Advanced Algorithms for MOS Surveys

Adam S. Bolton

Department of Physics and Astronomy
The University of Utah

Multi-Object Spectroscopy in the Next Decade
Isaac Newton Group, La Palma
2015 March 02

What I will talk about

Computational data processing and statistics for inference and discovery in MOS surveys

Not comprehensive, but hopefully interesting

Will make over-generalizations for narrative effect

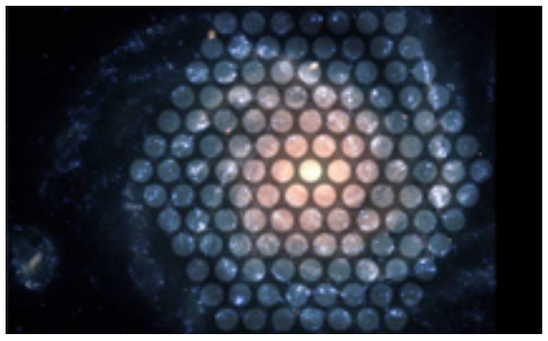
Will shamelessly take opportunity to focus on the work of those immediately around me

Input from the audience is welcome!

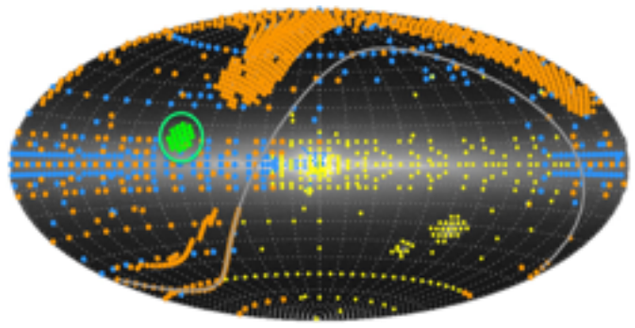
A view from within the SDSS

SDSS-IV (2014-2020)

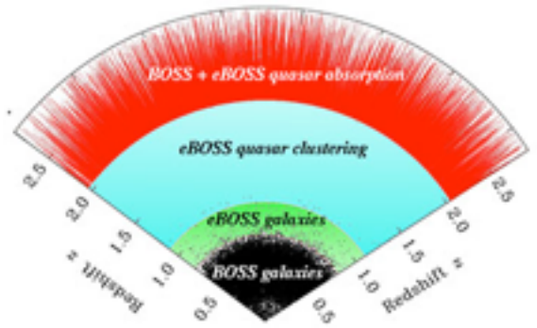
MaNGA



APOGEE-2



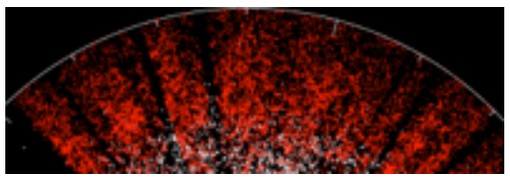
eBOSS



SDSS-III (2008-2014)

APOGEE

BOSS



MARVELS

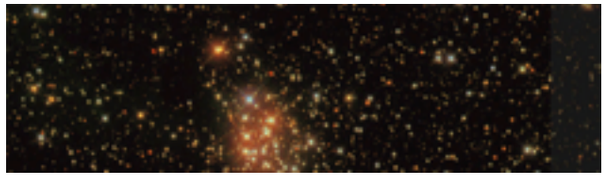
SEGUE-2

SDSS-I & -II (2000-2008)

Legacy Survey

Supernova Survey

SEGUE-1



An apparent golden age

Algorithmic/statistical astronomy now enjoys...

- Richness of instrumentation and surveys
- Broad appreciation for statistics and “data science”
- Value in astronomy for software along with hardware
- Many venues for communication
- Wide availability of powerful computing
- Lots of people actively developing the skills

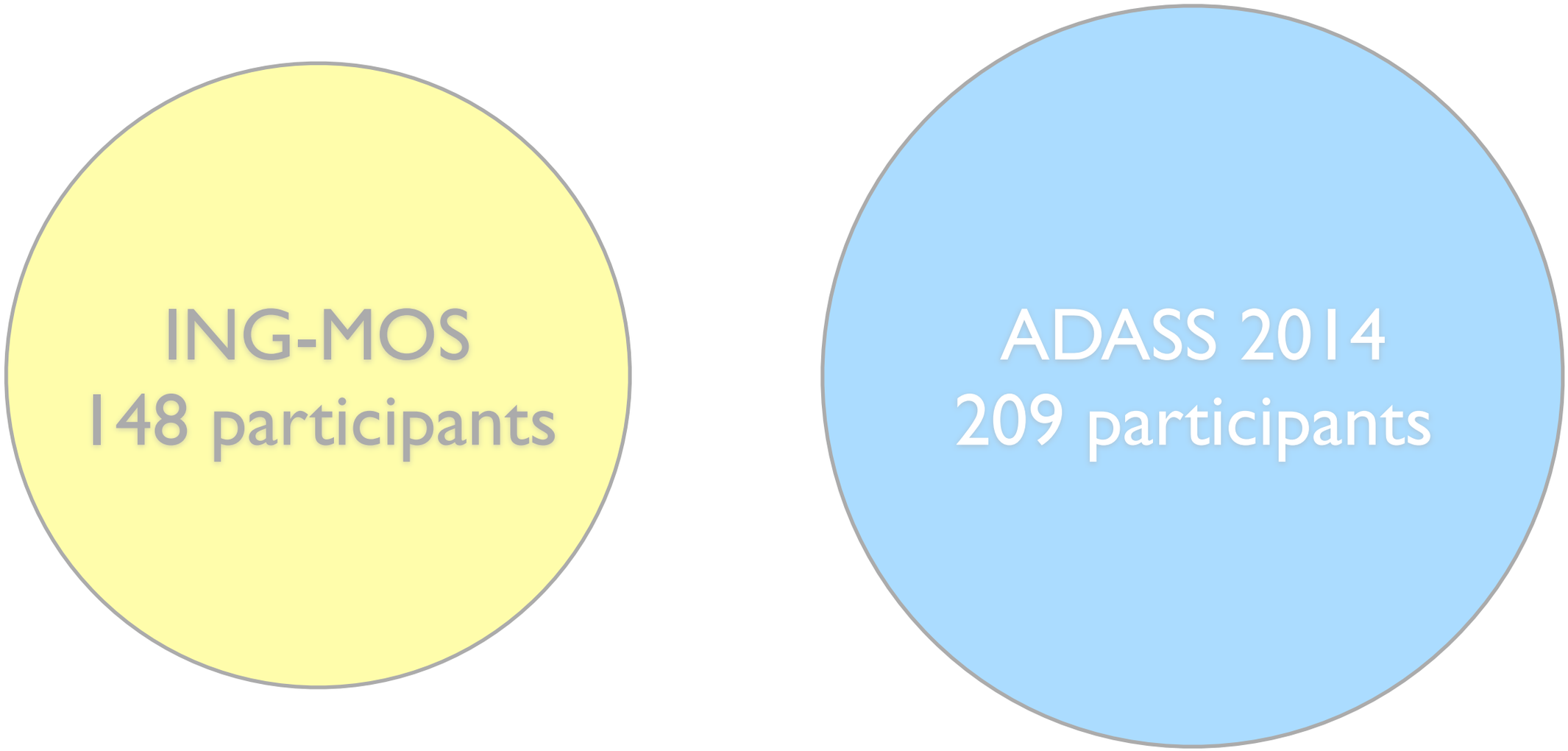
And yet...

