Data-driven models in the era of *Gaia*

David W. Hogg (NYU) (Flatiron) (MPIA),
and Lauren Anderson (Flatiron), Keith Hawkins (Columbia), Boris Leistedt (NYU),
Melissa Ness (MPIA), Hans-Walter Rix (MPIA)

Thank you, *Gaia*

- **Thank you** for the early data release (DR1) and steady data releases.
- Impact will be huge (it already is).
- We recognize and appreciate how much work these early releases are.
    - (But can we also get trial data to, say, train new models? *cf*. Steinmetz)

Gaia Sprints

- Hack for **one intense week** on the project of your choosing.
- Enforced policy of openness.
- Already produced 12 refereed papers!
  - (including all *Gaia* results in this talk)
- Next one is the week of **2018 June 03** in New York City.
  - We will pay travel expenses for *Gaia* team members.
  - http://gaia.lol/

(my) *Gaia* Mission

- My vision: A precise parallax for **every star of the billion**!
- But: *Gaia* parallaxes are only precise for nearby stars.
- But: *Gaia* delivers amazingly precise spectrophotometry.

(my) *Gaia* Mission

- Calibrate stellar models at close distances?
- Use those models for photometric parallaxes at all distances?
- *But:* I **don’t trust** the numerical simulations!

The astrometrist’s view of the world

- Geometry **>** Physics
- Physics **>** Numerical simulations of stars
    - (even **spectroscopic radial velocity measurements are suspect**!)

What can I contribute?

- You **don’t have to use physics** to build an accurate stellar model.
- Data **>** Numerical simulations of stars!

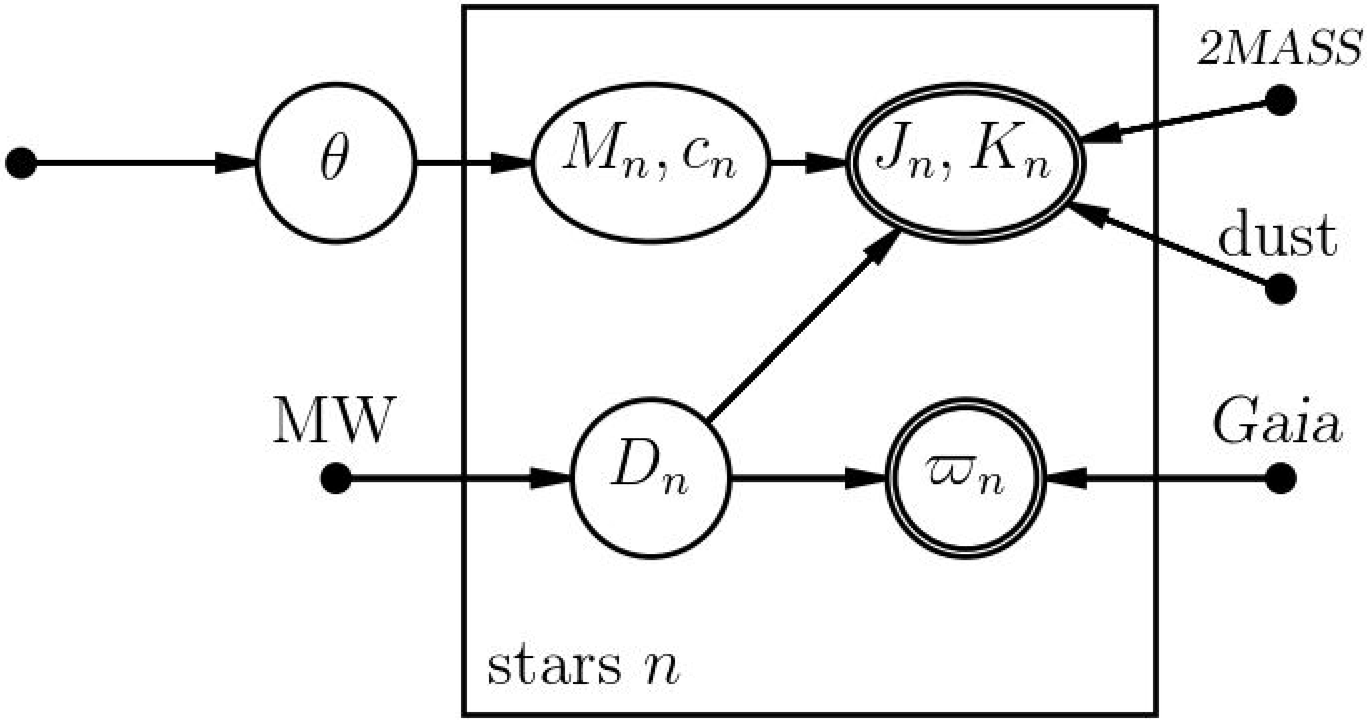
Statistical shrinkage

- If you observe a billion related objects, every object can contribute some kind of information to your beliefs about every other one.

