw frames. The same doesn't hold for the analysis.
The list of analysis operations here is clearly incomplete, because they are in principle countless. Even more importantly, they do not form a linear workflow: in a typical analysis session, one wants continuously to go back and revise operations in the light of the new information that has been extracted. This of course must be reflected in the way the data treatment code is designed.



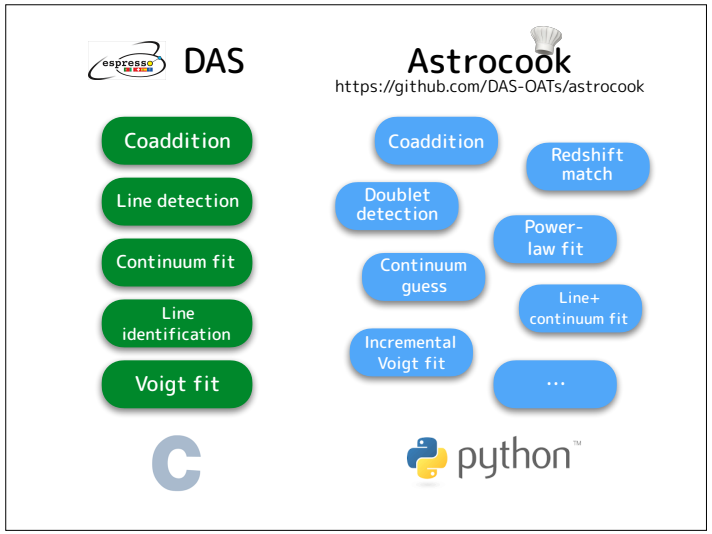
For ESPRESSO, we adopted an hybrid approach that was favored by the existing ESO data flow system: we designed the data analysis software as a pipeline, much like the data reduction. For the data reduction, a conservative approach was chosen, in which all pixels in the reduced spectra can be traced back to the original detector pixels they come from. This approach impacts on the analysis, too, as it allows the user to fit their models on data that were not previously rebinned (thus preserving the error statistics of the detector readings).



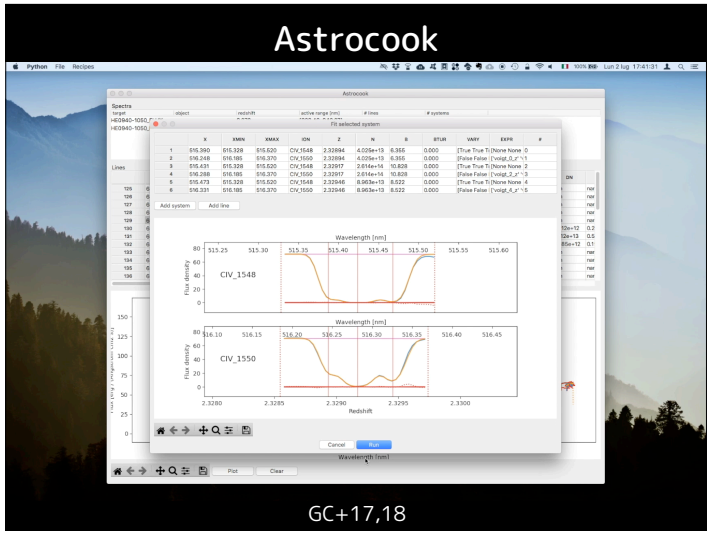
As for the quasar analysis part, we addressed the problem of interpreting simultaneously the emission an the absorption features in the spectra. In principle, one needs to first determine the level of continuum emission to fit the lines, but the continuum itself is well constrained only when all the lines have been fitted and removed. The problem is particularly serious in the Lyman alpha forest, were lines are crowded and blended and the neutral hydrogen opacity may account for a diffuse absorption not associated with detectable lines.
The ESPRESSO data analysis software solves the problem iteratively. A preliminary fit of the lines is performed with respect to a guess continuum (even a bare power-law) and is used to incrementally refine the continuum determination.



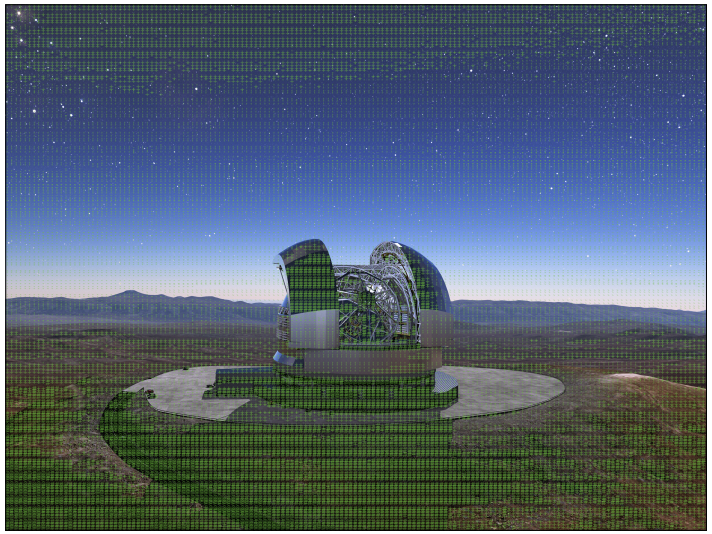
Here is a snapshot of the software at work, as it appears through the standard ESO Reflex pipeline interface.
You can see that after the continuum is refined, lines are associated with each other by discriminating possible coincidences between transitions (different species that happens to be at the same redshift). These coincidences are selected as “absorption systems” and fitted accordingly, with composite Voigt profiles, after imposing the right constraints between associated system components.



The system is working, but has two main drawbacks: (1) it can be applied only to ESPRESSO data, (2) it doesn't allow for fine tuning of the workflow. The basic operations (like the continuum determination) are closed boxes and cannot be split into steps, to implement better iteration schemes. Such approach is obviously not suited to the typical user case in the ELT age, when flexibility and synergy between different data sources are paramount.
This is the reason why we started a new project, dubbed "Astrocook". Astrocook is a new suite of procedures which is meant to provide "a thousand recipes to cook a spectrum". Compared to the DAS, the operations are more fine-grained and can be organized into custom-made workflows. The suite is currently implemented as a Python package and is built on NumPy and Astropy standard libraries. It is already interfaced with SDSS and ESO data formats and is equipped with a graphical user interface.



This is how the Astrocook graphical user interface looks like. At face level, it may look similar to the ESPRESSO software, but is much more flexible. As one can see, the operations in the "Recipes" menu become available as new information is produced (e.g.: continuum estimation become available once absorption lines are detected), and they can be iterated at will, taking advantage of the results of the operations already performed. This allows the user to maintain control over the procedure while fully automatizing the repeated tasks.
We are also planning to develop an interface to design workflows. The idea is that the workflow should be designed and executed on a test dataset, and then automatically applied to a larger data sample.
It is noteworthy that Astrocook runs equally on observed and simulated data. This already proved useful e.g. in assessing the completeness of doublet detection (results upcoming).



All this work is meant to explore the possibilities of data analysis tools before the advent of the 40-m class telescopes. The field is still quite new but looks really promising.
I am already involved with the Trieste team in the design of the data analysis software for the ELT HIRES spectrograph. With Astrocook we are aiming to test and validate a set of good data treatment practices. To me, this is the only way to make all the exciting science cases you heard about feasible.