There remains significant unrealized potential

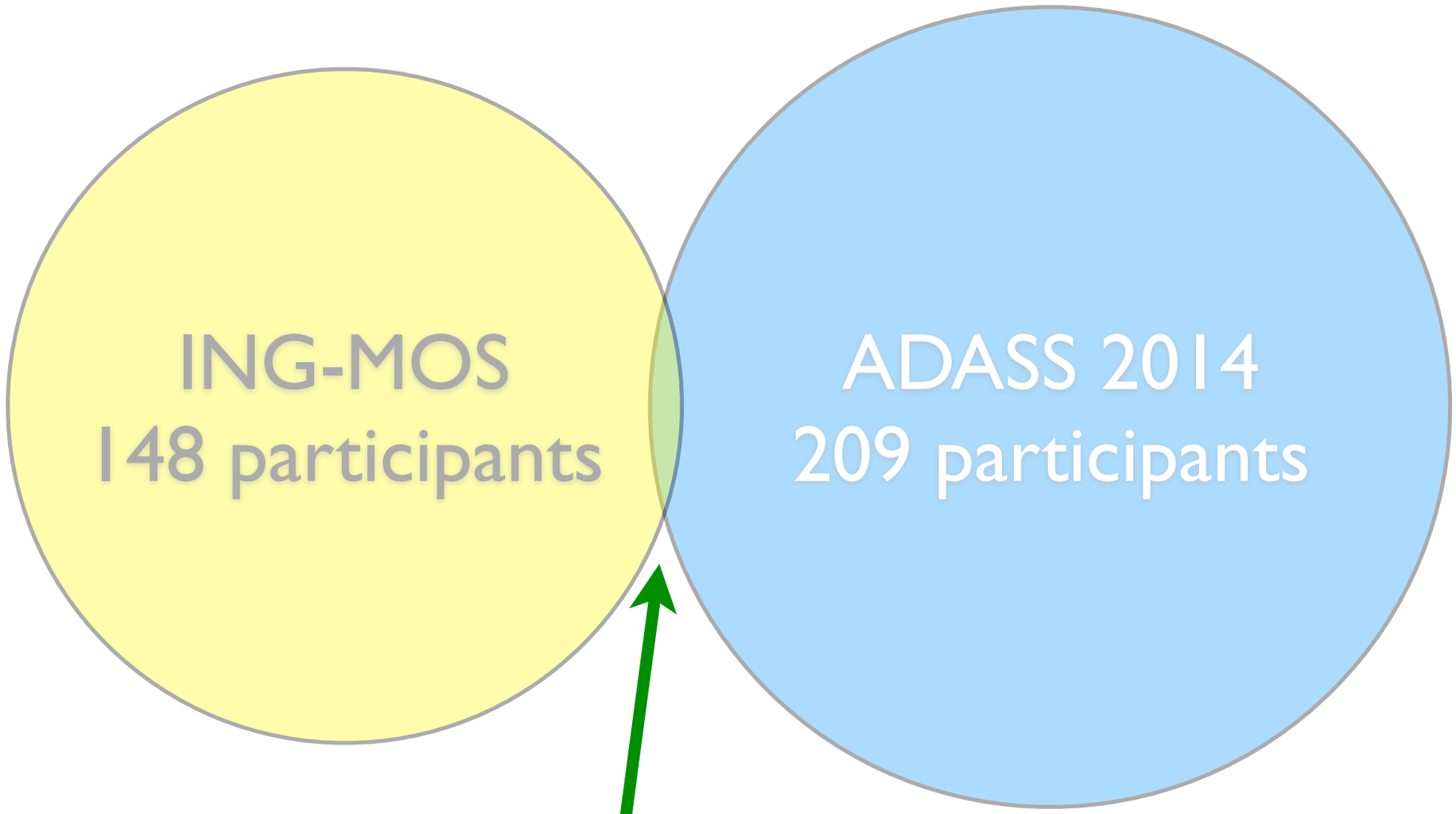
- *To make the scientific efforts of MOS surveys more algorithmically and statistically sophisticated*

- *To drive developments in the computational and statistical communities with MOS science*

ING-MOS overlap with ADASS 2014



ING-MOS overlap with ADASS 2014



Overlap = 2 participants

Spectroscopy and statistics

Among astronomers, spectroscopists are particularly un-Bayesian

The spectroscopist's classic perspective is:

"Don't talk to me about *inference*. If you want to know something, just ____ing *measure* it. That's what a spectrograph is *for!*

Statistics and spectroscopy?

Are Bayesians (& statisticians in general) unSpectroscopic?

E.g.:

"Don't talk to me about *measurement*. If you want to know something, just ____ing *infer* it. That's what a statistics are *for!*

History, Part 0

Theorists have always been hackers, even before there were computers to hack with

THE ASTROPHYSICAL JOURNAL

AN INTERNATIONAL REVIEW OF SPECTROSCOPY AND ASTRONOMICAL PHYSICS

VOLUME 94 NOVEMBER 1941 NUMBER 3

ON THE CLUSTERING TENDENCIES AMONG THE NEBULAE

II. A STUDY OF ENCOUNTERS BETWEEN LABORATORY MODELS OF STELLAR SYSTEMS BY A NEW INTEGRATION PROCEDURE

ERIK HOLMBERG

History, Part 0

Theorists have always been hackers, even before there were computers to hack with

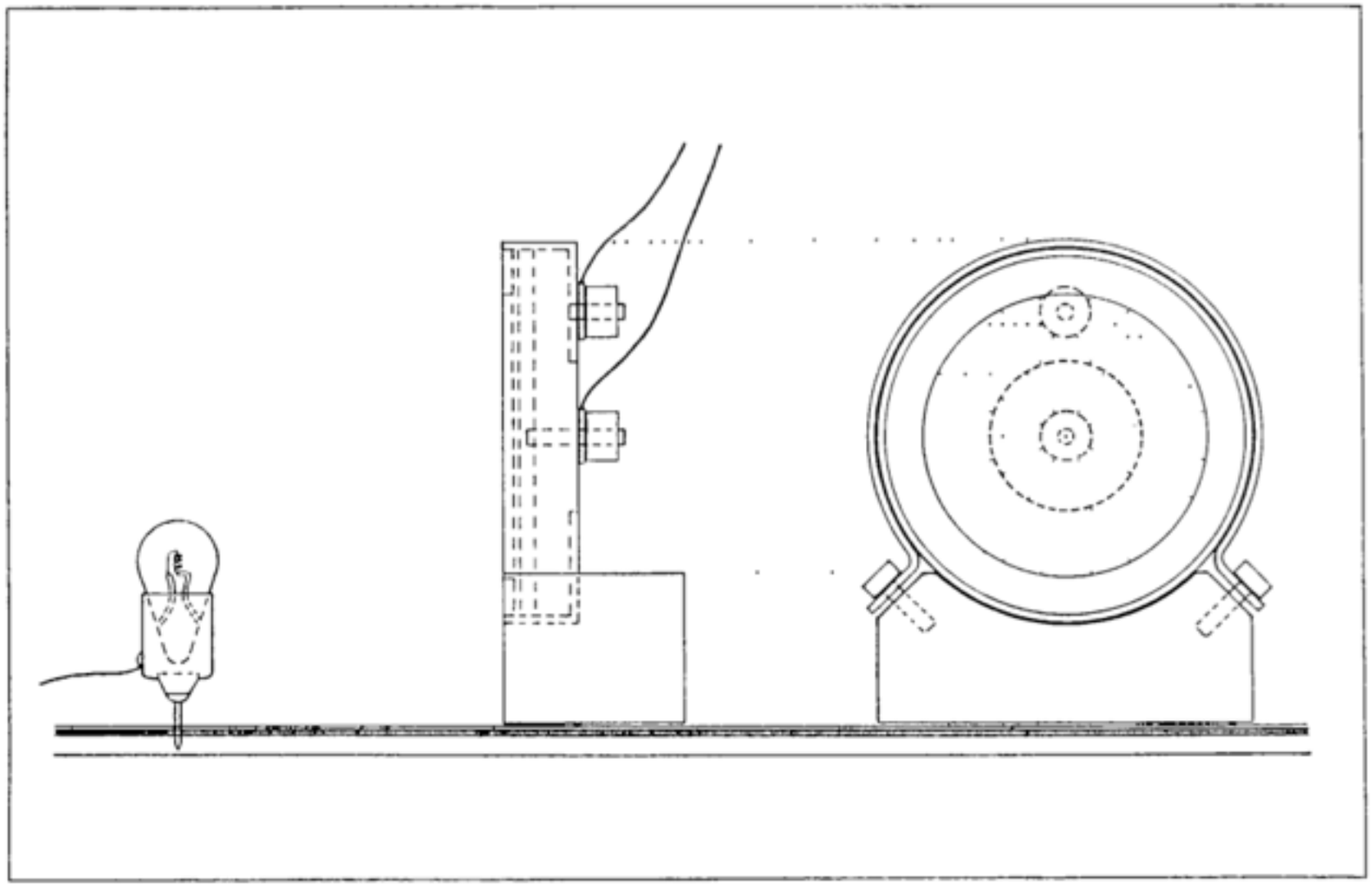


FIG. 1.—Cross-section of light-bulb and photocell (half-size)

Holmberg 1941

History, Part 0

Theorists have always been hackers, even before there were computers to hack with