Causal structure

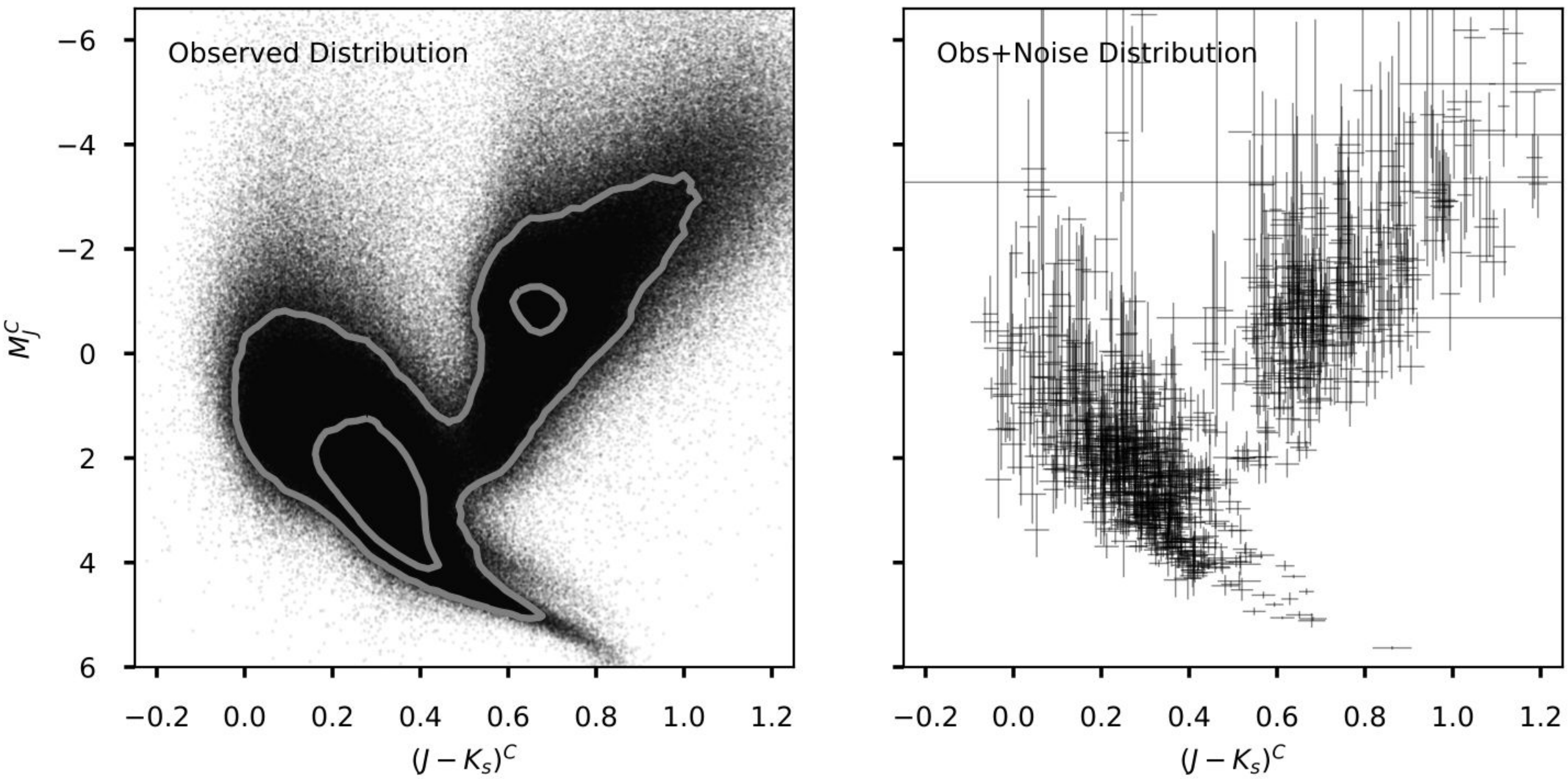
- To capitalize on shrinkage, you must impose the causal structure in which you strongly believe.
- For example: Geometry & relativity.
- For example: *Gaia* noise model.

