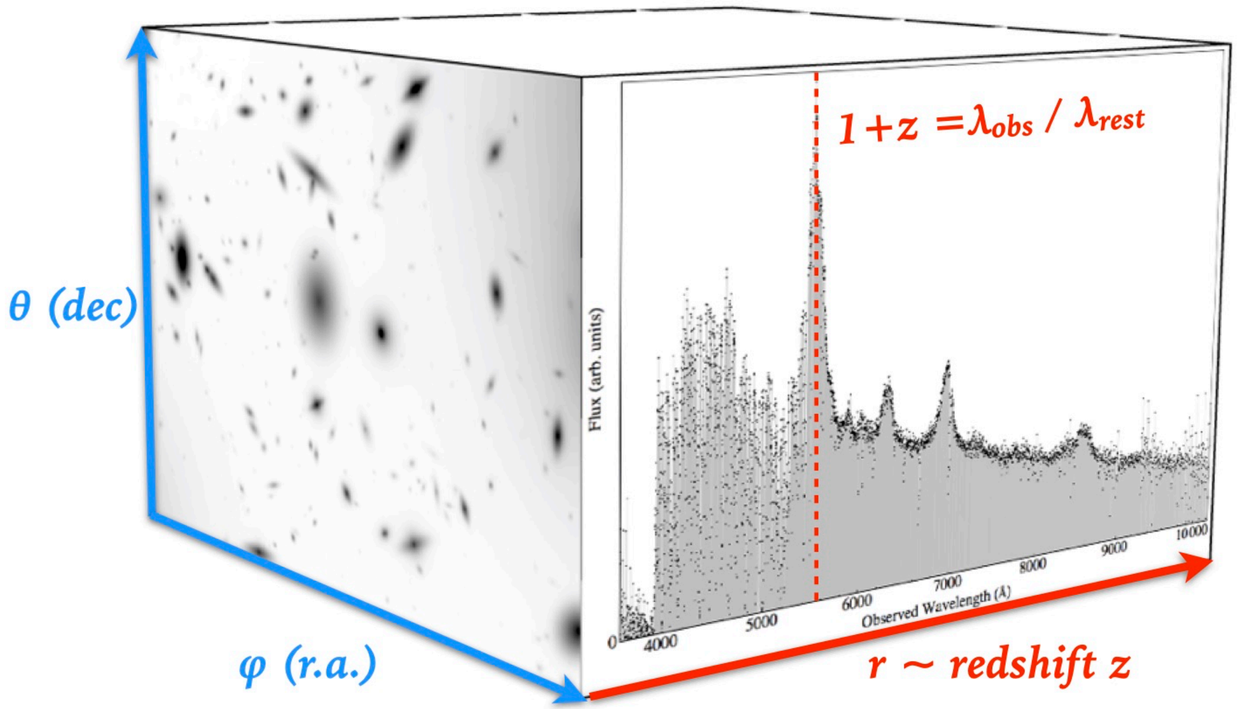
Open data with DESI - Dark Energy Spectroscopic Instrument

Eva-Maria Mueller

14th May 2026

What is our data?



The optical sky is a data cube: $\theta \times \varphi \times r$

How is the data analysed?

