Is machine learning good or bad for astrophysics?

These slides are at: https://dwh.gg/a3d3

David W Hogg

NYU / MPIA / Flatiron

http://cosmo.nyu.edu/hogg/

What I'm going to say

- Negative side:
  - Current ML methods **cannot be trusted** and cannot be interpreted (by construction).
  - Their use exposes us to strong **biases** or **systematic errors**.
- Positive side:
  - ML helps with the **engineering systems** involved in astrophysics projects.
  - ML can be used on auxiliary components (nuisances), such as **calibration and backgrounds**.
  - In causal problems, flexibility is paramount (and interpretation is not).

What we’ve learned in astrophysics from ML

What we’ve learned in astrophysics from ML

[that’s it; that’s the list]

What we’ve learned in *science* from ML

“But ML solved protein folding!”

- *AlphaFold* (eg, PubMed/PMC8728224) can predict structure from sequence.
    - The main functional goal of protein folding.
- This success told us **literally nothing about how proteins fold**.
- They solved an engineering problem: *Given a sequence, what is the corresponding fold?*
    - It didn’t answer any open question in the physics, chemistry, or biology of protein folding.

Engineering and science

- My group does lots of engineering for big astrophysics projects.
    - Instrument calibration, observatory operations, data analysis pipelines, model building, optimization and inference systems, project management.
- Good engineering is an extremely important part of every project.
- *Engineering successes* are not the same as *science results*.
- Don’t get me wrong: **Great engineering makes all science possible**. I love engineering, and I do it.

What is machine learning?

- A *machine learning method* is a **method whose capability improves as it sees more data**.
  - Probably meaning: **Improves substantially faster than the square-root of *N*.**
- *Classic:* PCA, ICA, SVM, linear regression, Gaussian process, k-means, K-nearest-neighbor, KDE
- *Contemporary:* MLP, deep CNN, transformer, diffusion

What is (supervised) machine learning?

- You have a golden set of data containing N objects, each of which has a list x_i of *features* and a list y_i of *labels*. This is your **training set**.
- You try to find the function $f(x)$ that does “the best” job of predicting y in this data set. This is the **training step**.
  - You give this function immense flexibility—often literally millions or billions (!) of parameters.
- You can now predict new labels y_* for any new data point x_* with $f(x_*)$. This is sometimes called the **test step** or **prediction**.
  - Note the deep assumption that *the new data are similar to the training data*.

The uses of ML in astrophysics

- Classification
  - Which pipeline to apply to which object? Which objects to observe further?
- Outlier detection
  - Find moments when the observatory has issues; find unique objects.
- Dimensionality reduction
  - Stars and galaxies live in low-dimensional spaces!
- Regression for label transfer
  - I know the parameters of these stars, can I get parameters for 200M more stars?
- Emulation of expensive simulations
  - The Universe is hard to simulate; our carbon footprint is horrifying.

The philosophy of machine learning

- *Ontology:* **Only the data exist**; models predict *data from data*.
  - The latent structure is irrelevant; judged only on performance.
  - We don't need to understand the internals of $f(x)$.
- *Epistemology:* **Performance on held-out data** is the one arbiter of truth.
  - Compare this to the epistemology of physics!

Interdisciplinarity

- ML methods were (mostly) built by companies for commercial applications.
- They perform **incredibly well on those tasks**!
    - Have you seen TikTok recently?
- How is *presenting content to users* like or not like *doing astrophysics*?

ML vs astrophysics

- ML uses “train, validate, and test” frameworks.
  - These don’t really exist in astrophysics: **We are trying to find new things** (higher redshifts, lower masses, novel signatures of atmospheric chemistry).
- ML takes the data as given.
  - We care about experimental design, noise models, and **selection effects**.