Graphical models

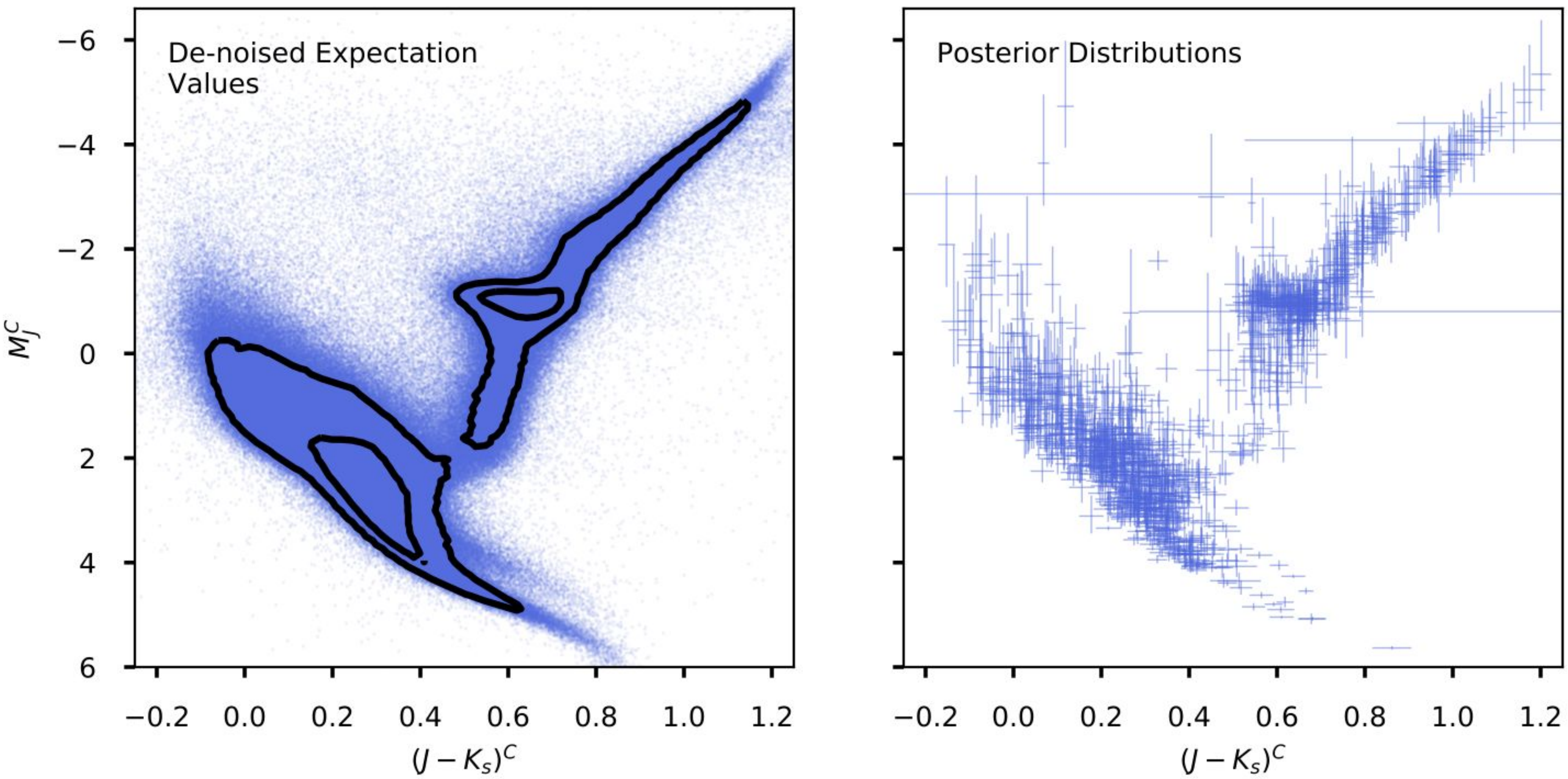


Anderson *et al* 2017 *arXiv:1706.05055*

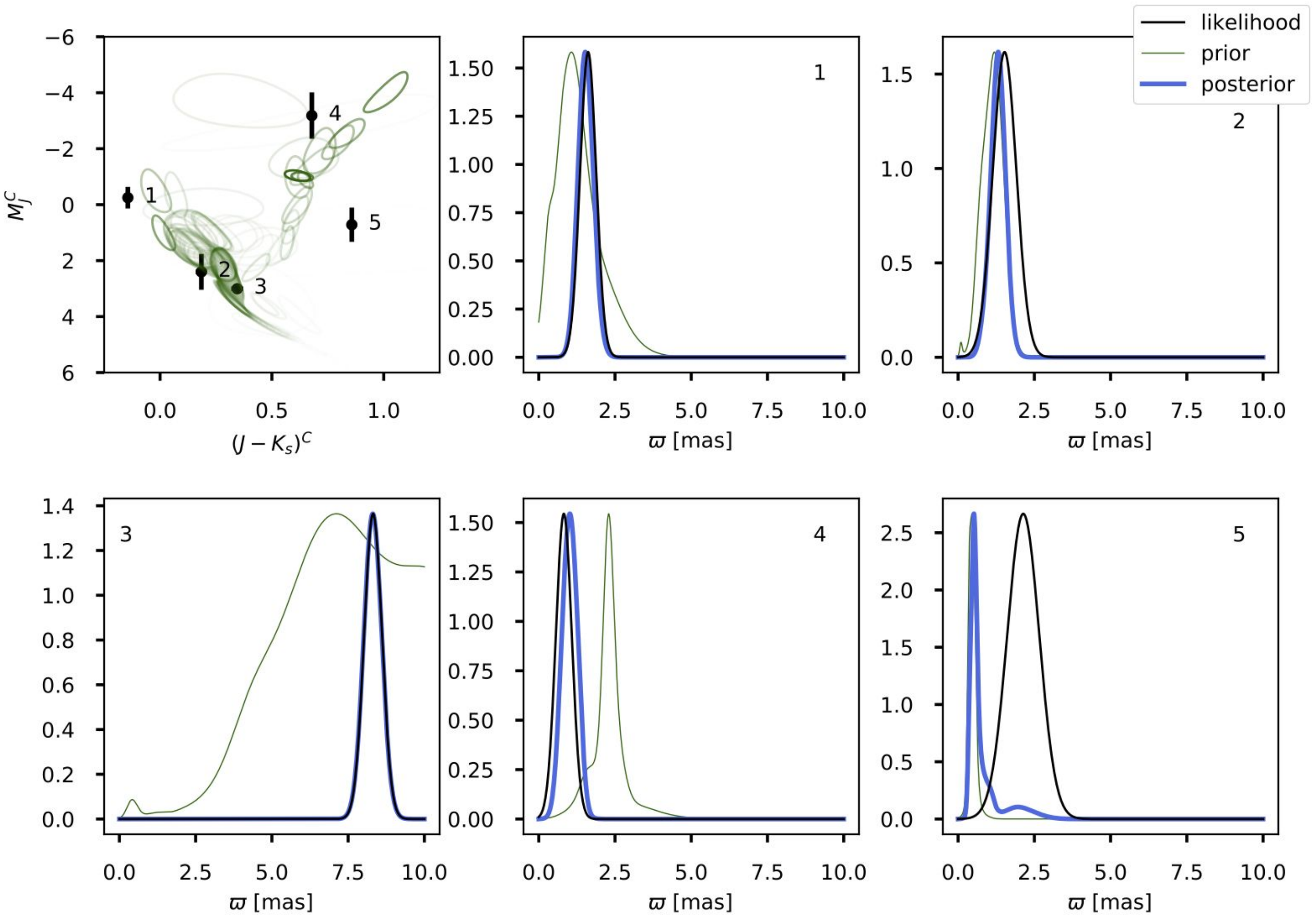
- Flexible mixture-of-Gaussian model for the **noise-deconvolved** color–magnitude diagram.
- Using *Gaia TGAS* parallax and *2MASS* photometric noise (uncertainties) responsibly.
- Using rigid dust model (from Green *et al*).
- ...Then use the CMD model to get **improved parallaxes**.



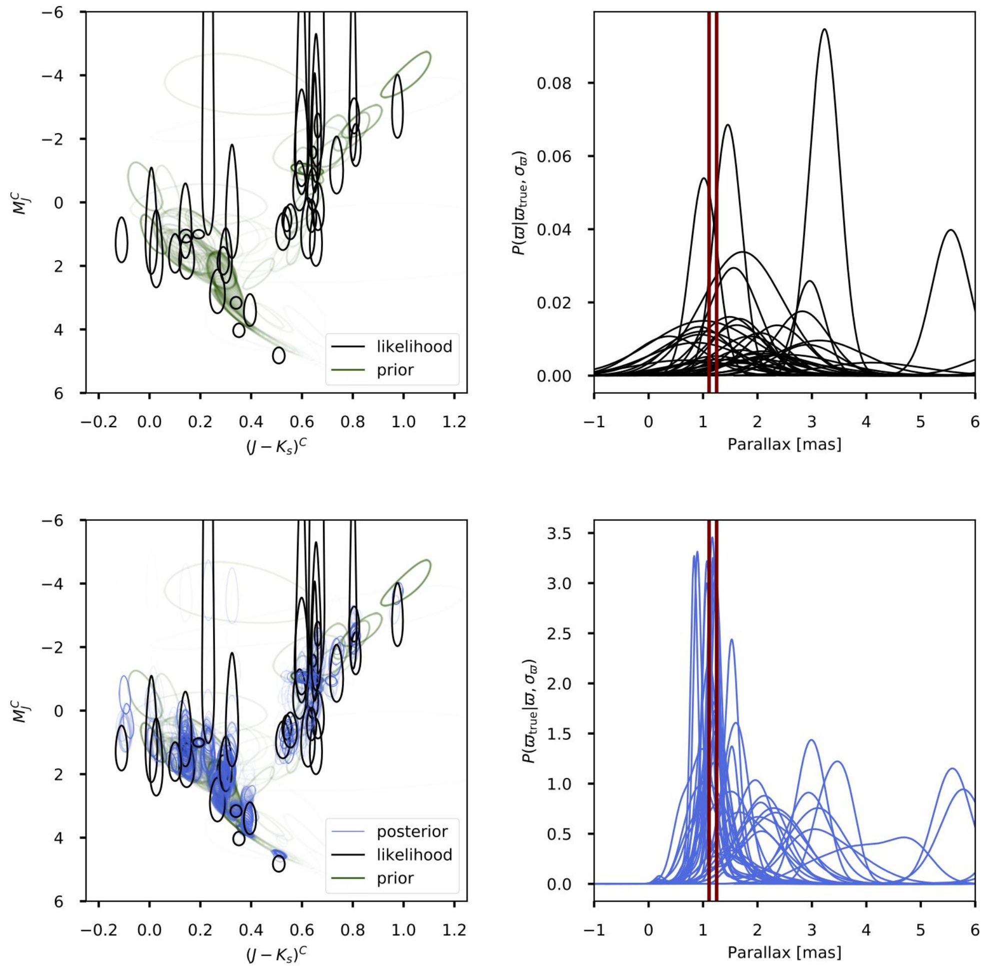
Observed Distribution
Obs+Noise Distribution
M_J^C
$(J-K_s)^C$
$(J-K_s)^C$
−6
−4
−2
0
2
4
6
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2



