Can machine learning be used to make measurements?

David W Hogg

NYU / MPIA / Flatiron

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

Work in collaboration with **Soledad Villar** (JHU).

What I’m going to say

- Negative side:
  - Regressions produce **biased predictions**, not measurements.
  - Simulation-based inferences with emulators amplify **confirmation biases**.
- Positive side:
  - ML should (actually *should*) be used on nuisances, such as **calibration and backgrounds**.
  - In causal-separation problems, flexibility is paramount (and interpretation is not).

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, large 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*.

What is a measurement?

- Hot take: A measurement is a peak in a **likelihood function**!
  - If you want your measurement to be used downstream, it has to be either unbiased,
  - or (at least) **can be combined** with other measurements.
- Even if you are *strictly Bayesian*, you should agree with this.
  - Information comes from data via a likelihood function (likelihood principle).
  - If you want to use someone else’s measurement, you want their LF, not their posterior.
  - A measurement is not an *expression of your belief*! It is a statement about the data.

How is machine learning used to make measurements?

- **Regression**: Train a function that predicts labels from data.
  - Execute that function on a new datum: New measurement of that label. *Bad!*
- **Emulation**: Speed up simulation-based inferences or likelihood evaluations.
  - This makes measurements possible that wouldn't otherwise be. But *also bad!*
- **Statistics discovery**: Find sufficient (or good) statistics of the data.
  - Maybe a good idea? Great place for symmetries, etc. I'm not going to talk about this.
- **Causal separations**: Model and remove backgrounds and instruments.
  - A great idea, especially when executed hierarchically.

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!