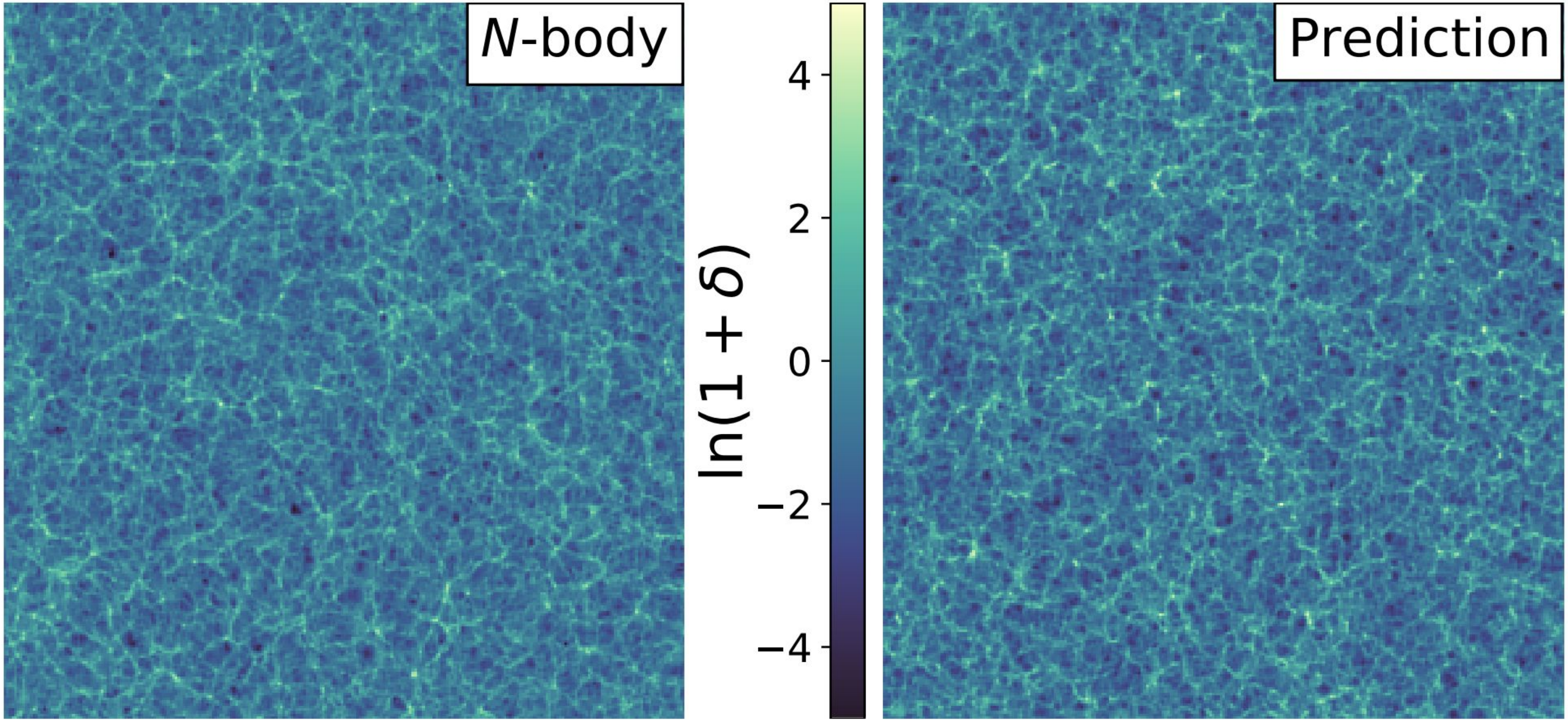
Trust issues

- Fundamentally you can’t know what an ML method is doing, internally.
    - (this is controversial; many experts would disagree)
- Interpretability is much discussed, but *is currently a failure*.
    - Even linear regression is generally uninterpretable once the number of features gets large.
    - I believe that interpretability is doomed to failure, because it is at odds with model capacity..

The question

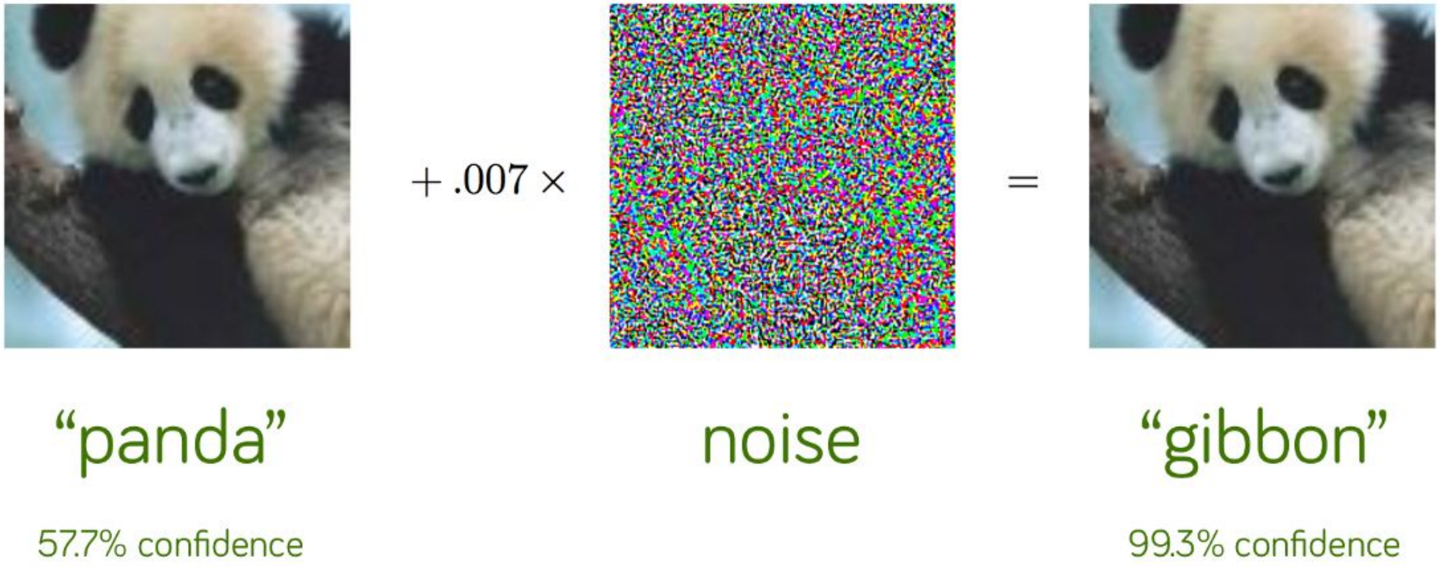
- Where in science can you use a model that you don’t understand?