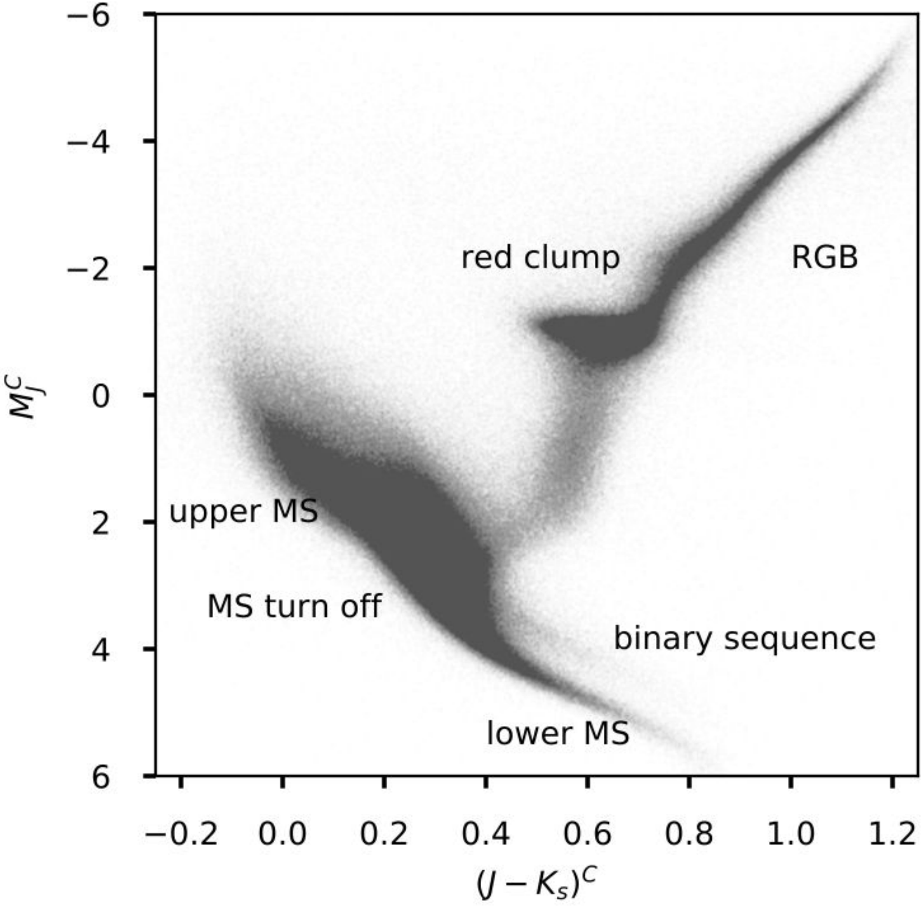
De-noised Expectation Values
Posterior Distributions
M_J^C
$(J-K_s)^C$
$(J-K_s)^C$
−6
−4
−2
0
2
4
6
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2



likelihood
prior
posterior
M_J^C
$(J-K_s)^C$
ϖ [mas]
1
2
3
4
5



M_J^C
$(J-K_s)^C$
likelihood
prior
$P(\varpi|\varpi_{\rm true}, \sigma_\varpi)$
Parallax [mas]
posterior
likelihood
prior
$P(\varpi_{\rm true}|\varpi, \sigma_\varpi)$
Parallax [mas]