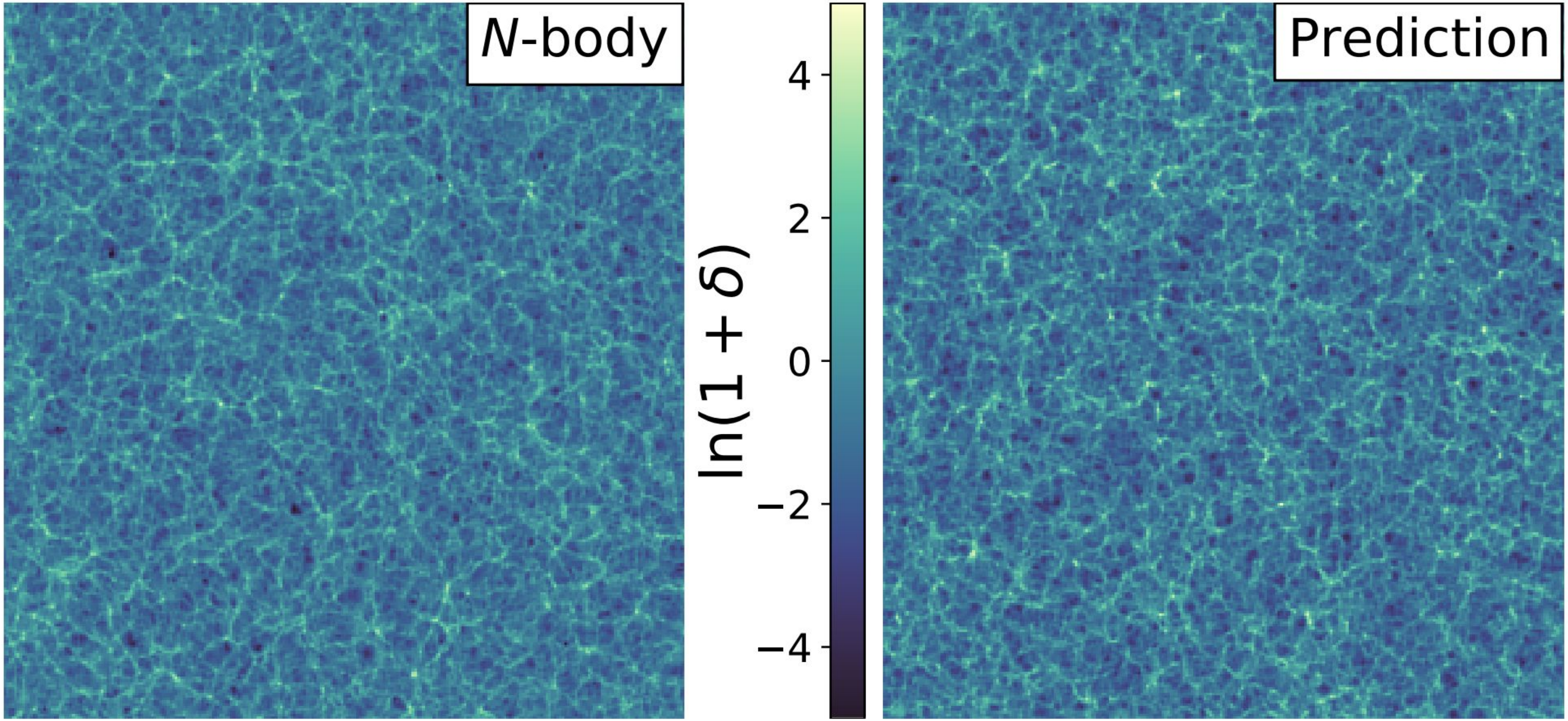
Trust issues

- 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

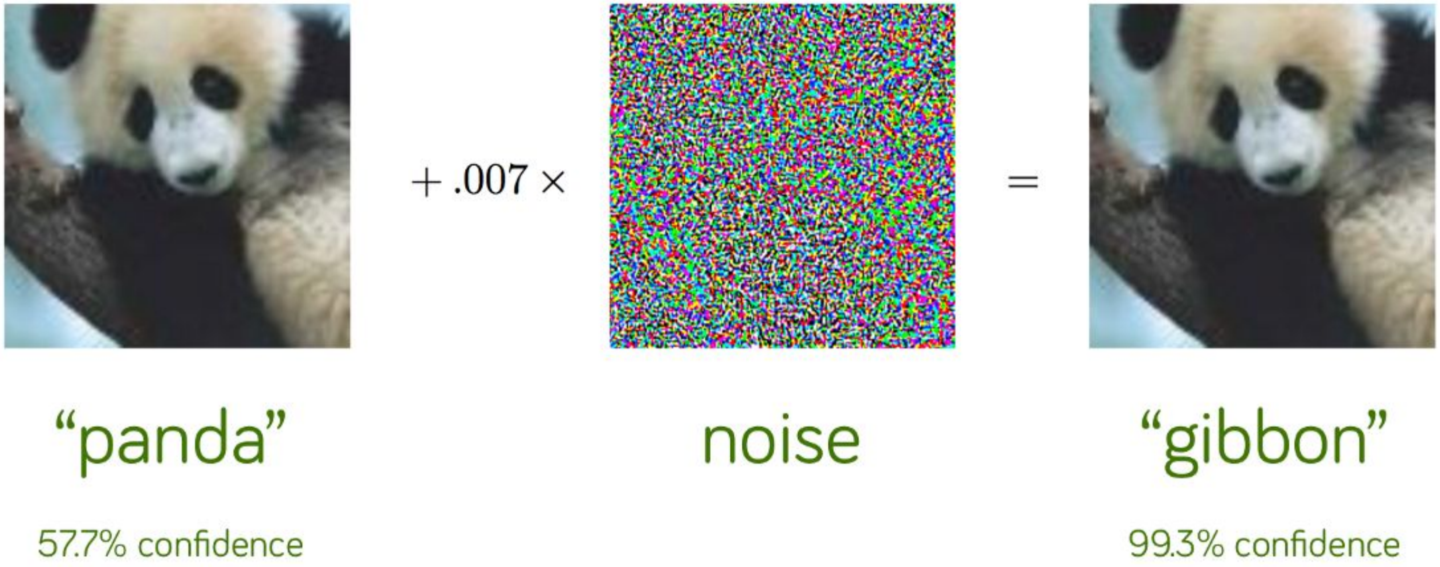
- Where in science can you use a model that you don’t understand?

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



10

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.

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.