Example: Emulation (Piras *et al* arXiv:2205.07898)



16

Adversarial attacks (Goodfellow *et al* ICLR 2015)



What do adversarial attacks reveal?

- They are carefully tuned, so *they don't represent generic failure modes*.
- But they reveal that **the model is not doing what we think it is doing**.
  - In scientific applications, that's pretty disturbing.

Technical point: Confirmation bias

- *Simulations are expensive, so let's replace them with an ML emulator!*
  - *Really expensive!* In cosmology and in ocean science, *eg*, the requirements **exceed the computing capacity of the United States**.
- … [grind on your scientific problem using those emulations as your theory] …
- Now you discover something really really surprising. What do you do?
  - Checking your result is very expensive (by construction), so **you will only check if the result is very surprising**.
- This is the very definition of **confirmation bias**.
  - Emulation forces us inevitably into a confirmation-bias setting.

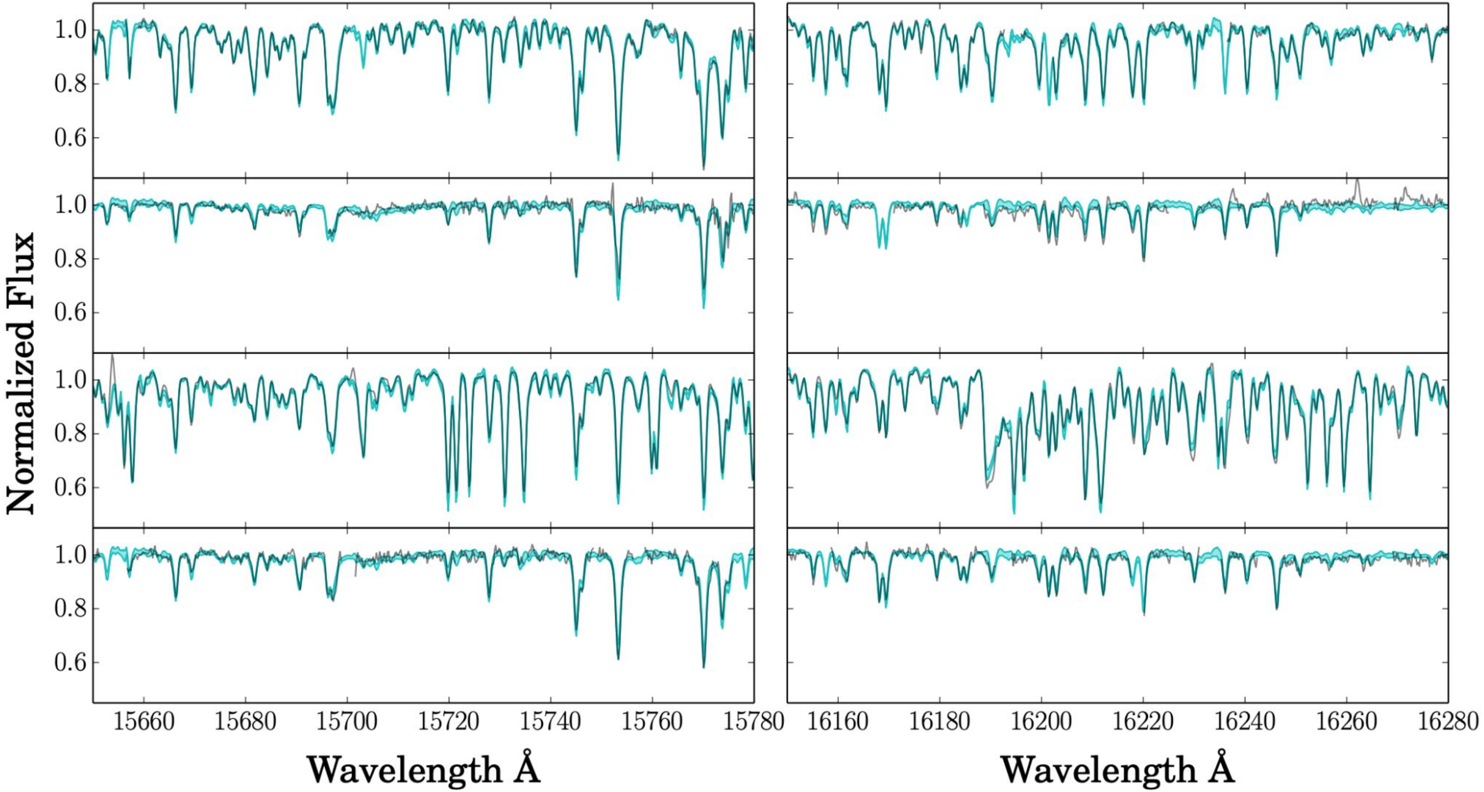
Technical point: Confirmation bias

- *I don’t have a solution for this problem.*
  - (But I’ll return to it at the end.)

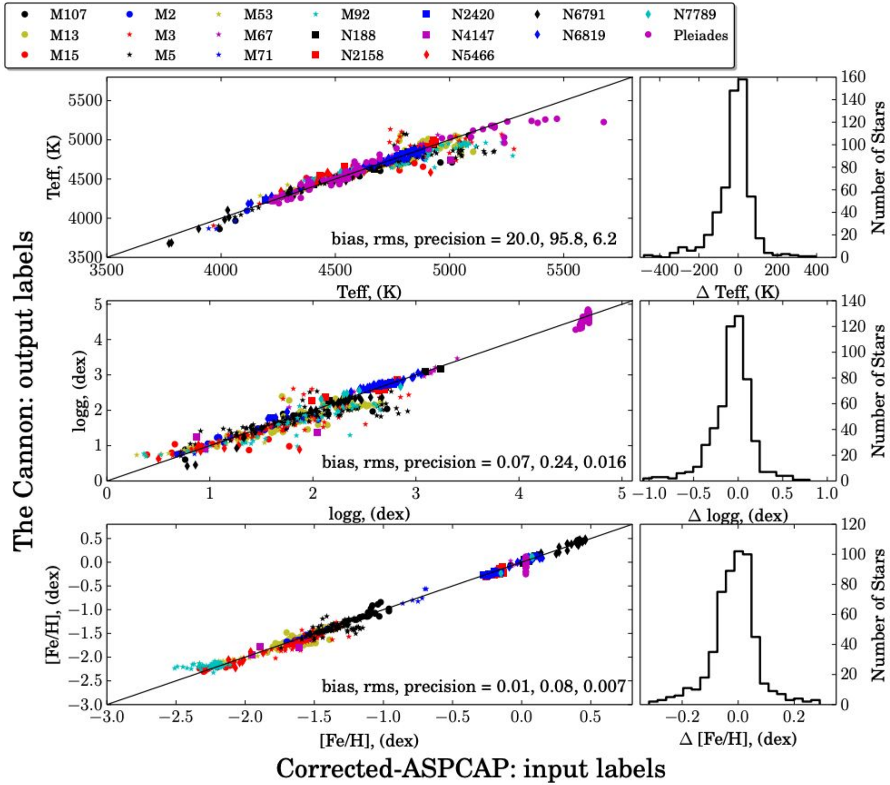
Stellar parameters

- Take a spectrum of a star, infer the mass, age, and composition of that star.
- Very hard to do; requires excellent data, good judgement, and a whole lot of computation.
- So we label a few stars, and then use ML regression to label the rest.
  - With the ESA Gaia Mission data, this has become a cottage industry.