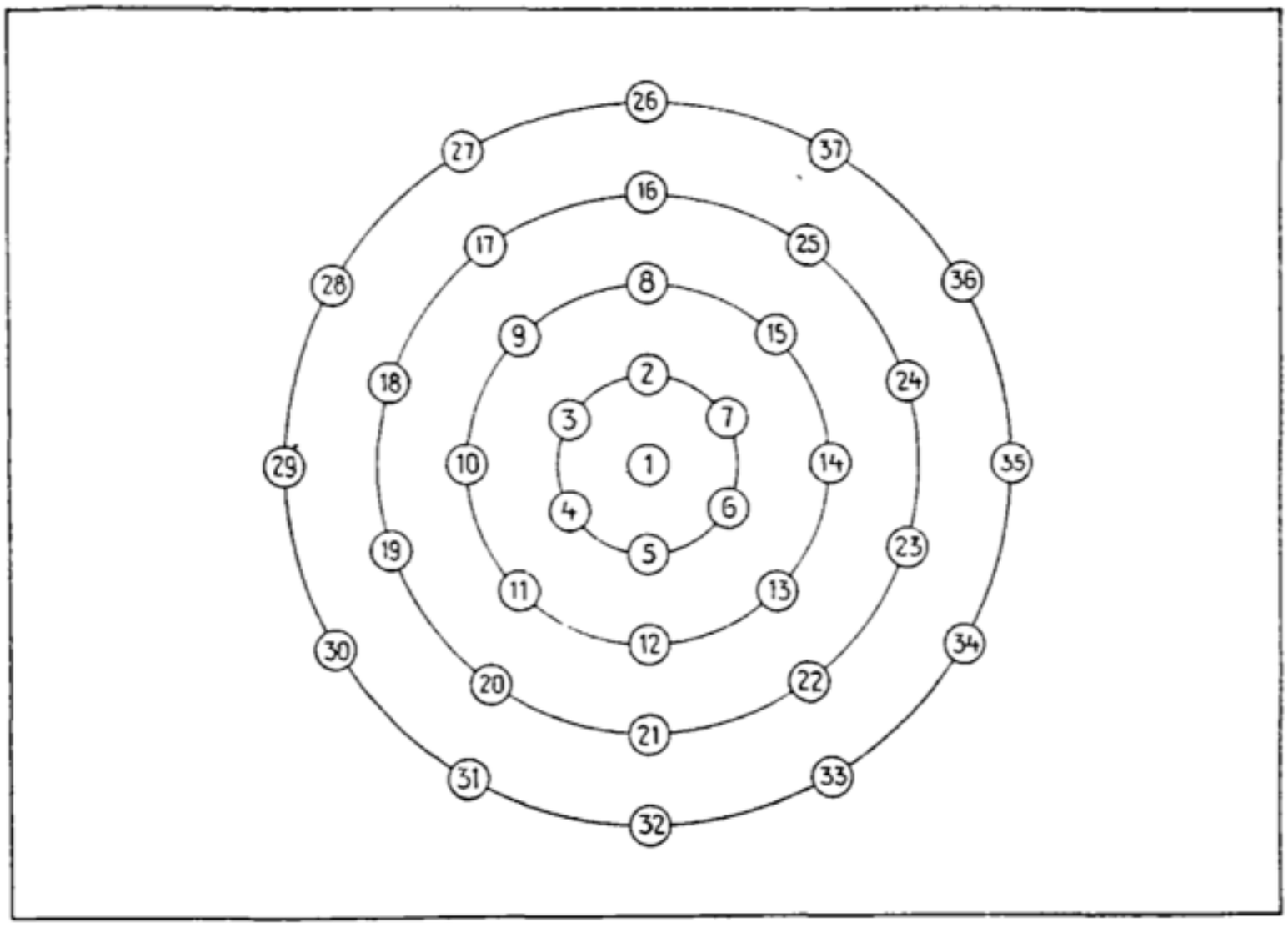


FIG. 3

Holmberg 1941

History, Part 0

Theorists have always been hackers, even before there were computers to hack with

Holmberg
1941

History, Part 1 (1980's)

CCDs arrive, turning observational astronomy and data processing into a digital and computerized undertaking

STARLINK and IRAF projects establish the model of observational astronomer as both programmer and software end-user

Important algorithmic standards are set (e.g., "optimal extraction" of Horne 1986, also Hewett et al. 1985)

Our astro-data-hacker culture of today dates to this period originally, and is a natural evolution from it.

History, Part 2 (ca. 2000)

SDSS: digital astronomy on a massive scale

Robust, scalable, automated, and statistically well-characterized spectroscopic data reduction

Scientific impact of massive, homogeneous, high-quality digital spectroscopic data sets released publicly

Establishes precedent for coupling powerful hardware and software for survey spectroscopy

Survey science requires automated inference!

History, Part 3 (ca. now!)

As evident from this conference, we are at the leading edge of an explosion of digital survey spectroscopy

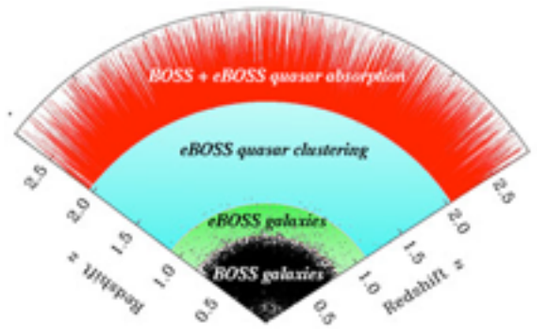
We all agree that we need software

Realizing the scientific promise of this epoch will require an evolution/revolution in sophistication of data analysis to match ambitious instruments and surveys, specifically in regard to:

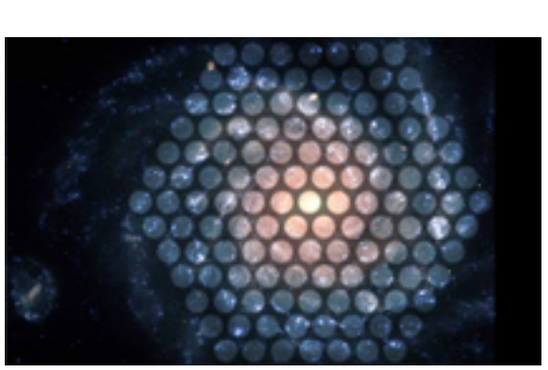
- *Low-level modeling and understanding of raw data, calibrations, and experimental characteristics, and*
- *High-level methods that connect the results of data analysis to physical theory and phenomenology*

Major challenges in the coming decade

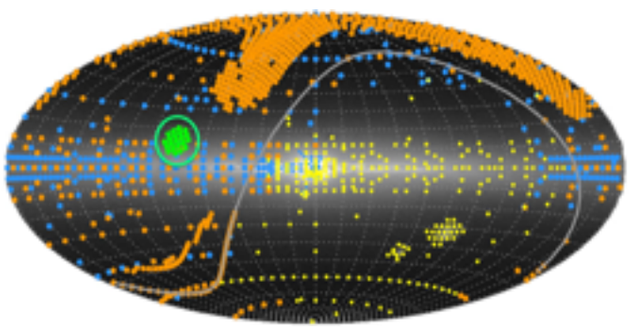
- **Distant galaxies at large N and low S/N**