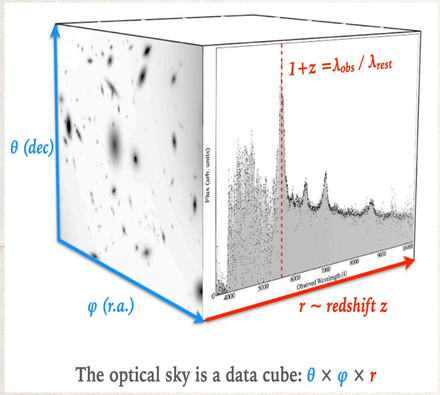
Spectra

Catalogs

Compressed Statistics

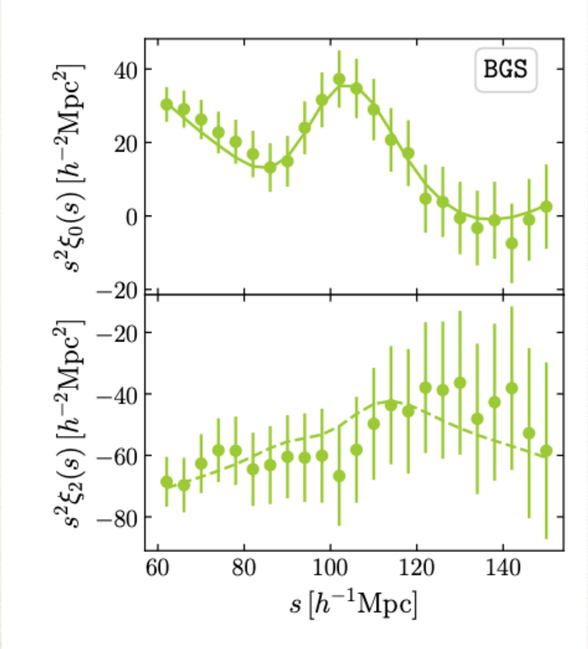
Likelihoods

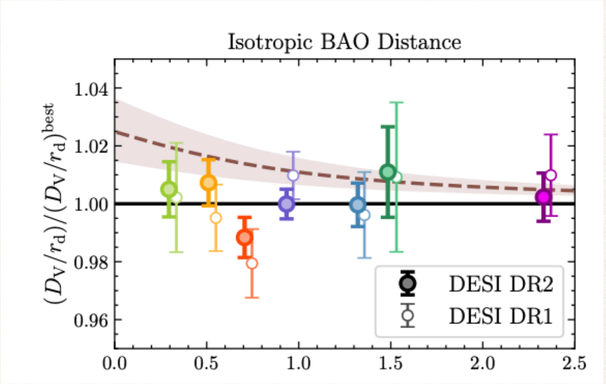
Cosmological Interpretation

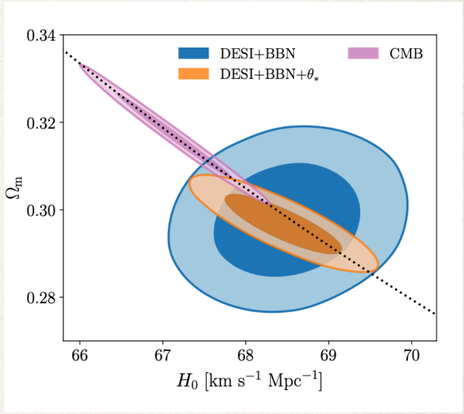


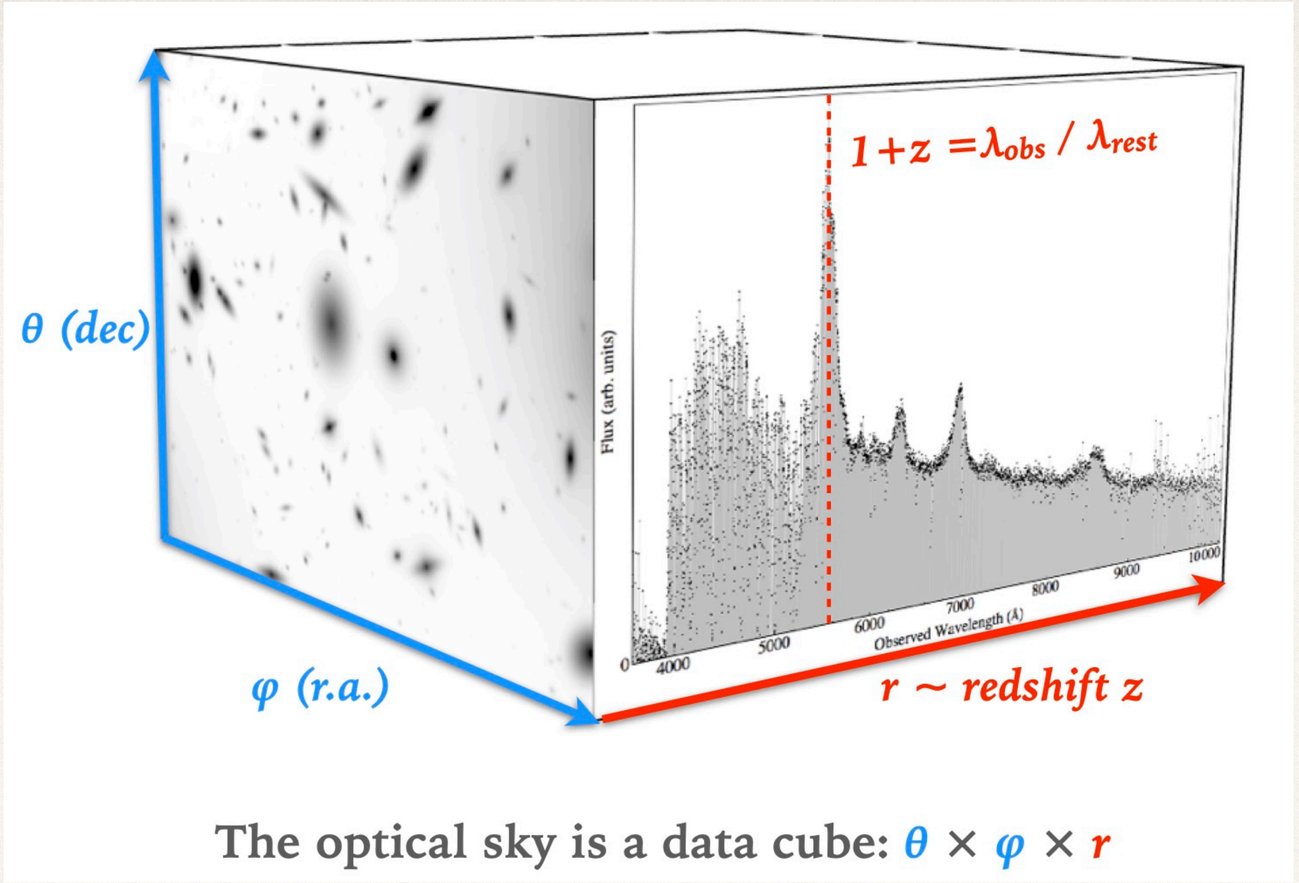
The optical sky is a data cube: $\theta \times \varphi \times r$