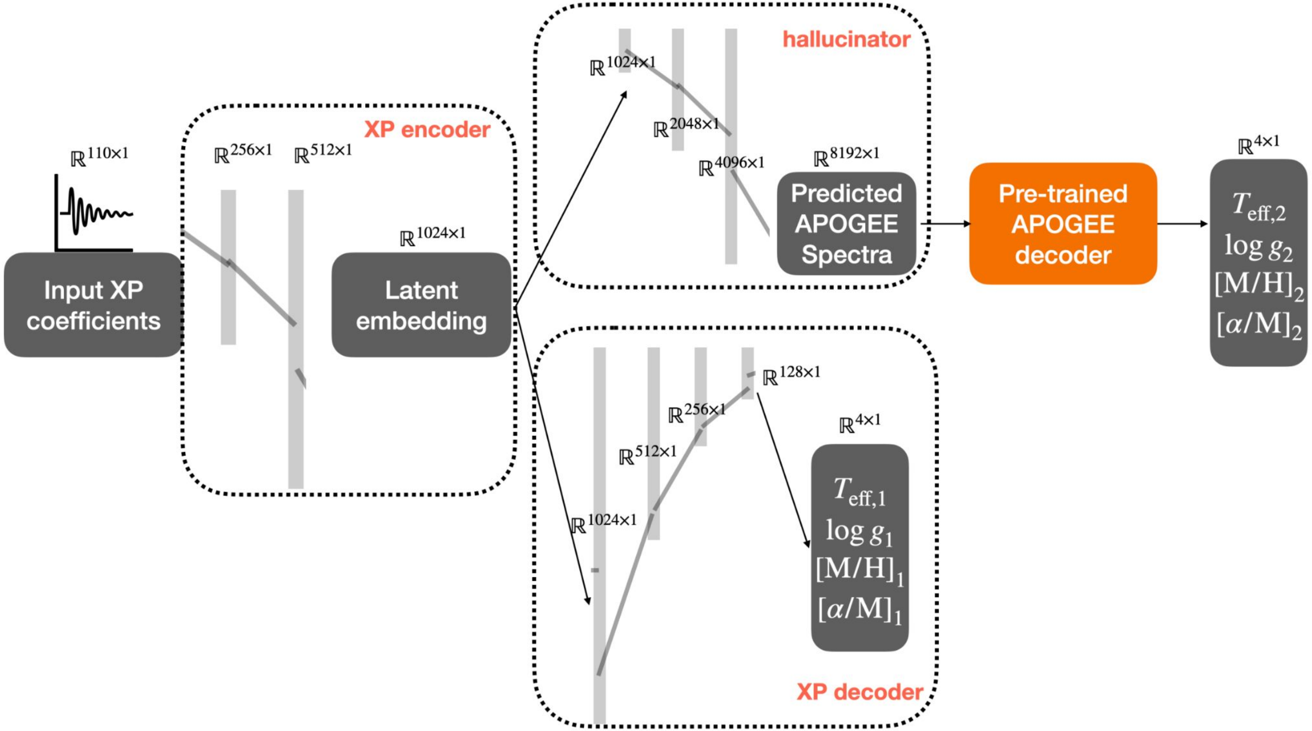
Example: The Cannon (Ness *et al* arXiv:1501.07604)



Example: The Cannon (Ness *et al* arXiv:1501.07604)



Example: AspGap (Li *et al*, arXiv:2309.14294)



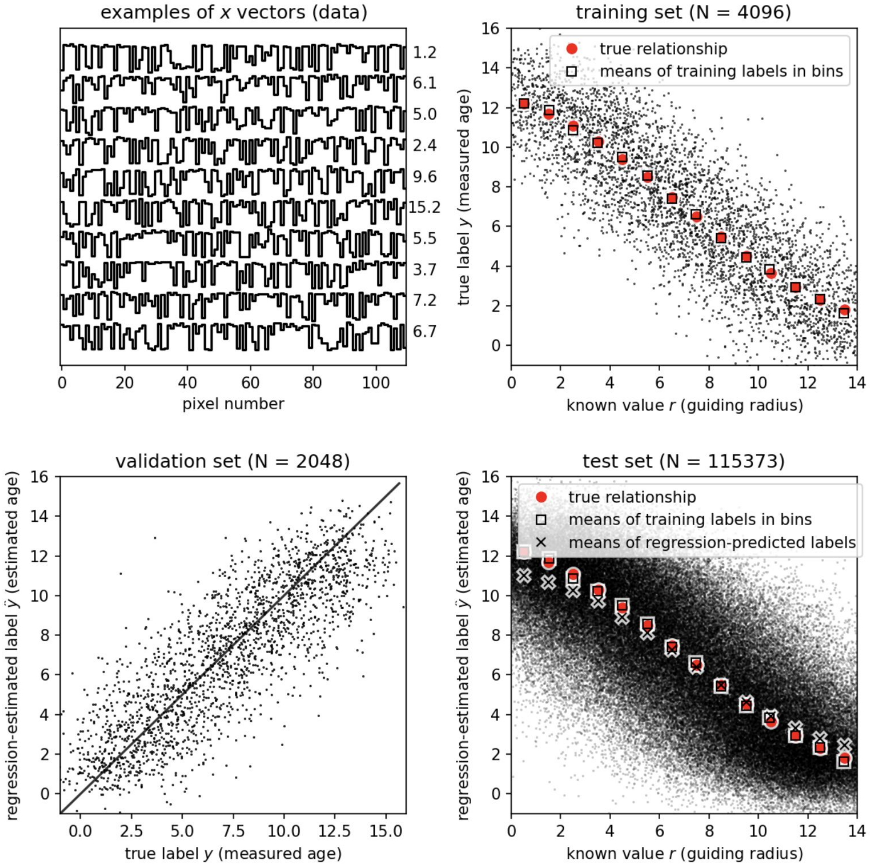
Technical point: Population and joint analyses