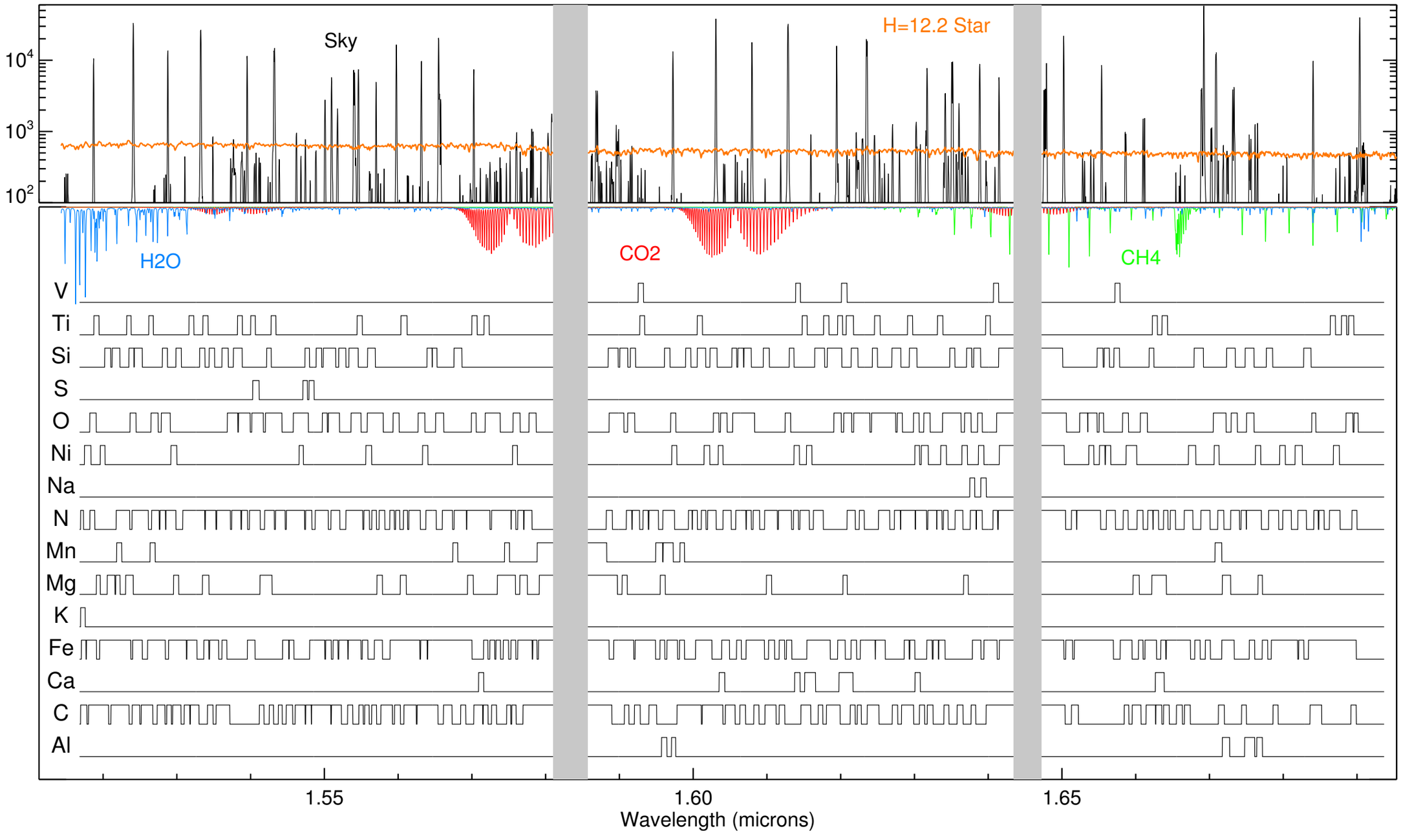
- **Nearby galaxies at large N in 3-D**



- **Stars at large N, high S/N, and N-D**

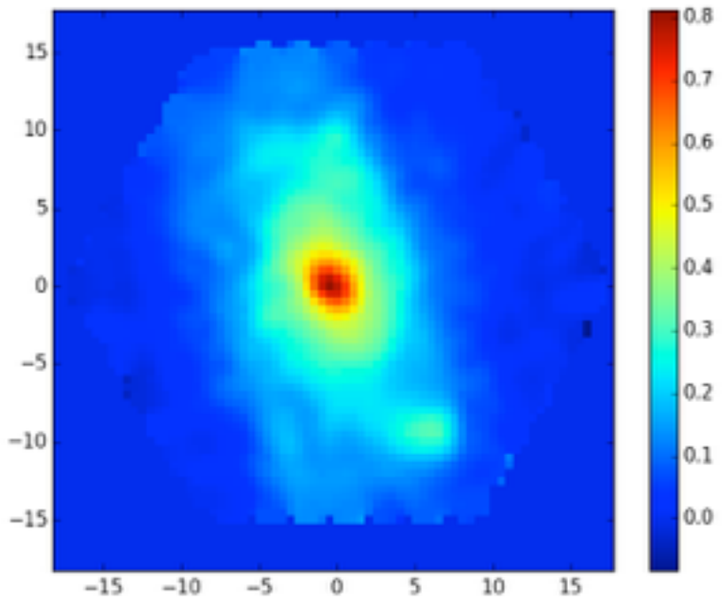


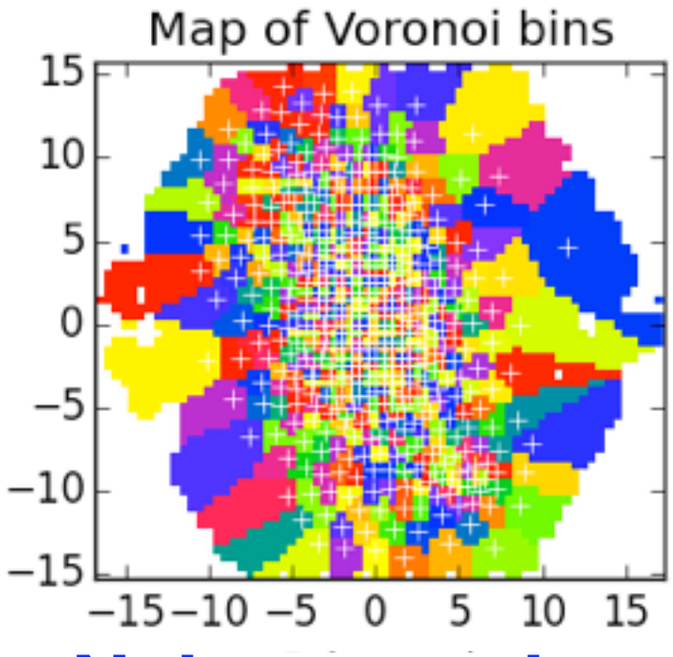
Many-D stellar abundance inference



APOGEE Figure courtesy of David Nidever, from Garcia-Perez et al., in prep.

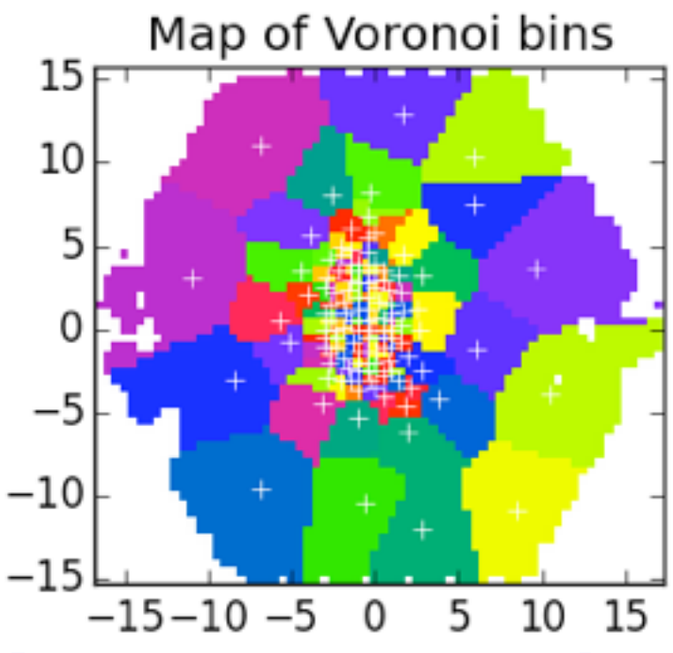
Automated inference in galaxy datacubes





Naive covariance

MaNGA Figures from
Kyle Westfall



Accurate covariance

Low-level: 2D PSF extraction

- Bolton & Schlegel 2010
- Accurate extraction for unbiased sky subtraction
- Lossless compression of spectrum likelihood functional
- Requires extremely accurate and stable calibration!
- Very difficult in BOSS; under development for DESI