List of:
Object, θ, ϕ, z









The optical sky is a data cube: $\theta \times \varphi \times r$

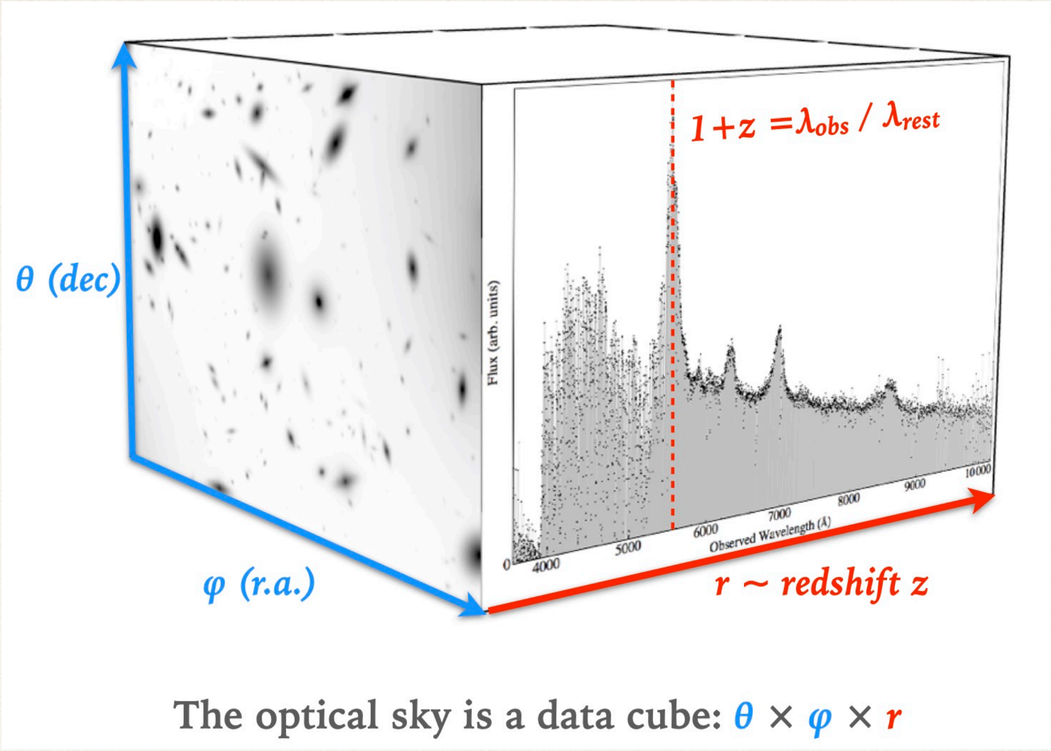
List of:
Object, θ, ϕ, z

Data Products: Raw Spectra + Catalogs

Fit spectra to get redshift aka. distance

Data Products: Value added Catalogs (VACs)

- Curated subsets of data with additional analysis beyond the core spectroscopic data processing pipeline outputs
- Many different types of VACs each accompanied with a publication