- If you want to perform joint analyses on multiple objects (or multiple data sets), you have to combine their likelihood functions.
    - If you try to combine their posterior pdfs, you will end up exponentiating your prior pdfs.
- Almost no ML regressions or classifications return quantities related to likelihood functions.
    - They tend to return posterior quantities, where the training set takes the role of the prior.

Technical point: Population and joint analyses

- Example: You have 1000 stars in some region of the Galaxy. What is their average age?
- If you take the average of **maximum-likelihood estimates** of their ages, you get an **unbiased estimate** of the average age.
- If you take the average of **posterior estimates** of their ages, you get a **highly biased estimate**.
  - It’s like you took your prior to the 1000th power.
  - ML regressions generally return posterior estimates.

Technical point: Population and joint analyses



Technical point: Population and joint analyses

- *I don't have a solution for this problem.*
  - (well actually, some ML methods—like *The Cannon*—return maximum-likelihood estimates)

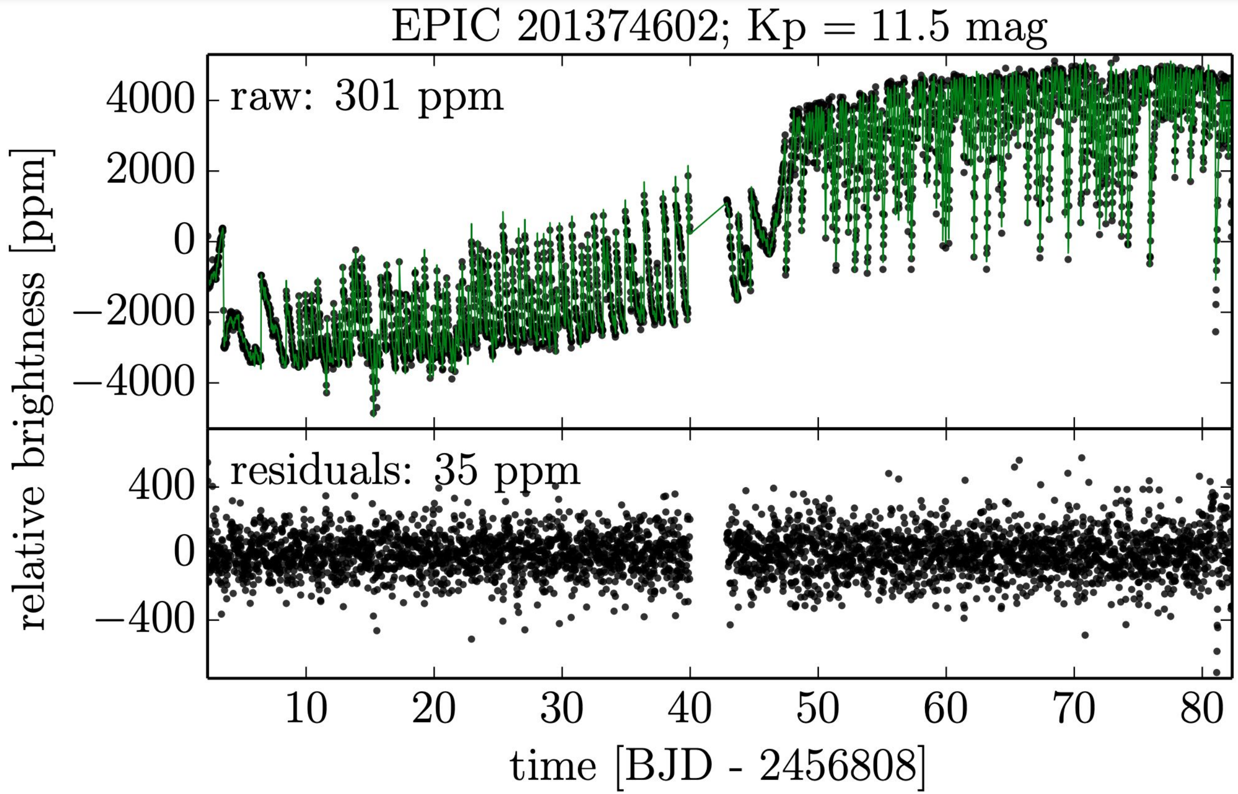
Causal inference in astrophysics?