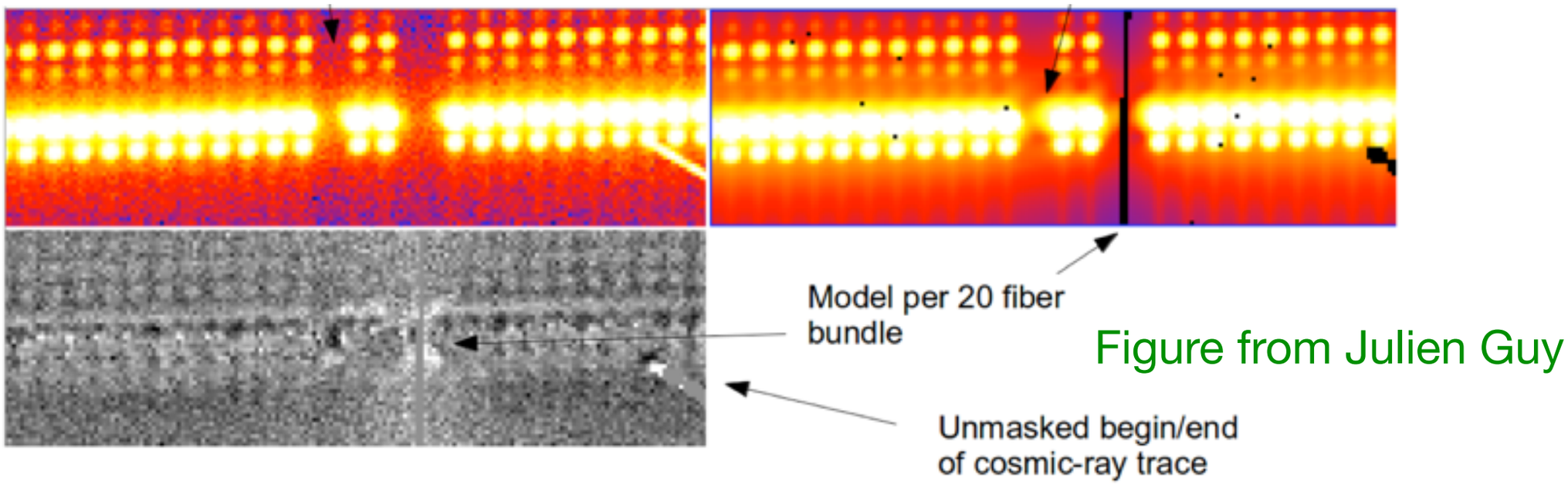
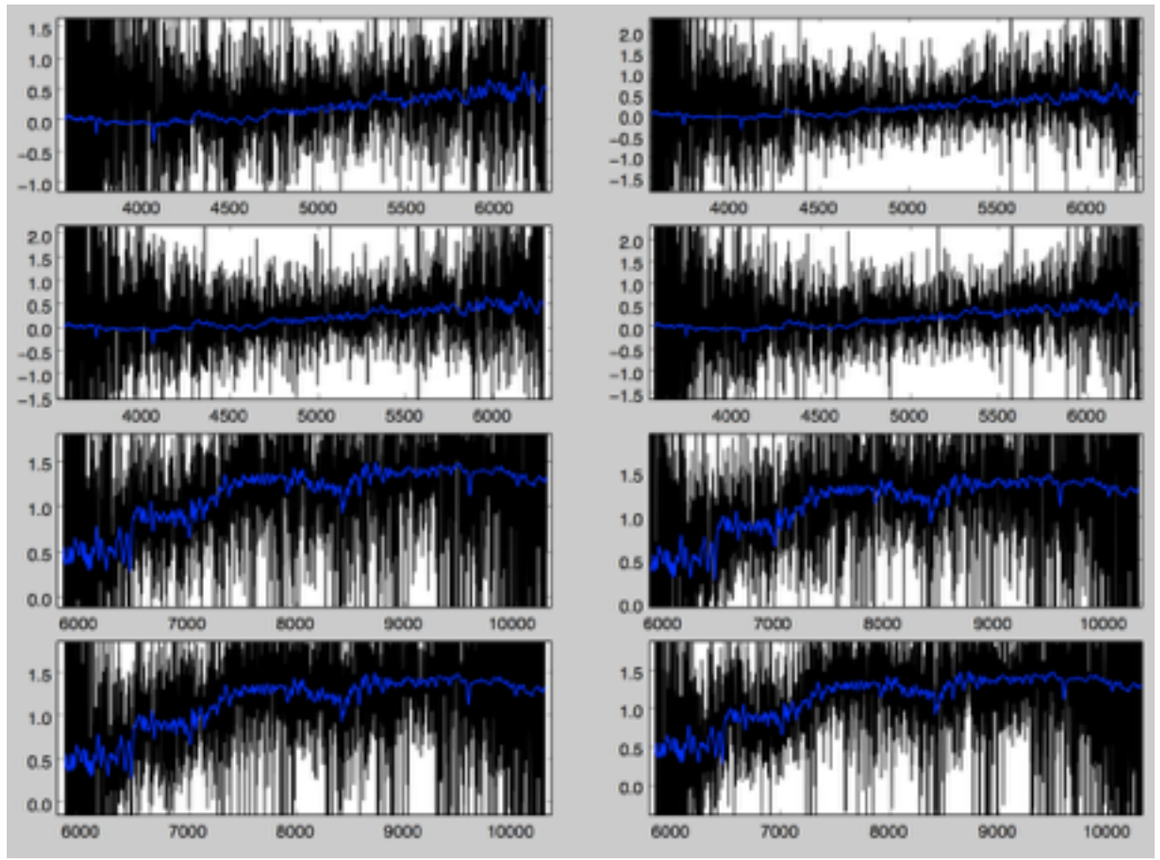


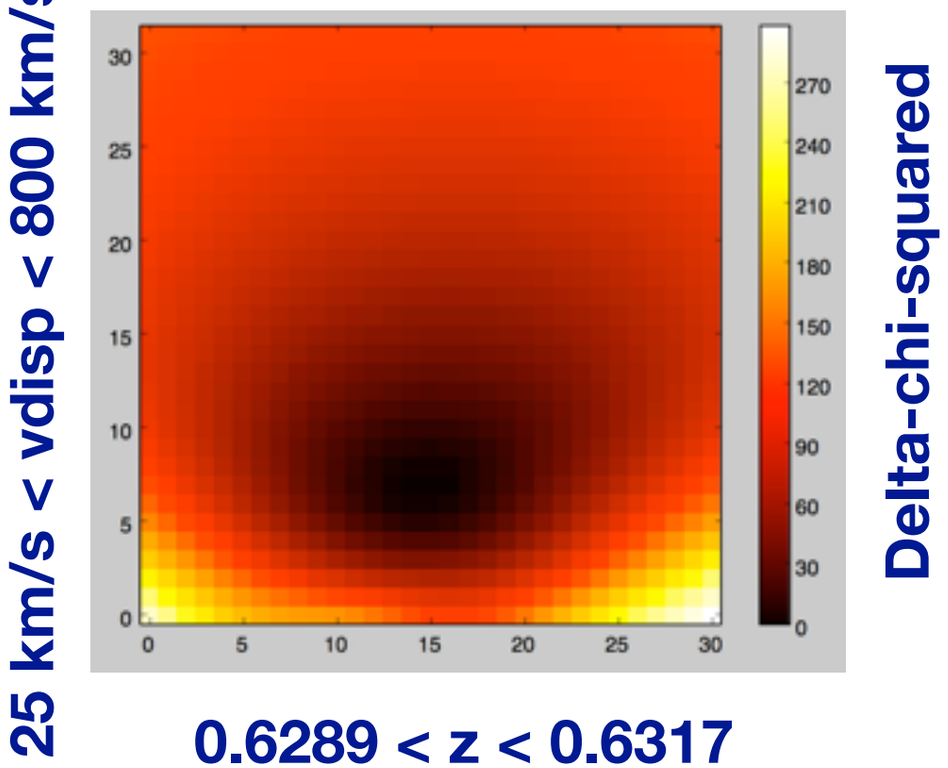
Figure from Julien Guy

Figure 3: Comparison between an arc lamp image data and the model (camera r1).

Low-level: take redshift models to data

- Hutchinson, ASB, et al. in prep ("redmonster")
- Fit multiple exposures & arms simultaneously
- Project templates through λ-dependent resolution
- Evaluate & compute χ^2 with native (unrebinned) pixels