The optical sky is a data cube: $\theta \times \varphi \times r$

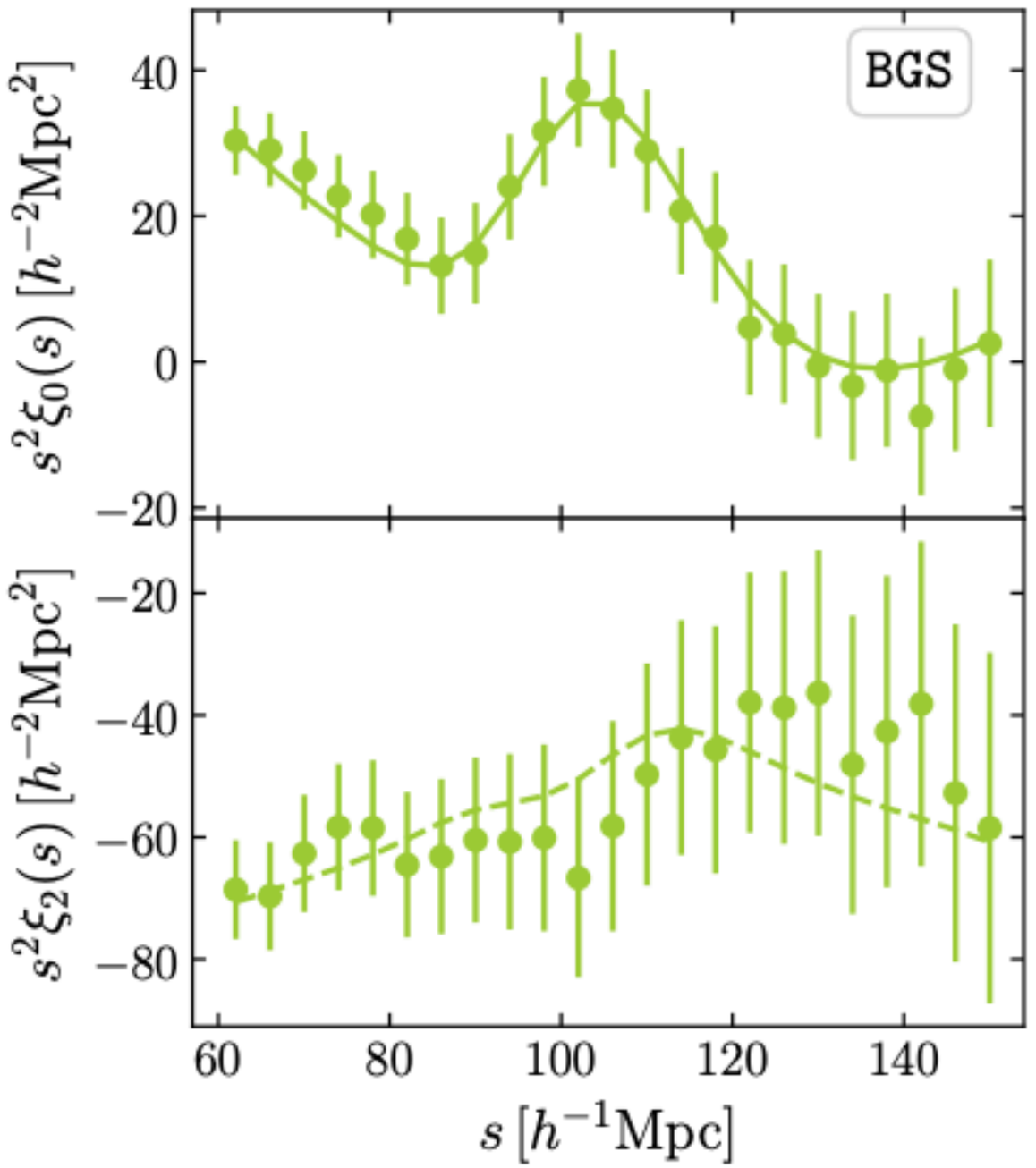
x velocity
x mass
x temperature …

Data Products: Large Scale Structure Catalogs

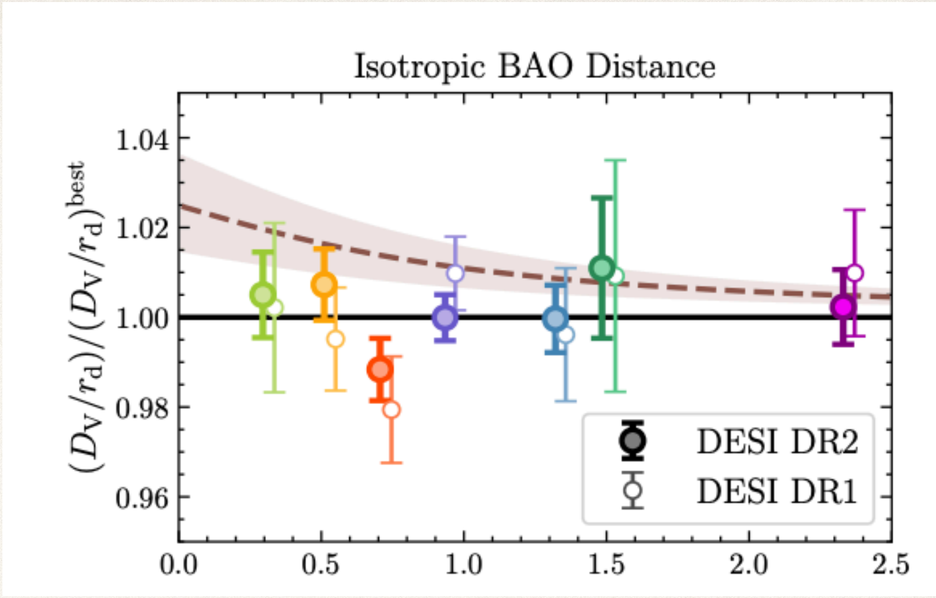
- “Cosmology-ready" catalogs
- Data “cleaned" of systematics
- Cosmology VAC

Data Products: Compressed Statistics

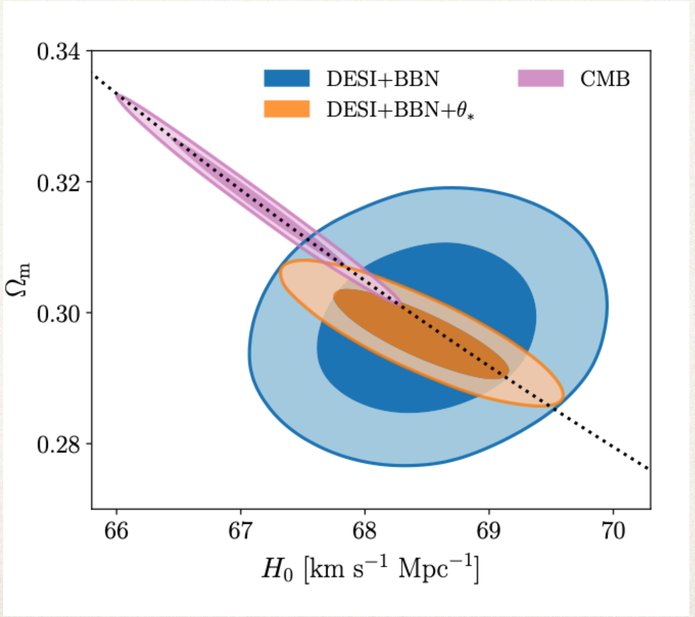
- 100s of data bins instead of millions of objects
- Contains most of the information on the distribution of galaxies in the Universe
- Intermediate step processing the data



Data Products: Likelihoods