Confirmation bias

- *I don’t have a solution for this problem.*

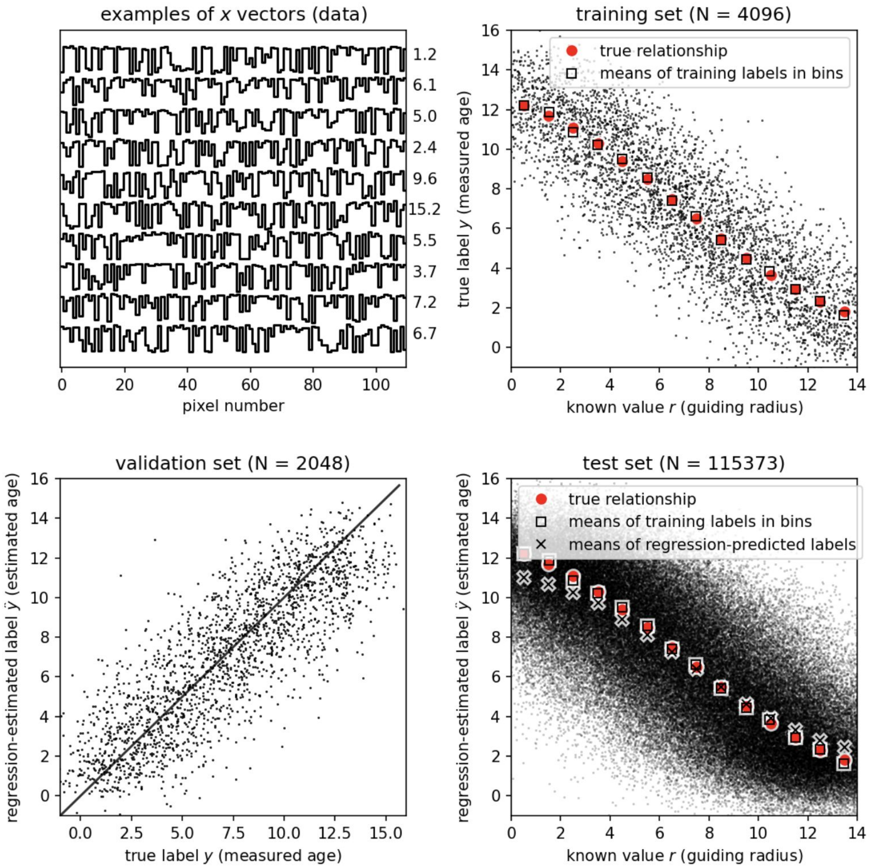
Population and joint analyses

- 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 something akin to **posterior quantities**, where the training set takes the role of the prior.

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.

Population and joint analyses (Hogg & Villar arXiv:2405.18095)



Population and joint analyses

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

The question

- I asked: *Where in science can you use a model that you don't understand?*
- But in astrophysics *we use instruments we don't understand* all the time.
    - Example: Almost all infrared detectors, including those on ESA *Euclid*.

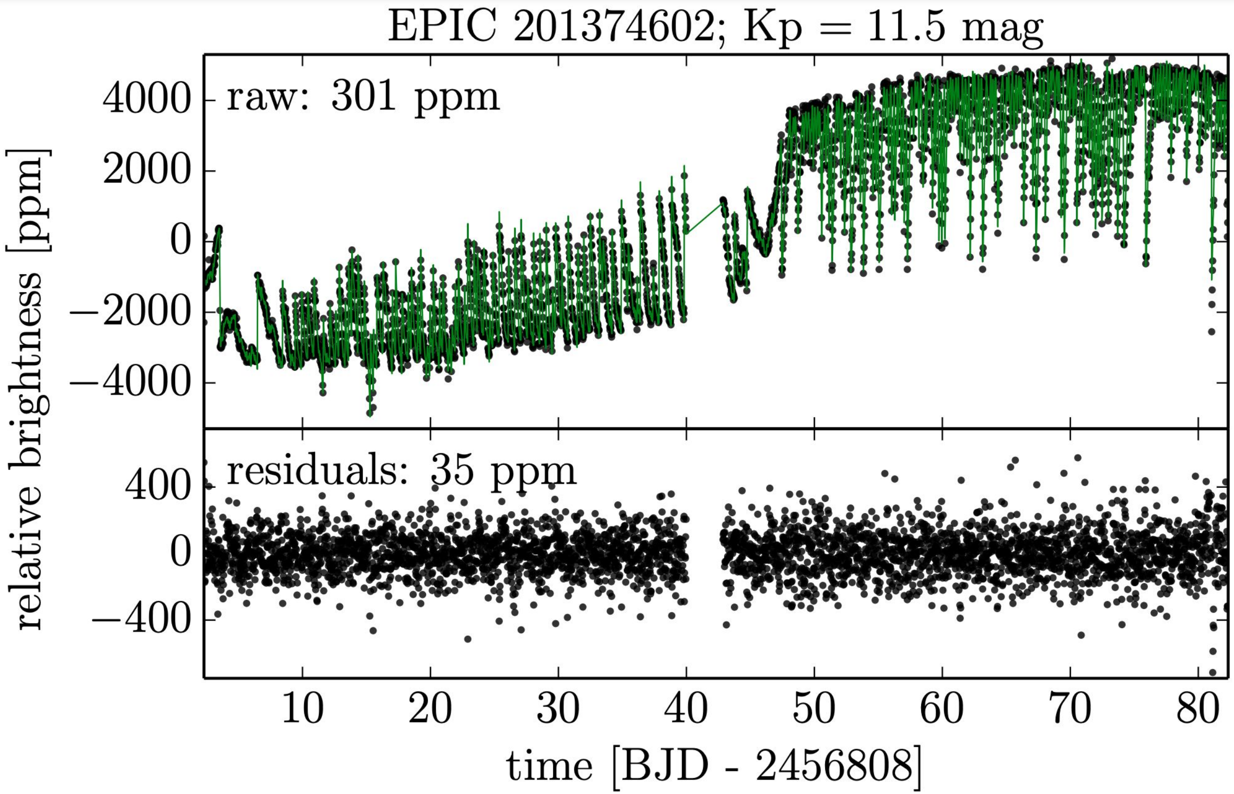
Causal inference in astrophysics?