- Social sciences and health sciences often foreground **causal inference**.
- Physical sciences less so, but:
  - Was this data feature produced by the star, or by the atmosphere? Or by my instrument?
  - Is that a signal or just a background effect?
  - If I had observed for longer, what would I have seen?

Instrument calibration

- Say we are measuring the brightness of a star extremely sensitively.
- What variations are due to the star, what are due to the instrument?
  - And what are due to any planets?
- You make the best argument that the signal is due to the star, when you have **given your instrument model a lot of flexibility**.
- Often (but not always), **you don't need to interpret** your instrument model.

Example: Planets in NASA *K2* (Foreman-Mackey et al, arXiv:1502.04715)



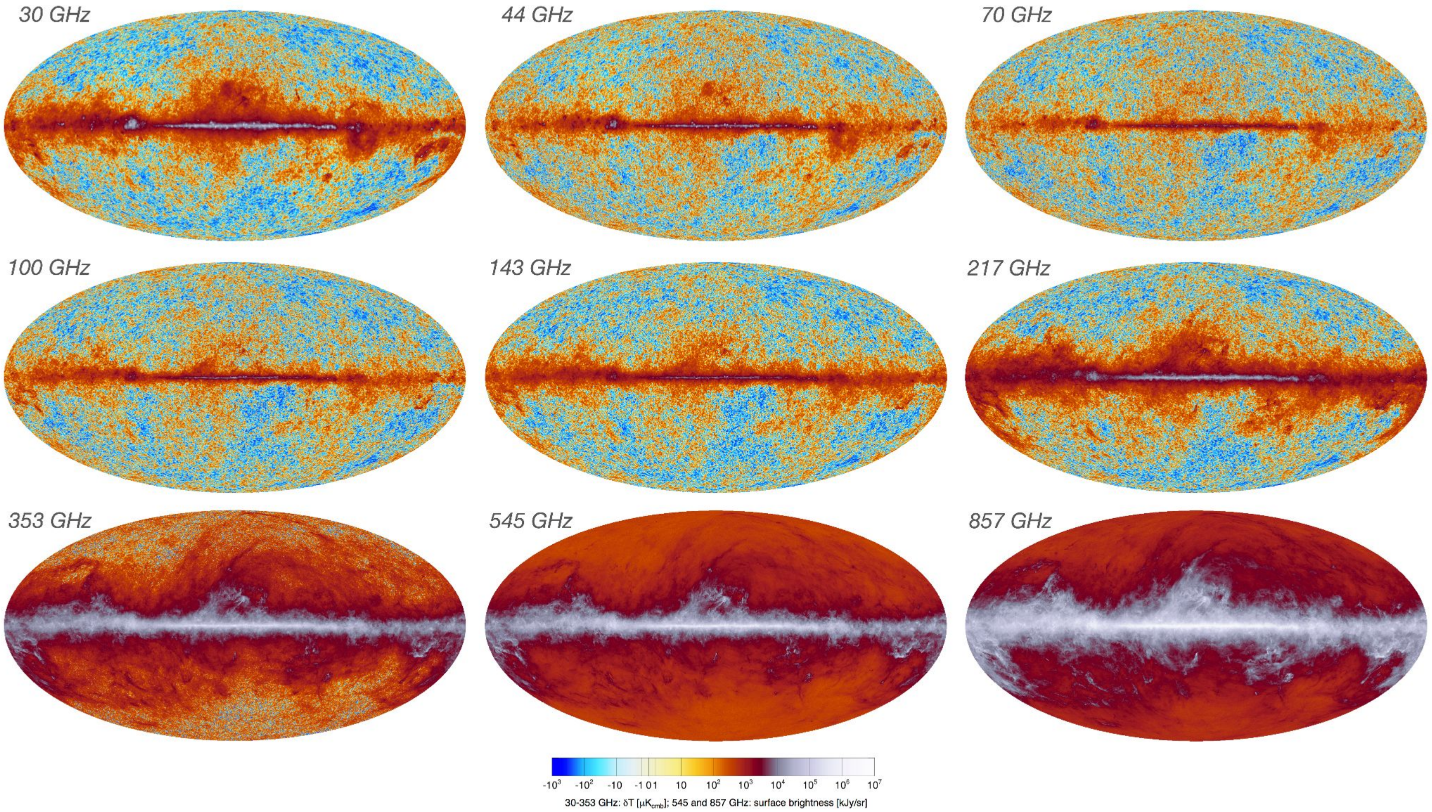
Instrument calibration

- Note the connections to engineering.
  - ML is useful in instrument calibration precisely because instrument calibration is part of the engineering infrastructure of the scientific project.