- Sorry, we are mostly Bayesians (there is only one observable Universe)…
- A further compression step for speed



How is the data used in the community?

Spectra → Catalogs → Compressed Statistics → Likelihoods → Cosmological Interpretation

- New / different value added catalogs
- New / different redshift fitting

- New / different compressed statistics

- New / different model for systematics
- New theory model of the compressed statistics

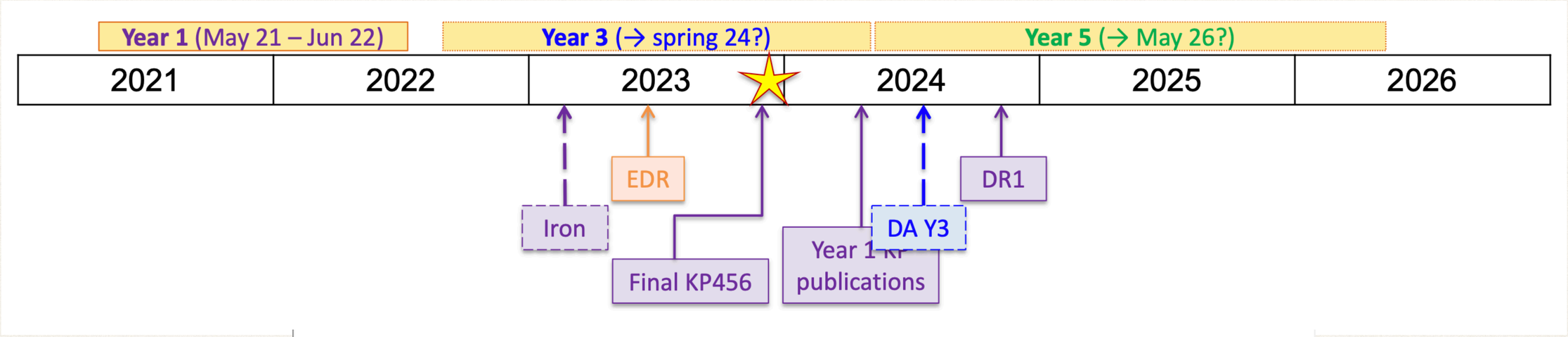
- New / different cosmological model
- Combination with external data sets