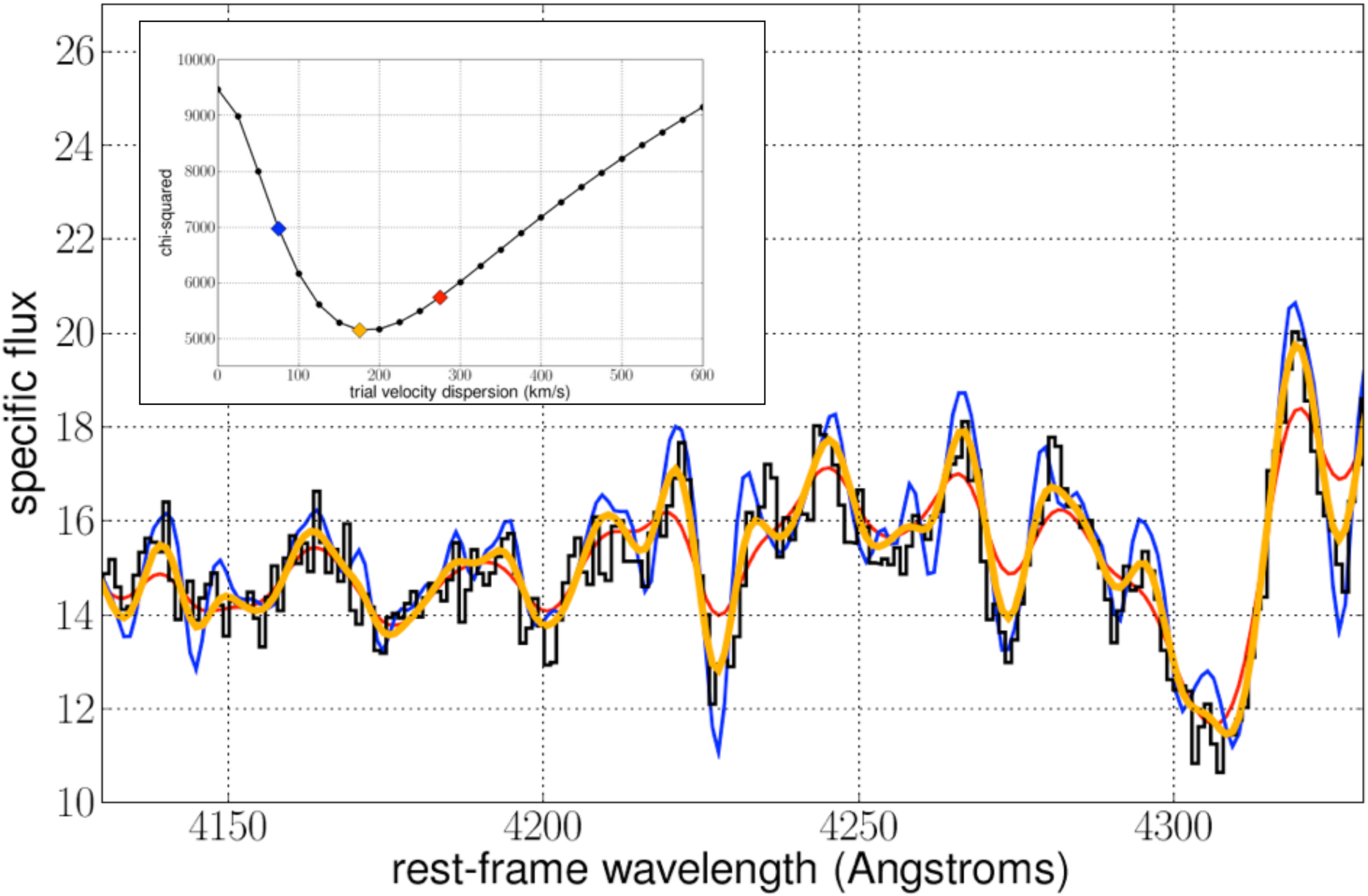




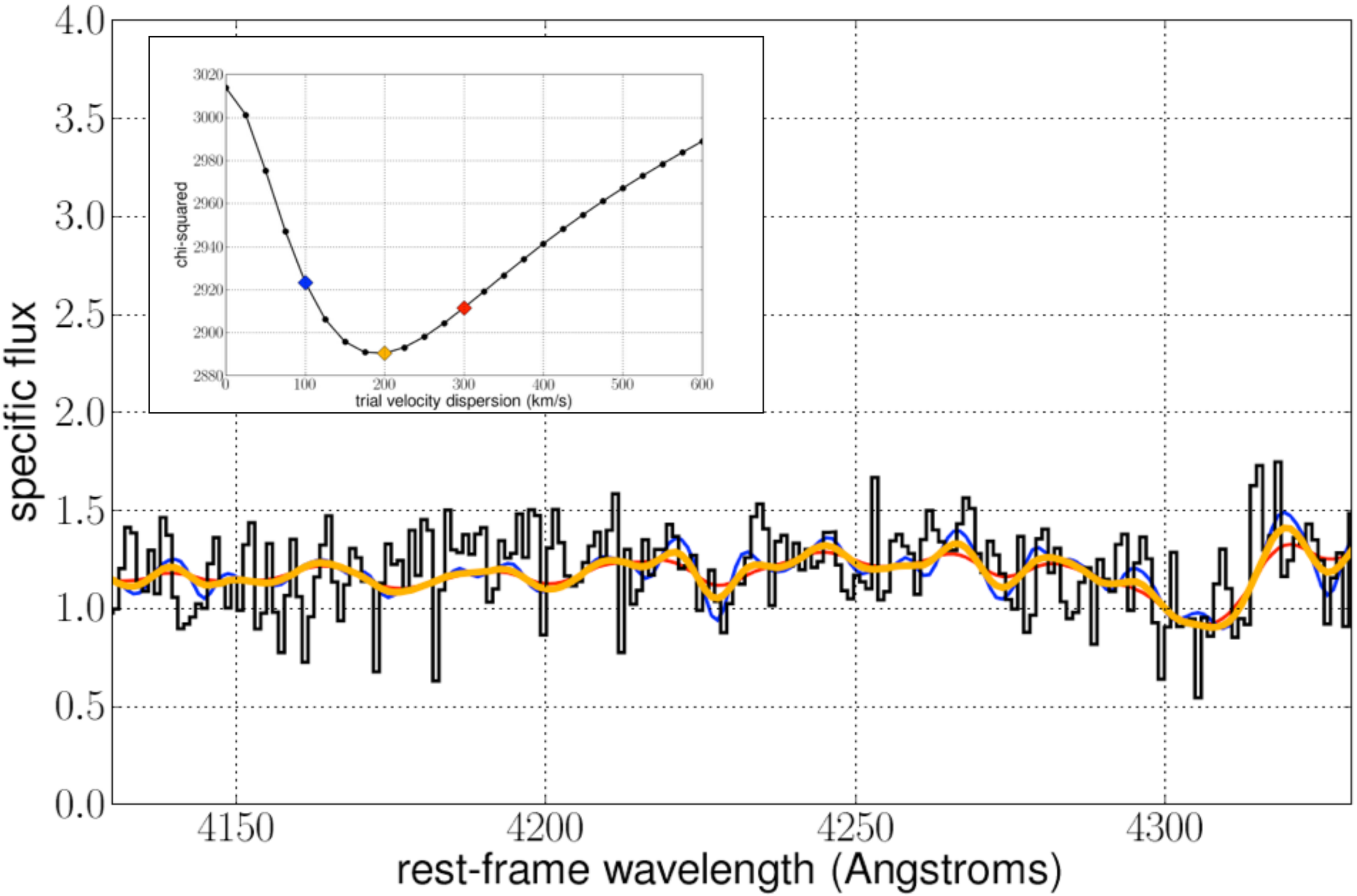
Galaxy evolution in dark-energy surveys

- Cosmological surveys like BOSS obtain >1 million spectra of massive galaxies
- The catch: spectroscopic signal-to-noise (S/N) ratio lower than what is typically "required" in astronomy
- Information content is nevertheless huge:
  - 1 million spectra with S/N ~ 3 comparable to
  - 100,000 spectra with S/N ~ 10
- Large-N / low-S/N is in fact *more efficient* for learning about the population
- Requires new methods for measuring distribution of galaxies in physical parameter space: $\Phi(\mathbf{p}|\mathbf{z};\theta)$

Velocity dispersion at “high” S/N



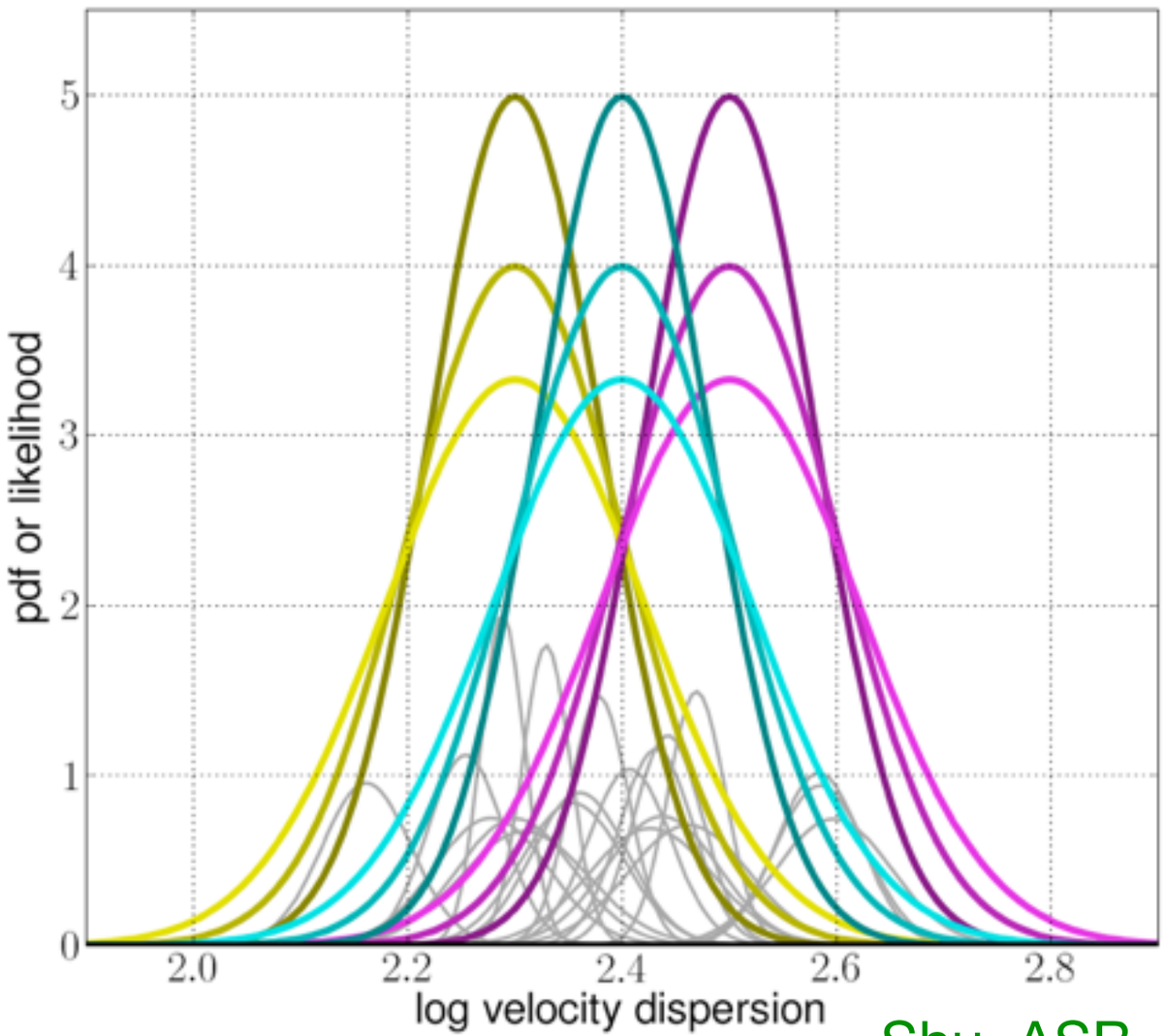
Velocity dispersion at low S/N



Hierarchical population modeling

When point estimates are just not good enough...