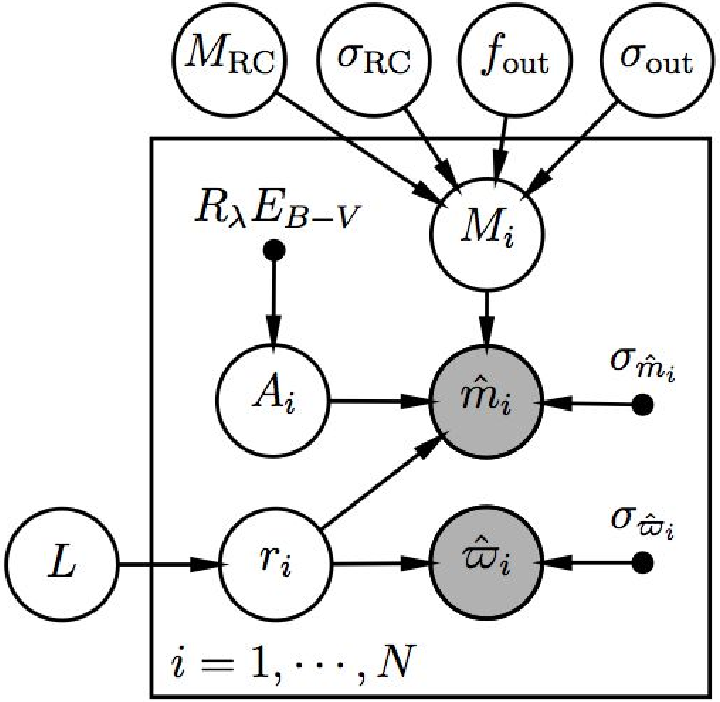


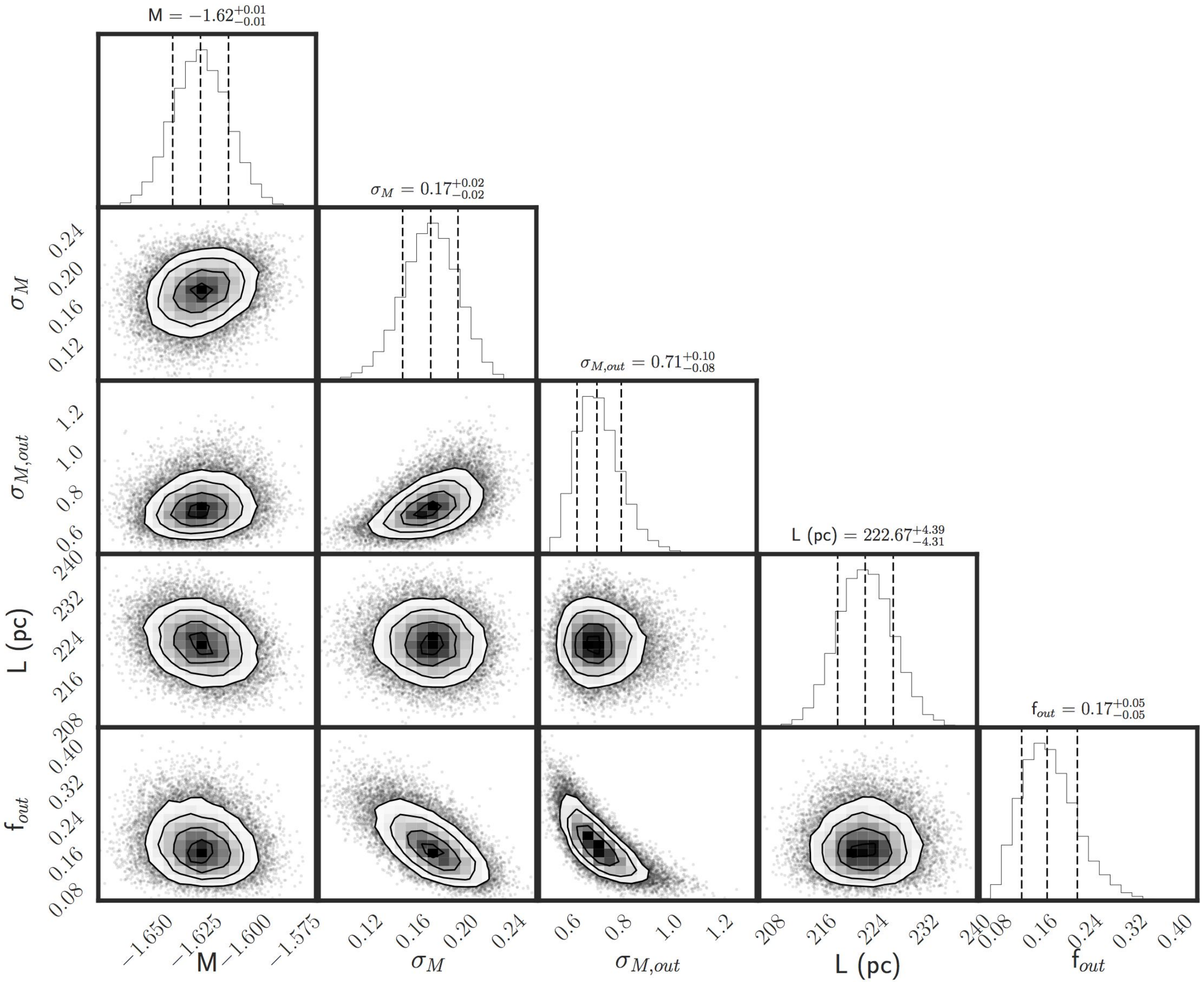
−6
−4
−2
0
2
4
6
M_J^C
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
$(J - K_s)^C$
red clump
RGB
upper MS
MS turn off
binary sequence
lower MS

Hawkins *et al* 2017 *arXiv:1705.08988*

- How precise are red-clump stars as standard candles?
- Build a mixture model for RC stars and contaminants.
- Fit for mean and dispersion of RC absolute magnitudes, taking account of the *TGAS* and photometric uncertainties.
- ...Find 0.17 mag dispersion.