- 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?

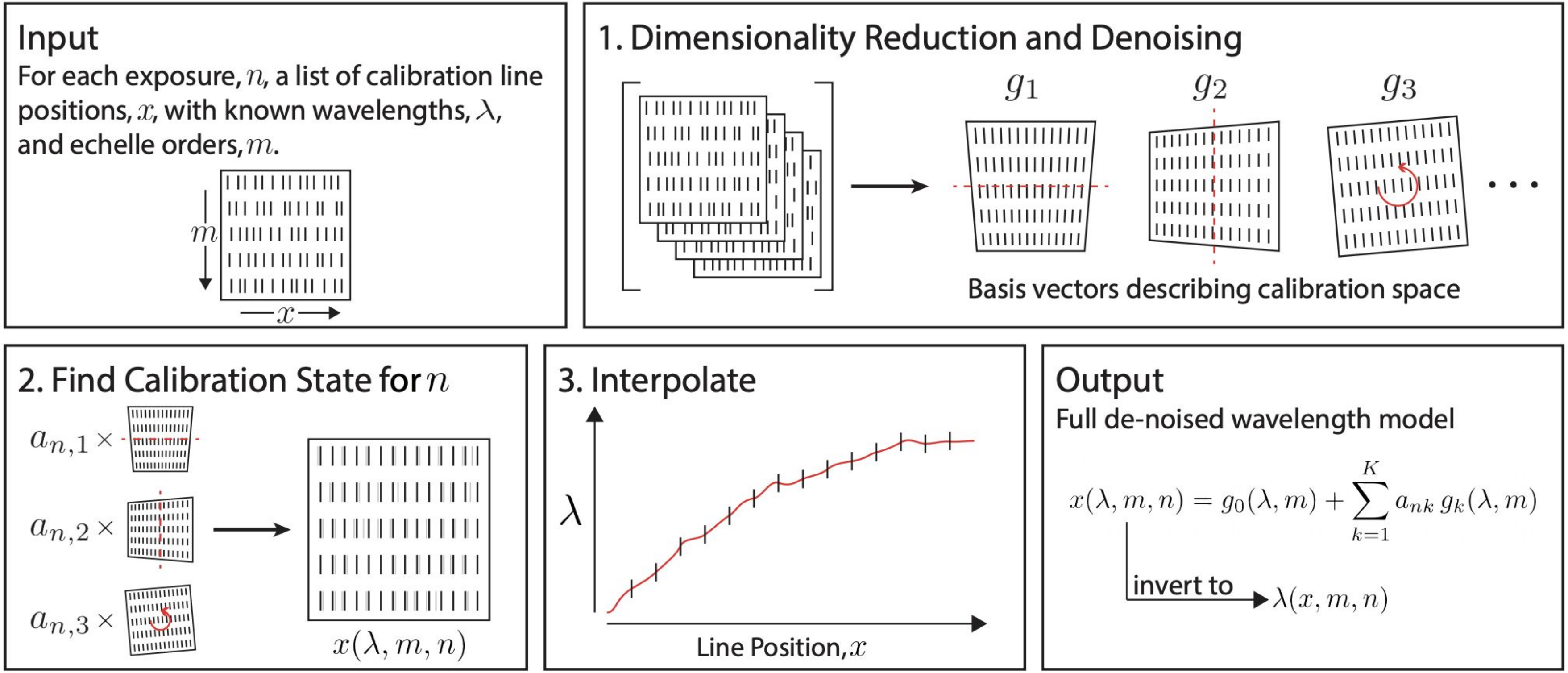
Instrument calibration

- 