Parameterize distribution, marginalize over physical parameters, and optimize population *hyperparameters*



***m* = population mean:**

s = population scatter:

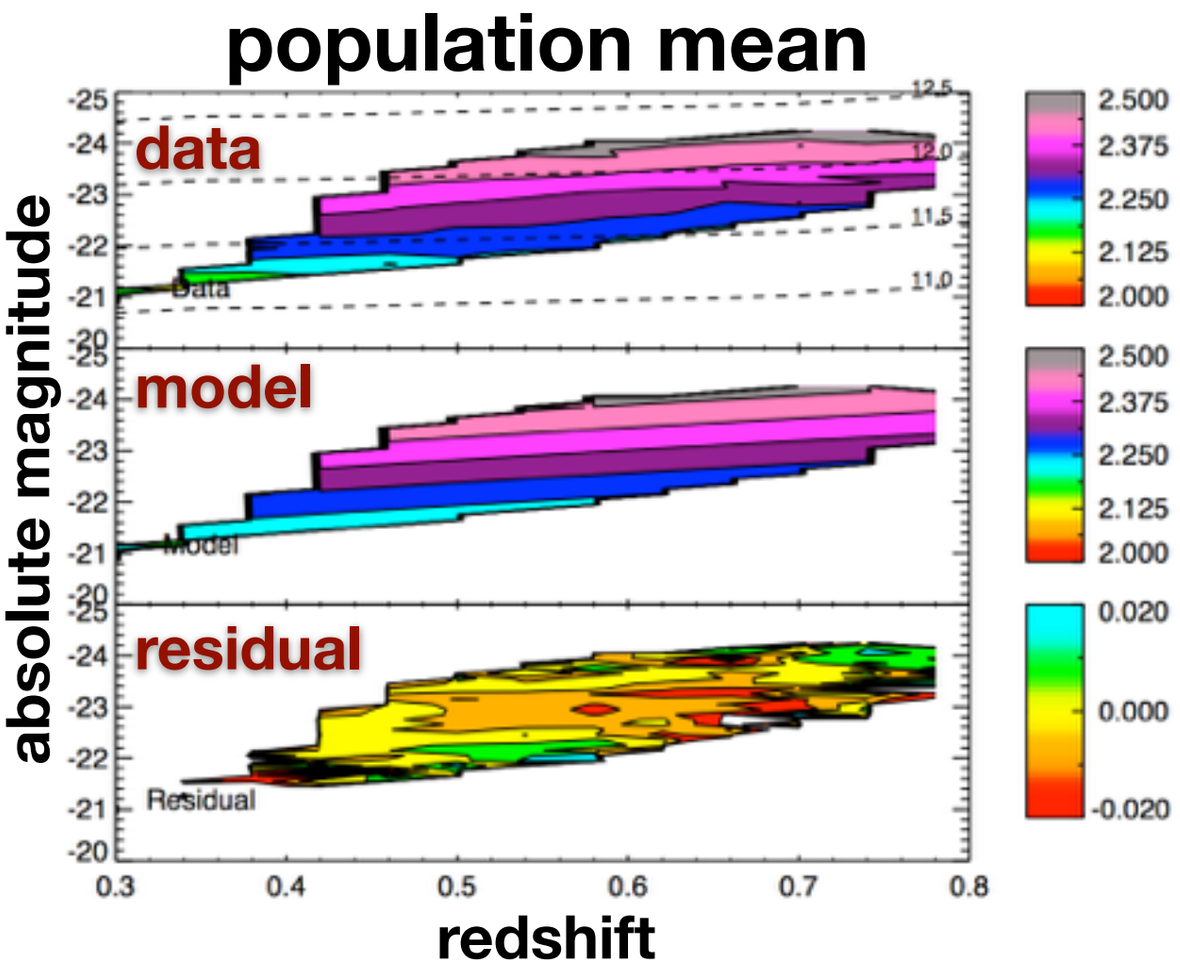
$$\mathcal{L}(m, s|\{\mathbf{d}_i\}) = p(\{\mathbf{d}_i\}|m, s) = \prod_i p(\mathbf{d}_i|m, s) = \prod_i \int d\sigma\, p(\mathbf{d}_i|\sigma)\, p(\sigma|m, s)$$

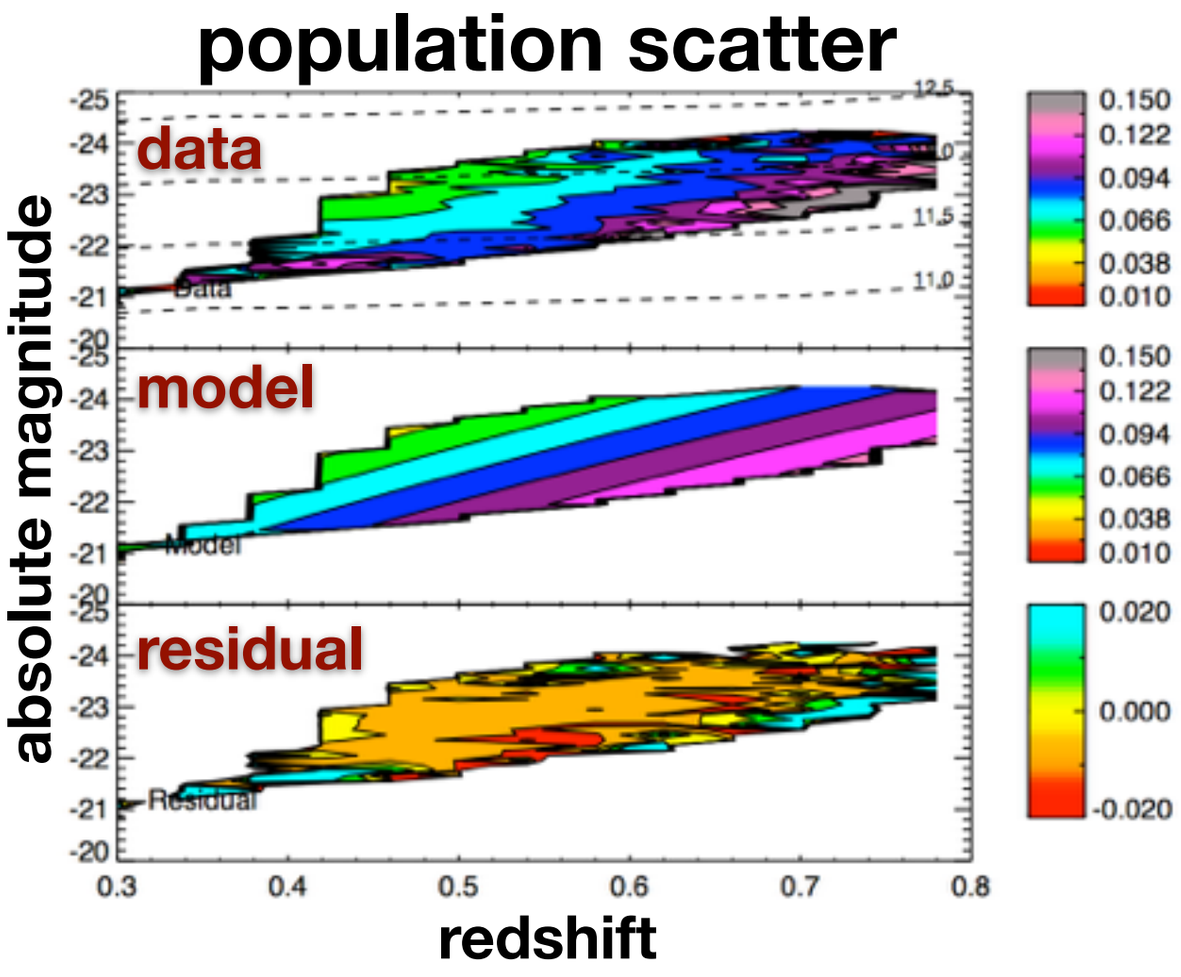
Shu, ASB, et al. 2012

Hierarchical population modeling

It works!