Hawkins *et al* 2017 *arXiv:1705.08988*

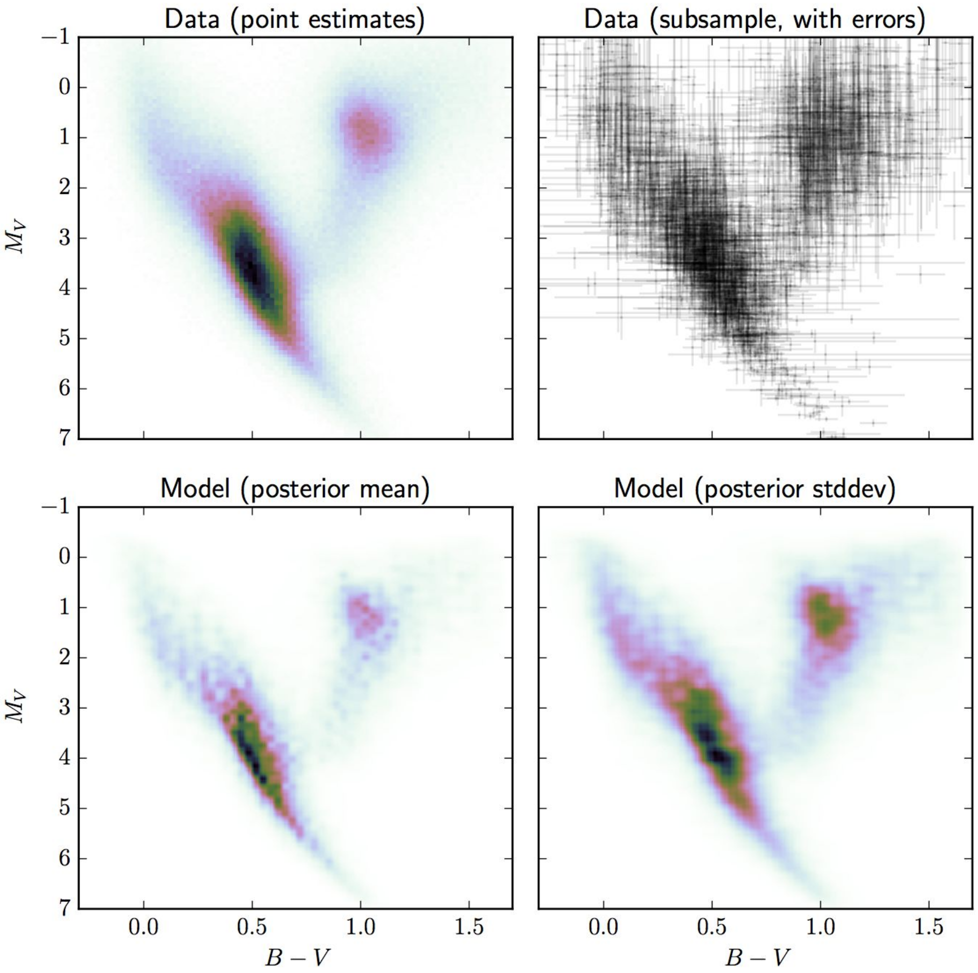




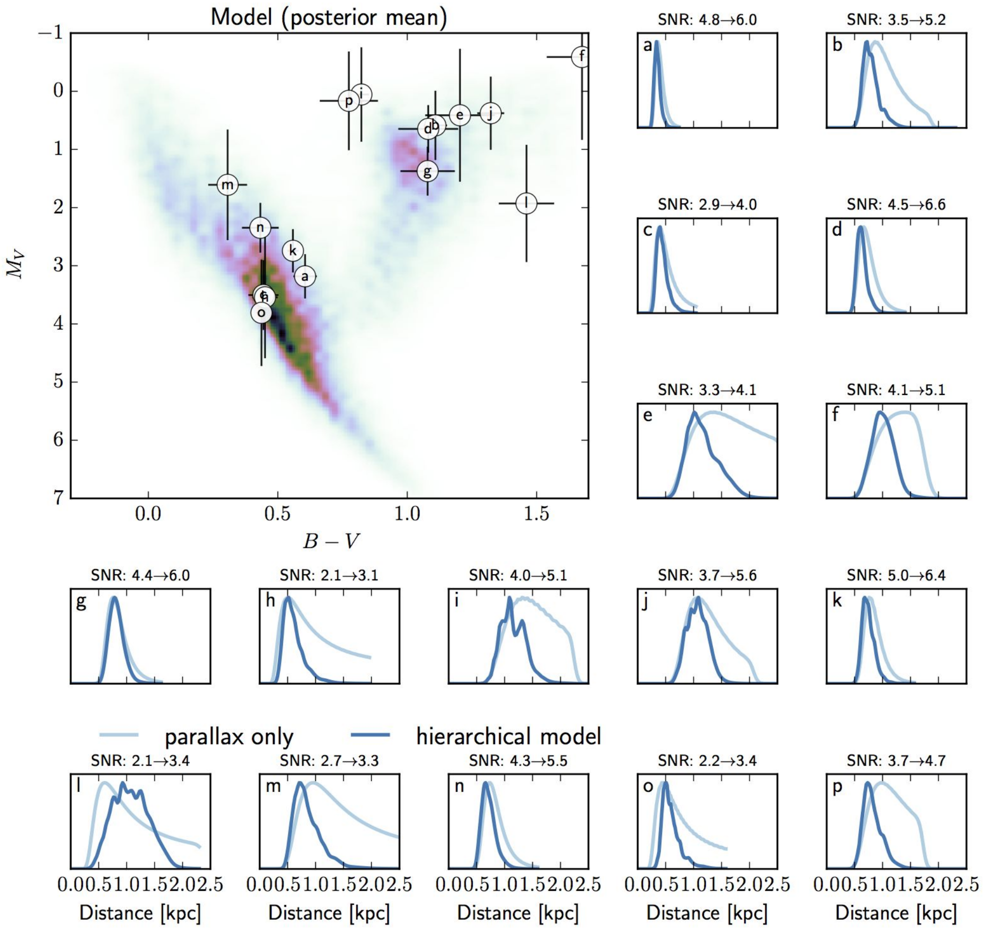
M = $-1.62^{+0.01}_{-0.01}$
$\sigma_M = 0.17^{+0.02}_{-0.02}$
$\sigma_{M,out} = 0.71^{+0.10}_{-0.08}$
L (pc) = $222.67^{+4.39}_{-4.31}$
$f_{out} = 0.17^{+0.05}_{-0.05}$
M
σ_M
$\sigma_{M,out}$
L (pc)
f_{out}
−1.650 −1.625 −1.600 −1.575
0.12 0.16 0.20 0.24
0.6 0.8 1.0 1.2
208 216 224 232 240
0.08 0.16 0.24 0.32 0.40

Leistedt *et al* 2017 *arXiv:1703.08112*

- Similar to Anderson *et al*, but fully Bayesian.
- Model is less flexible, but it is tractable as a sampling problem.
- ...Now distance posteriors are fully marginalized with respect to CMD models!



Data (point estimates)
Data (subsample, with errors)
Model (posterior mean)
Model (posterior stddev)
M_V
$B - V$
−1
0
1
2
3
4
5
6
7
0.0
0.5
1.0
1.5



Model (posterior mean)
M_V
$B-V$
SNR: 4.8→6.0
SNR: 3.5→5.2
SNR: 2.9→4.0
SNR: 4.5→6.6
SNR: 3.3→4.1
SNR: 4.1→5.1
SNR: 4.4→6.0
SNR: 2.1→3.1
SNR: 4.0→5.1
SNR: 3.7→5.6
SNR: 5.0→6.4
parallax only
hierarchical model
SNR: 2.1→3.4
SNR: 2.7→3.3
SNR: 4.3→5.5
SNR: 2.2→3.4
SNR: 3.7→4.7
Distance [kpc]

So: Just throw machine learning at the problem?

- **No!**
  - missing data.
  - heteroskedasticity.
  - generalizability.
- Every good data-driven model will be **bespoke**.