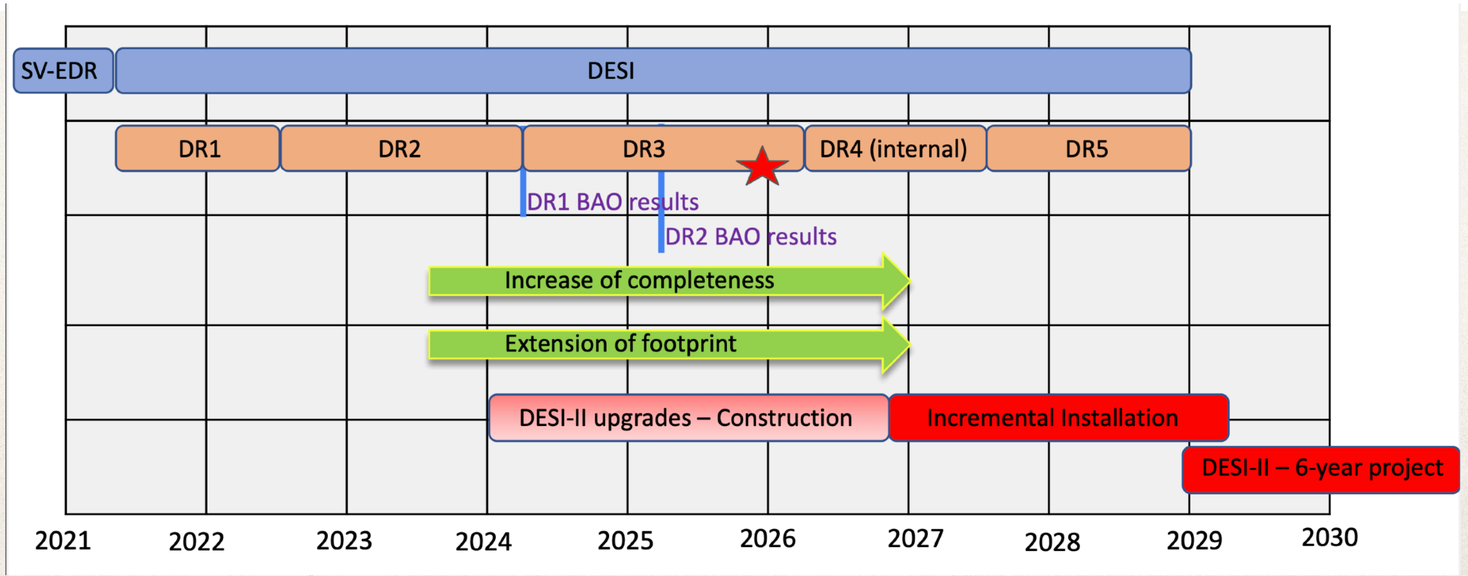
Astronomy application:

- galaxy formation
- …endless list

Public: Data + Software

Timeline


Year 1 (May 21 – Jun 22)
Year 3 (→ spring 24?)
Year 5 (→ May 26?)
2021
2022
2023
2024
2025
2026
Iron
EDR
Final KP456
DA Y3
DR1


SV-EDR
DESI
DR1
DR2
DR3
DR4 (internal)
DR5
DR1 BAO results
DR2 BAO results
Increase of completeness
Extension of footprint
DESI-II upgrades – Construction
Incremental Installation
DESI-II – 6-year project
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030

DESI policies

- For key projects (papers): data and analysis pipeline must be made public
- Data availability statement for each publication
- Zenodo repo for all DESI publications

Data Releases

- Website (Access, summary, information etc)
- Publication for each data release
- Github repository to recreate figures of the data release paper (notebook, processed data, figures)
- Github repo of all the software used to process the data
- Jupyter notebook tutorials on how to use the data