Shu, ASB, et al. 2012



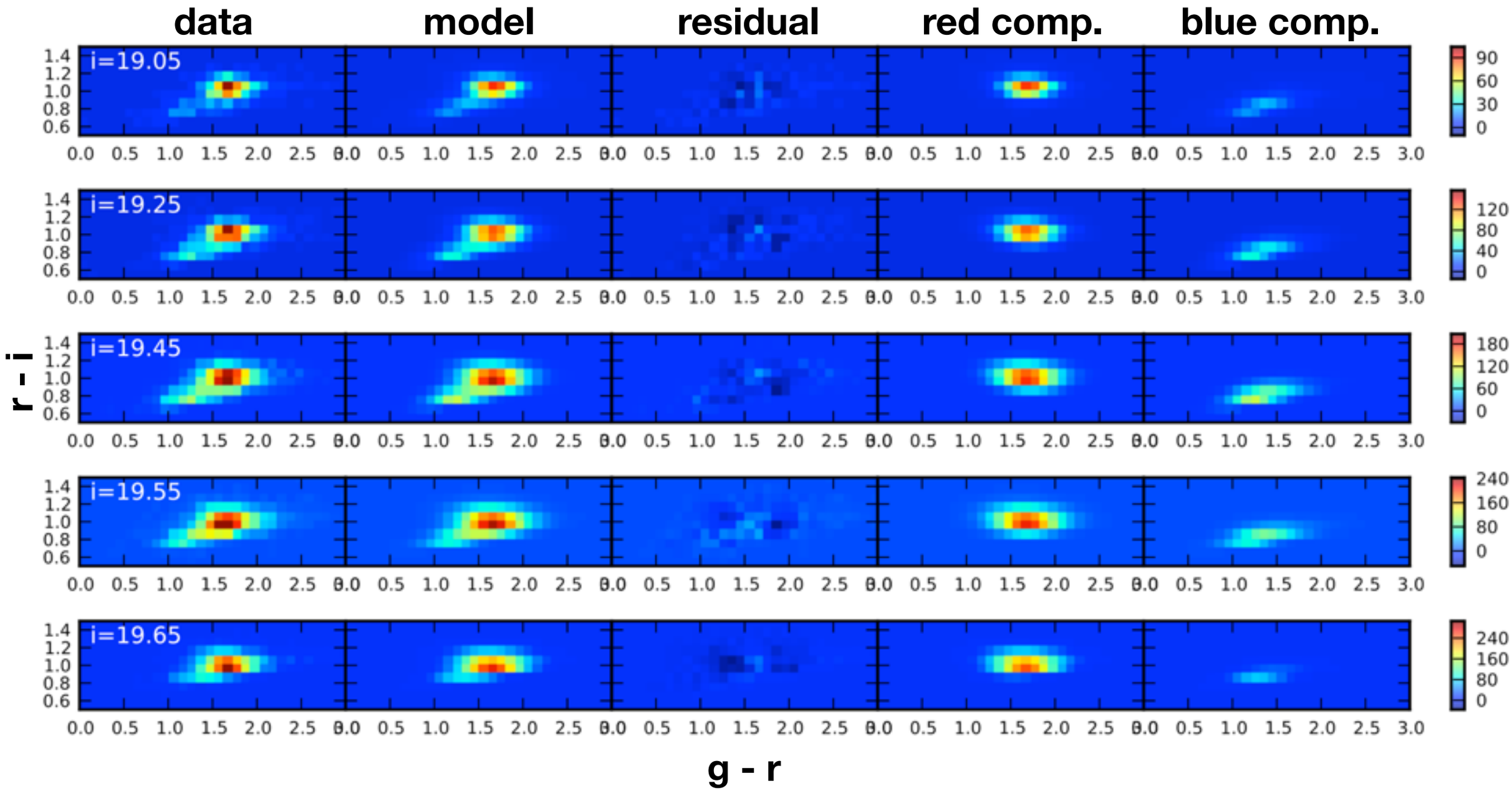


However, interpretation requires full modeling and understanding of survey selection function...

Forward modeling of luminosity f'n

Data & models in redshift slice 0.545 < z < 0.555

Montero Dorta, ASB, Brownstein, et al., submitted



Mapping statistics onto astronomy

Can build joint pdf of velocity dispersion and luminosity from independent analyses of the two quantities:

$$p(\sigma, L)\, d\sigma\, dL = p(\sigma \mid L)\, d\sigma\, p(L)\, dL$$

joint pdf

"Faber-Jackson relation": a conditional pdf

Luminosity function: a marginal pdf

Extended across multiple observational samples and multiple physical parameters for full

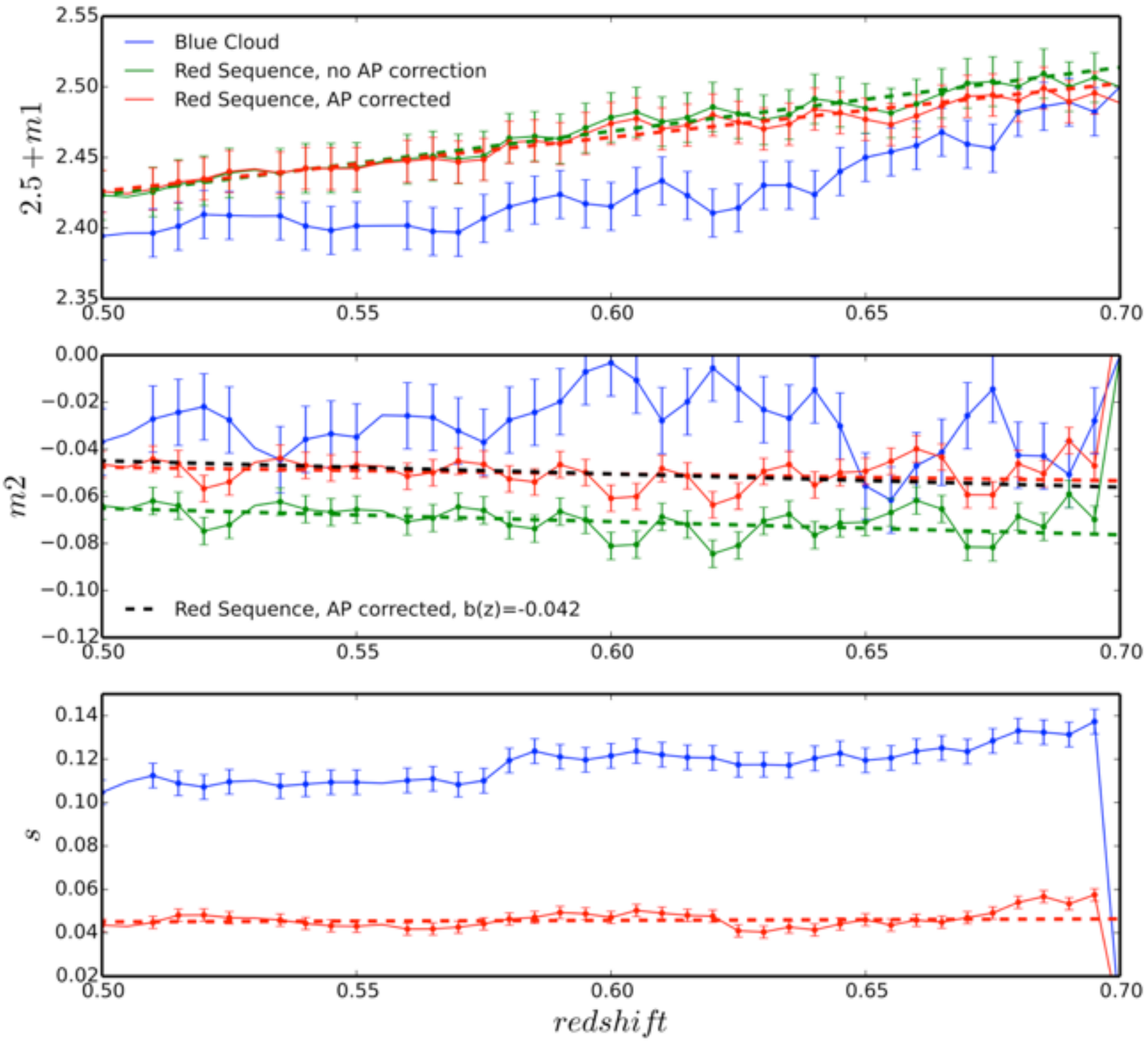
$$\Phi(p|z;\theta)$$

p(σ | L) for massive galaxies from BOSS

Zeropoint

Slope

Scatter



Montero-Dorta, Shu, ASB, et al. in prep

Closing thoughts

There has been significant cultural change even over the past 10+ years that I've been paying attention

Pipeline papers are now regularly published, especially by young people, to good professional effect

Can a "software instrumentalist" career path exist as well as a "hardware instrumentalist" path?

We must be driven by science, not by computation and statistics for their own sake. (Our goal is to learn about the universe and what's in it!)

Be scientists who code, not coders who do science: we have a lot of difficult problems to solve!