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.

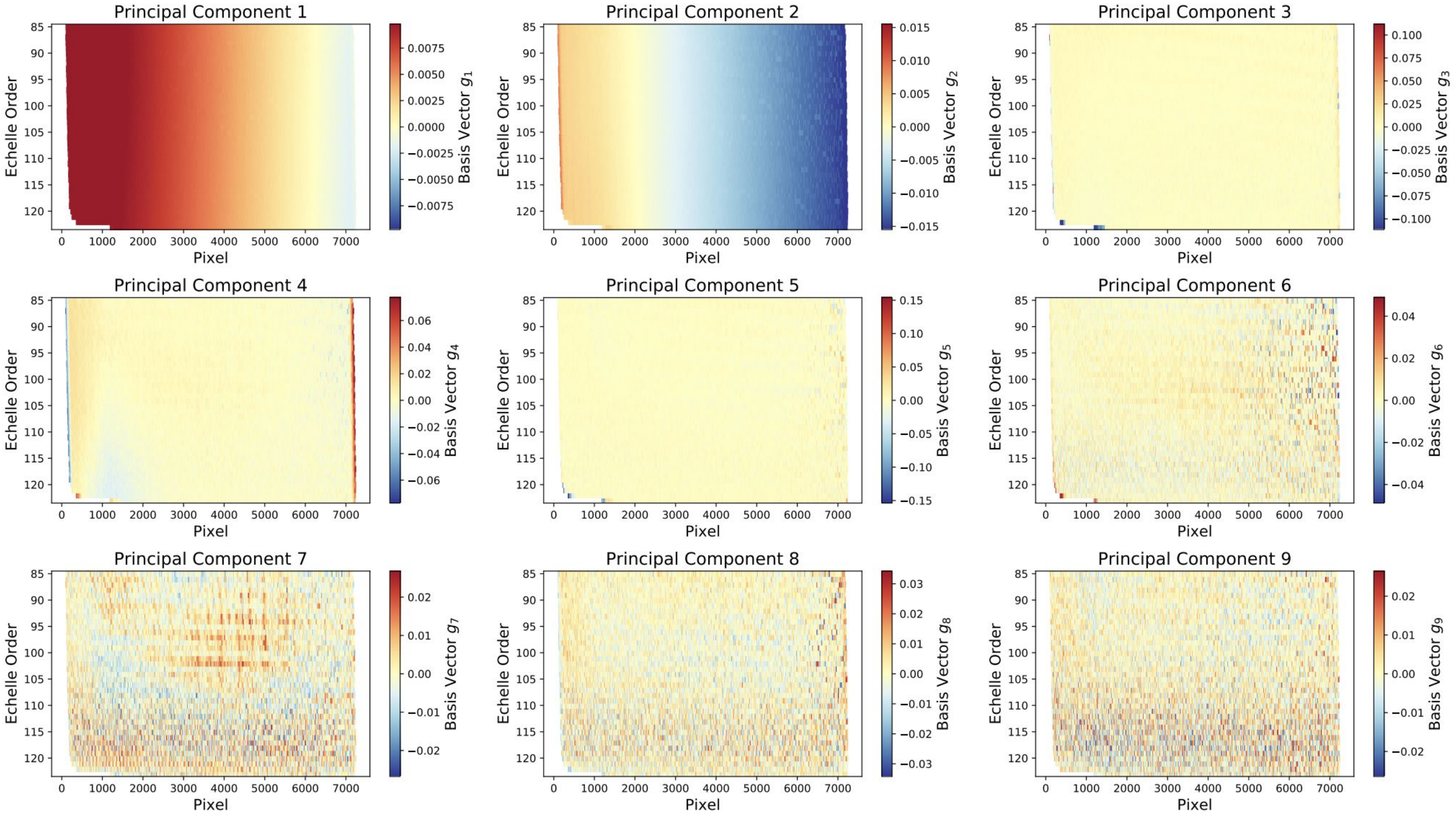
Example: Planets in NASA *K2* (Foreman-Mackey et al, arXiv:1502.04715)