Statistical shrinkage

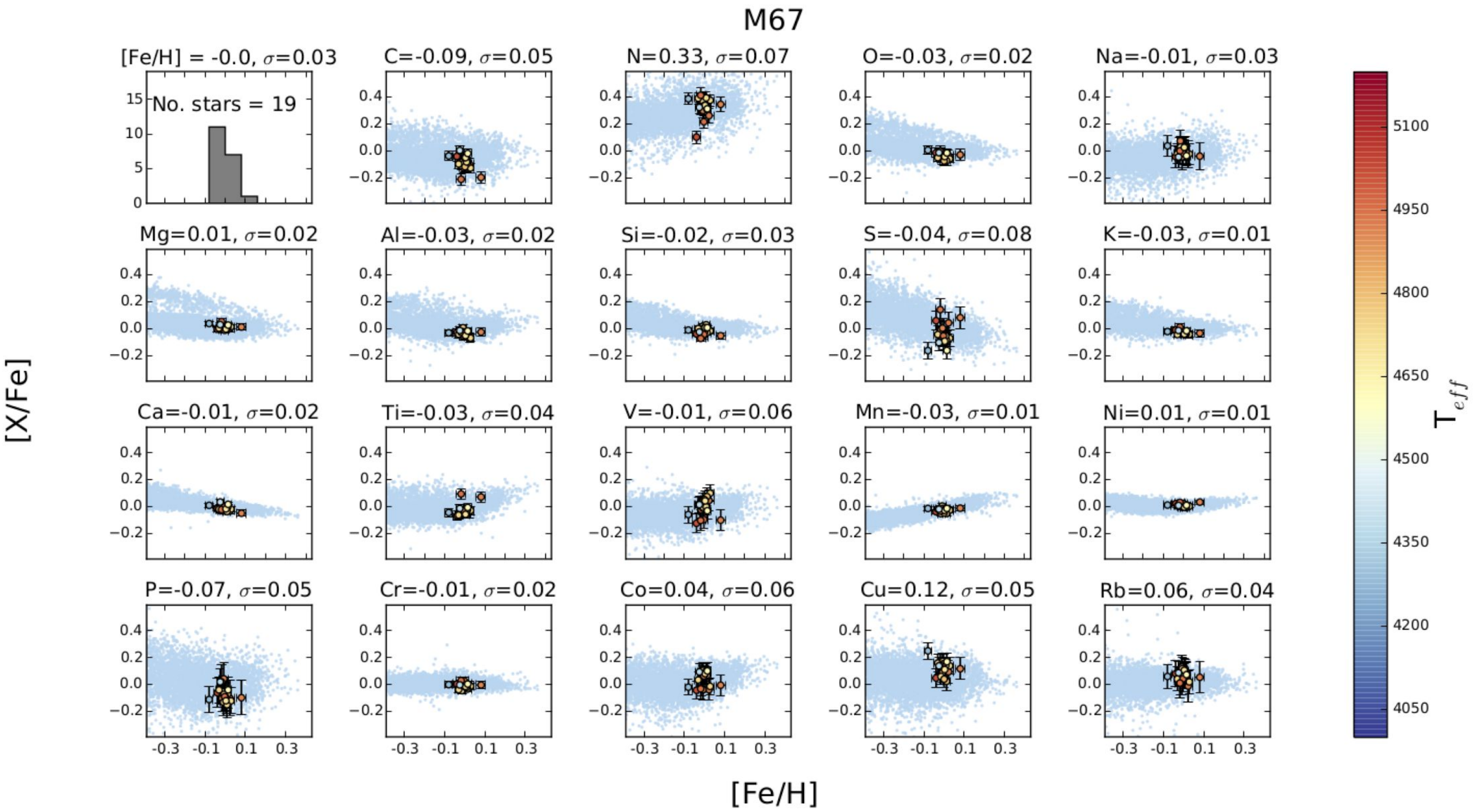
- A data-driven model can be **far more precise** than the data on which it was trained.
- (But **not more accurate**.)