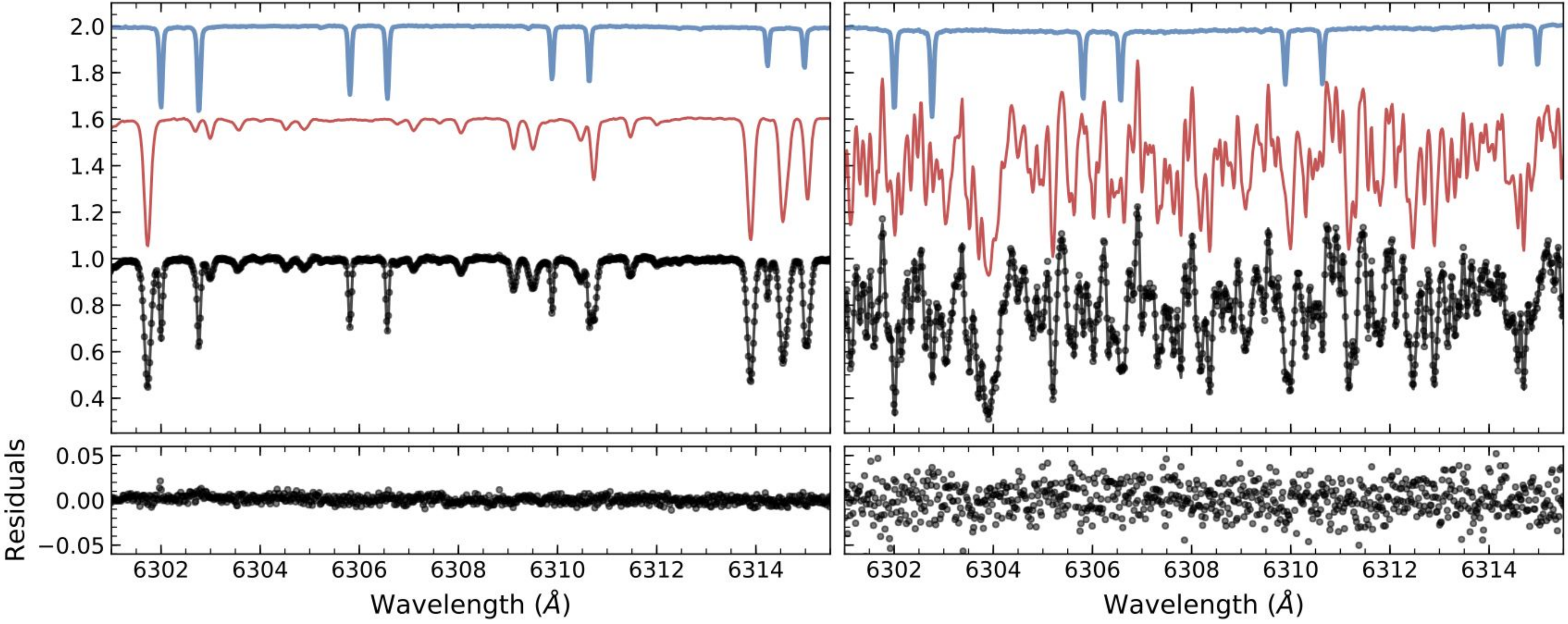
Backgrounds (or foregrounds)

- Most astrophysical data are contaminated by backgrounds and foregrounds.
- A subtle signal of interest is only believable when the background and foreground models have been given lots of flexibility.
- And by assumption, these are the signals **you don't care to understand**!

Example: Foregrounds in ESA *Planck*



Example: wobble spectral model (Bedell et al, arxiv:1901.00503)



Conservatism

- It is generally considered *cavalier*, and not *conservative*, to throw ML at your scientific data.
- However, in causal inferences, *the most conservative thing you can do* is give your nuisances and confounders maximum flexibility.
    - **ML can provide the most conservative possible approaches to these problems!**

Open question: Trust in emulators

- It is obvious that emulation of expensive simulations (and other expensive computation) is here to stay. It’s happening.
- So, we need to figure out ways to build trust systems for emulators.
    - We’re exploring methods involving exact symmetries.
    - We’re exploring methods built on adversarial training.
    - Maybe there are ways to introduce sanity checks and sparse resimulations?
    - (all joint work with Soledad Villar @ JHU)
- Many of these issues arise in **artificial intelligence** more generally.

What I said

- Machine learning tools are dangerous.
- Their use can lead to **badly biased outcomes**.
- However, there are contexts in which **ML methods are our only choice**, for computational reasons (*eg*, emulation), and for intellectual reasons (*eg*, calibration).
    - We have work to do if we are going to ensure that our scientific results remain **accurate**.