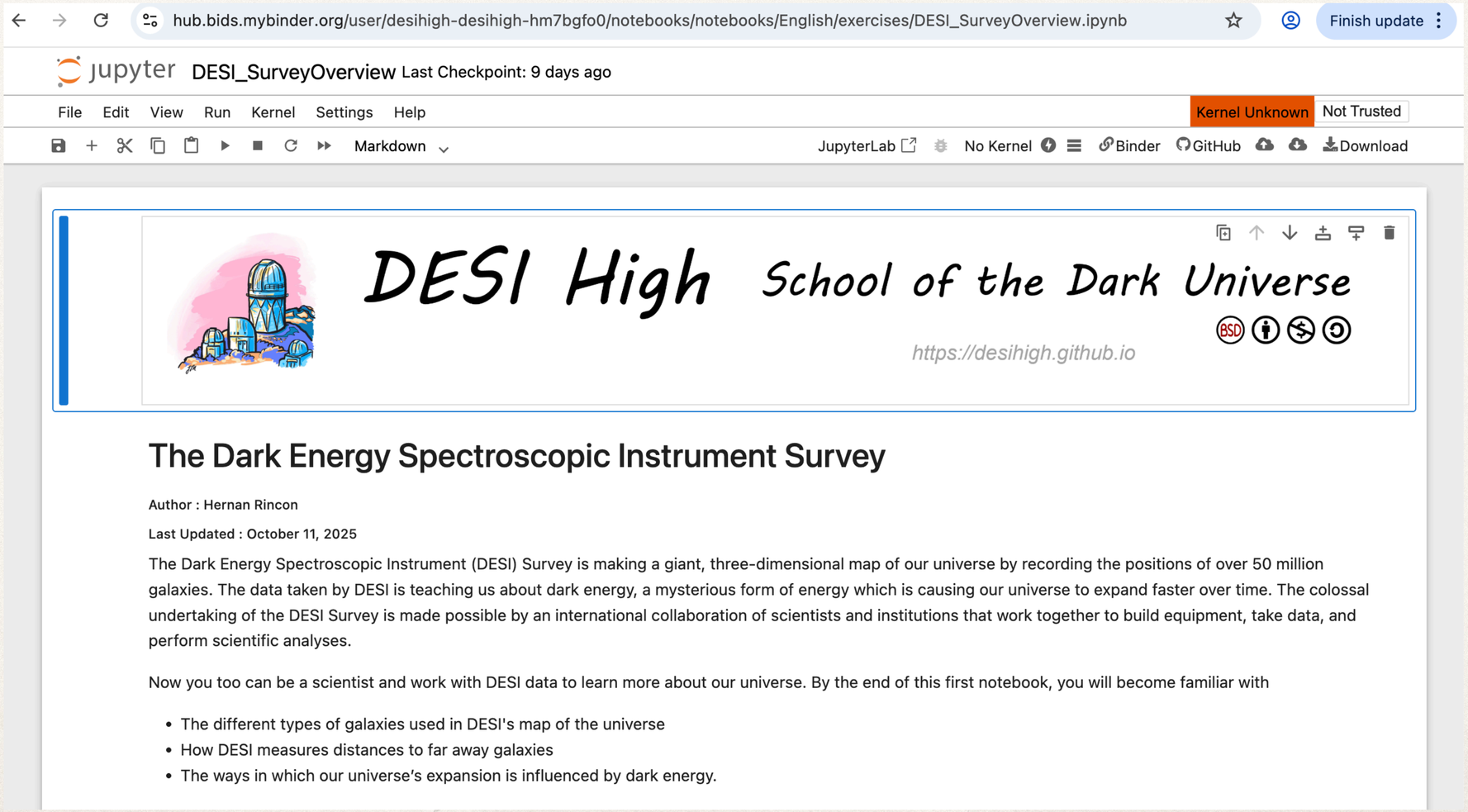
https://data.desi.lbl.gov/doc/

DESI management team

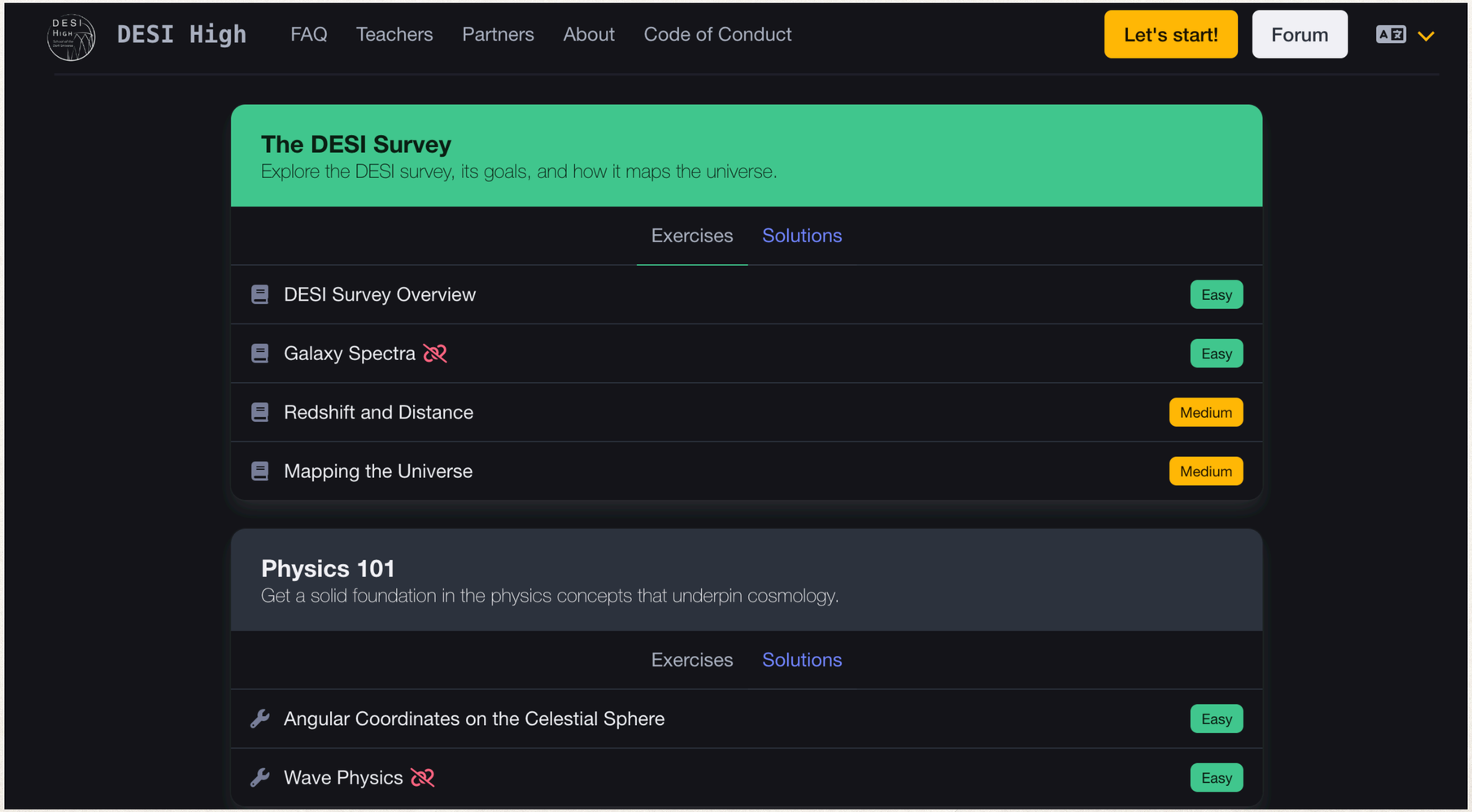
Data System (WBS 1.8)

- Stephen Bailey stephenbailey@lbl.gov - Manager
  - David Kirkby dkirkby@uci.edu - Data System Scientist
  - Kyle Dawson kdawson@astro.utah.edu - Survey planning and operations
  - Robert Cahn rncahn@lbl.gov - Fiber Assignment Lead
  - Julien Guy jguy@lbl.gov - Spectroscopic Pipeline Lead
  - Lado Samushia lado@phys.ksu.edu - Large Scale Structure Catalog
  - Benjamin Weaver - Data Transfer, Archive & Distribution
  - Stephen Bailey stephenbailey@lbl.gov - Integration & Testing of Data Systems
  - Dustin Lang dstn@cmu.edu - DESI Imaging Scientist
  - Adam Myers geordiemyers@gmail.com - Target Selection Pipeline

DESI High for High School Students



DESI High for High School Students

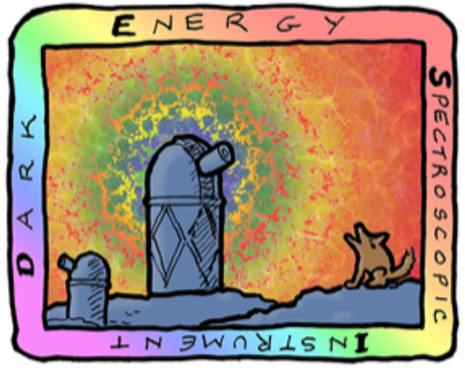


How can we avoid miss-usage of our data?

Data processing is complex…

Is a global platform for higher eduction resources using open data possible?

Each collaboration or research field has resources…the problem is knowing that it exists and where to find it!



DARK ENERGY SPECTROSCOPIC INSTRUMENT

U.S. Department of Energy Office of Science

Anyone should be able to reproduce key results!

Guiding principle

Key points

- Challenge: How to avoid miss-usage of our data?
- Future of open data in HE: Is a global platform for resources possible?
- Guiding principle: Anyone should be able to reproduce our key results