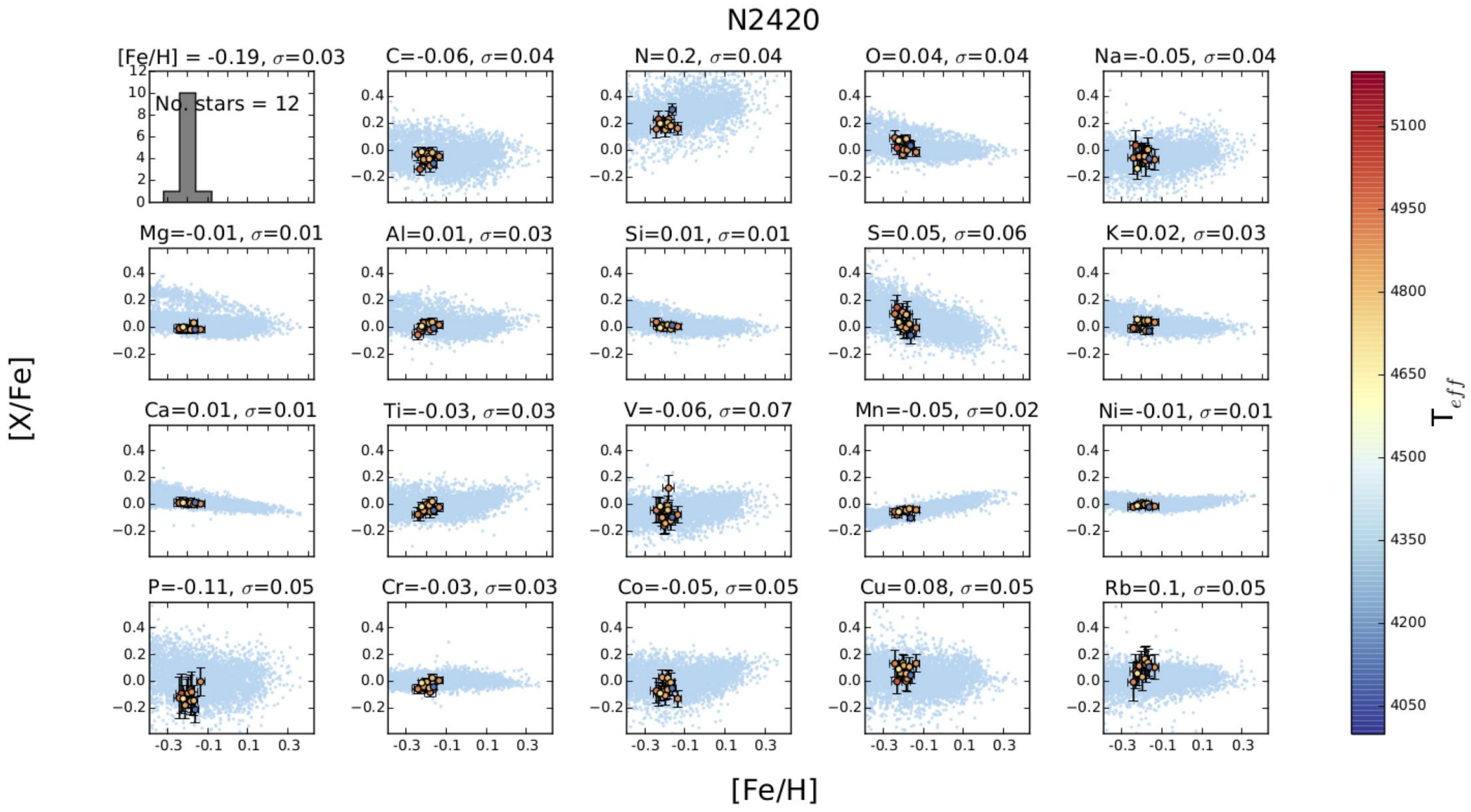
Statistical philosophy

- Pragmatism reigns.
  - Full Bayes (*eg,* Leistedt *et al*).
  - Maximum marginalized likelihood (*eg,* Anderson *et al*).
  - Maximum likelihood (*eg,* Ness *et al*).
- The important thing is the **causal structure**, not the statistical philosophy.

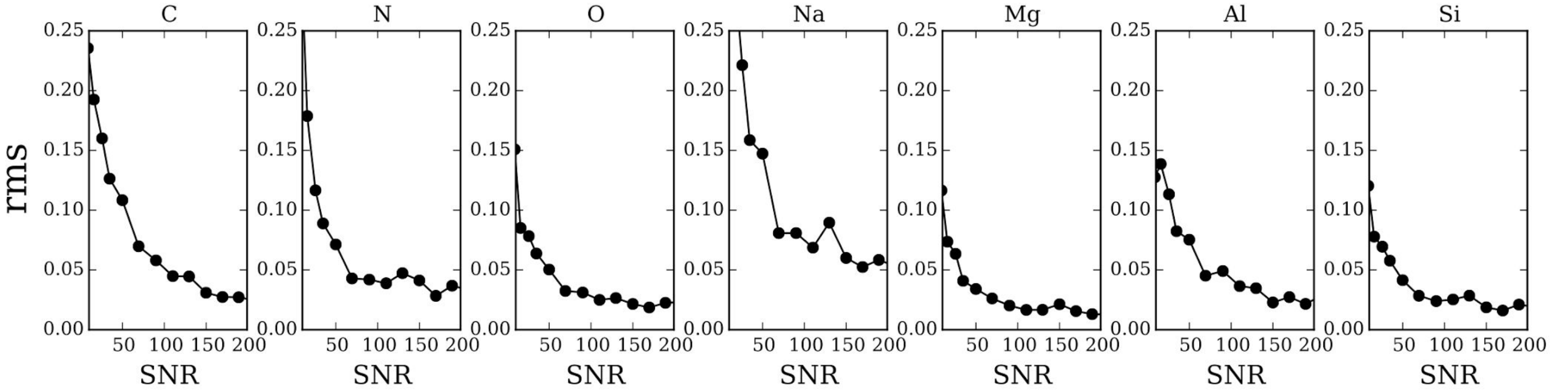
Ness *et al* 2017 *arXiv:1701.07829*

- Use high-SNR *APOGEE* spectra as training set.
- Train *The Cannon* (Ness *et al* 2015) to get detailed chemical abundances.
- Apply to low-SNR *APOGEE* spectra.
- ...Find **far more precise** chemical homogeneity among cluster stars than in the training data.
    - (also: better results at lower SNR)


M67
[Fe/H] = -0.0, σ=0.03
No. stars = 19
C=-0.09, σ=0.05
N=0.33, σ=0.07
O=-0.03, σ=0.02
Na=-0.01, σ=0.03
Mg=0.01, σ=0.02
Al=-0.03, σ=0.02
Si=-0.02, σ=0.03
S=-0.04, σ=0.08
K=-0.03, σ=0.01
Ca=-0.01, σ=0.02
Ti=-0.03, σ=0.04
V=-0.01, σ=0.06
Mn=-0.03, σ=0.01
Ni=0.01, σ=0.01
P=-0.07, σ=0.05
Cr=-0.01, σ=0.02
Co=0.04, σ=0.06
Cu=0.12, σ=0.05
Rb=0.06, σ=0.04
[X/Fe]
[Fe/H]
T_eff
5100
4950
4800
4650
4500
4350
4200
4050



N2420
[Fe/H] = -0.19, σ=0.03
No. stars = 12
C=-0.06, σ=0.04
N=0.2, σ=0.04
O=0.04, σ=0.04
Na=-0.05, σ=0.04
Mg=-0.01, σ=0.01
Al=0.01, σ=0.03
Si=0.01, σ=0.01
S=0.05, σ=0.06
K=0.02, σ=0.03
Ca=0.01, σ=0.01
Ti=-0.03, σ=0.03
V=-0.06, σ=0.07
Mn=-0.05, σ=0.02
Ni=-0.01, σ=0.01
P=-0.11, σ=0.05
Cr=-0.03, σ=0.03
Co=-0.05, σ=0.05
Cu=0.08, σ=0.05
Rb=0.1, σ=0.05
[X/Fe]
[Fe/H]
T_eff


C
N
O
Na
Mg
Al
Si
rms
SNR
0.00
0.05
0.10
0.15
0.20
0.25
50
100
150
200

Aside: Proper motions are like parallaxes

- Proper motions decrease with distance like parallaxes.
- With a position–velocity model for the MW, they can be combined.
    - *cf*. Floor's talk; *cf*. "reduced proper motion"
    - At large distances (and 10-year mission) we expect proper motions might dominate information.

Fundamental assumption of data-driven models

- **Stationarity**.
- *ie:* The causal structure is correct.
- *ie:* All non-trivial dependencies are represented in the graphical model.

Assumptions can be tested

- By construction, data-driven models are easy to validate.
- When the causal structure is insufficient, the failures appear in simple validations or visualizations.

Example: Halo stars are different from Disk stars

- Different distributions of metallicity -> different color–magnitude diagrams.
- Solution: Add kinematics and Galactocentric distance into the graphical model, and permit the model to discover this.

Summary

- There is no longer any reason to use numerical stellar models to generate photometric parallaxes.
- The billion-star catalog plus statistical shrinkage will deliver enormous precision (and accuracy), better than any physics models.
- Data **>** Numerical models of stars.