Example: Excalibur (Zhao *et al*, arXiv:2010.13786)



Example: Excalibur (Zhao *et al*, arXiv:2010.13786)



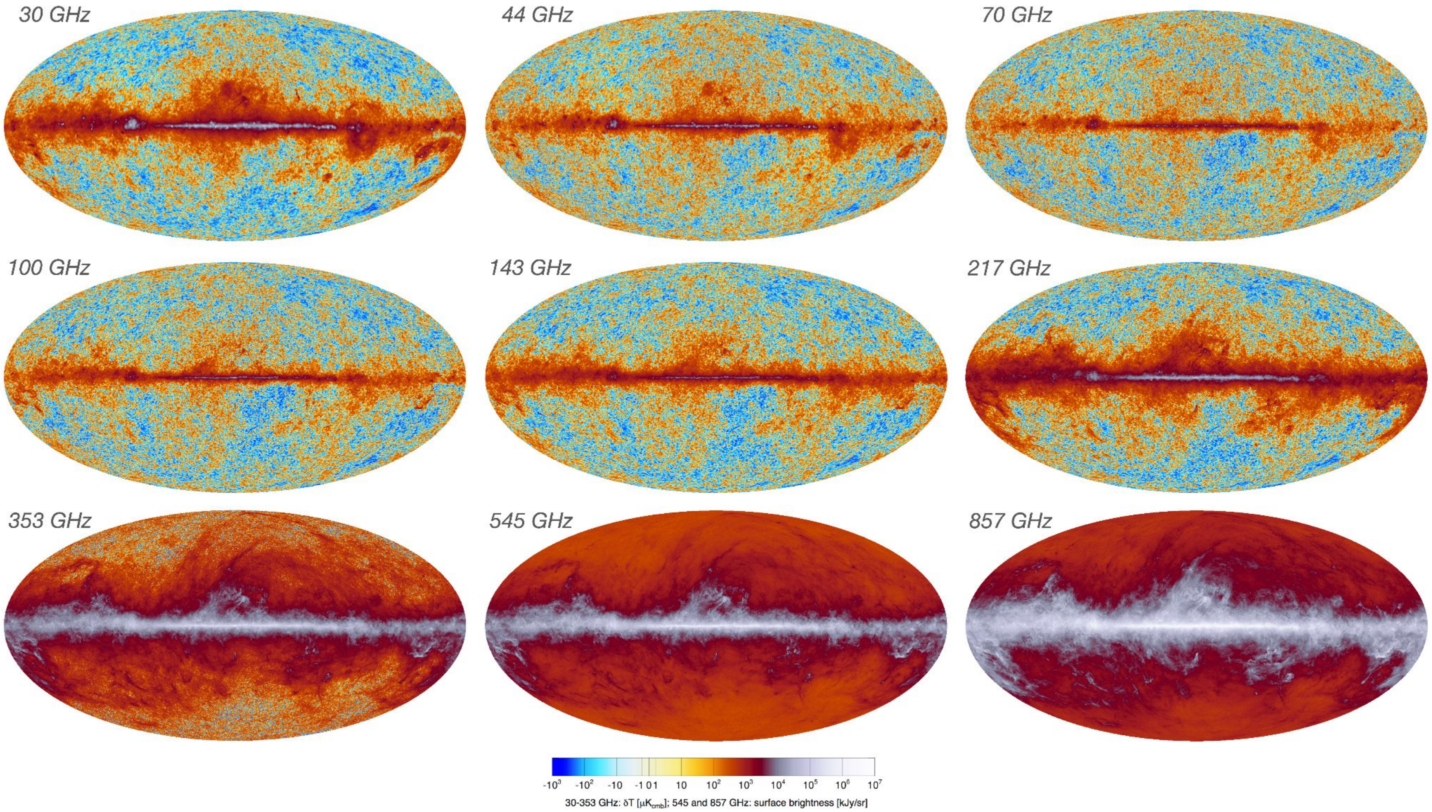
Instrument calibration

- Want **enormous flexibility** to capture unexpected instrument properties.
- Want **hierarchical structure** to restrict that flexibility appropriately.
  - Related to representation learning?

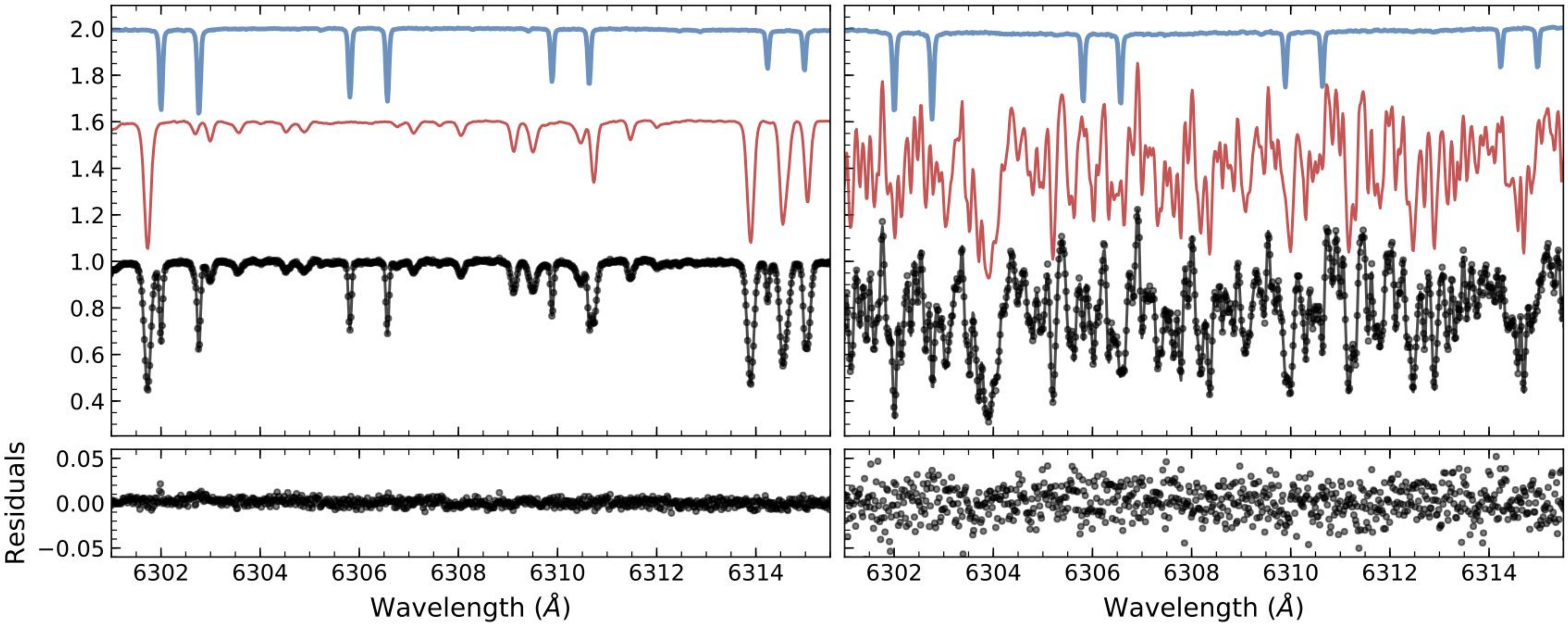
Backgrounds (or foregrounds)

- 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

- Negative side:
  - Regressions produce **biased predictions**, not measurements.
  - Simulation-based inferences with emulators amplify **confirmation biases**.
- Positive side:
  - ML should (actually *should*) be used on nuisances, such as **calibration and backgrounds**.
  - In causal problems, flexibility is paramount (and interpretation is not).