Statistical Modeling *and* Intro to ML

Adrian Bayer

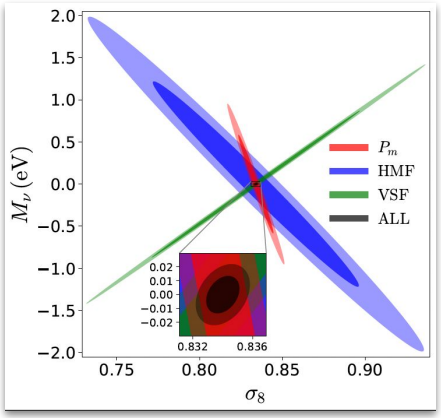
AstroAI Asian Network
Summer School 2025
KIAS

PRINCETON UNIVERSITY

SIMONS FOUNDATION

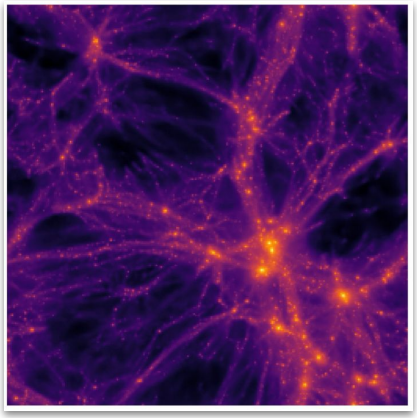
About Me

Postdoc @ CCA + Princeton
https://adrianbayer.github.io



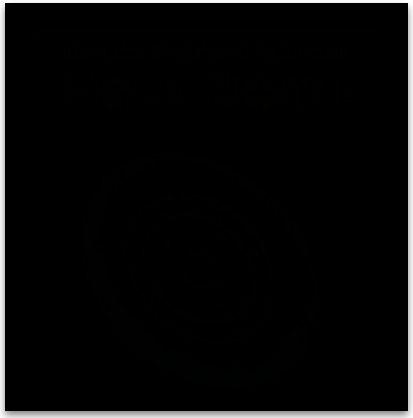
Neutrino mass from LSS

Lots of information from higher-order statistics for 3d matter field, but fake νs for galaxy / lensing.



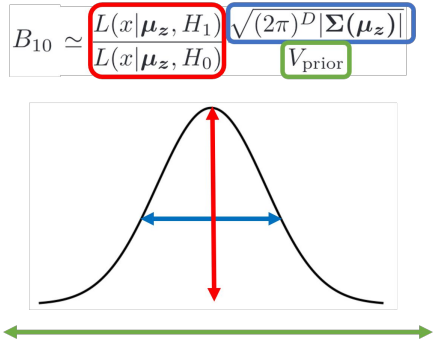
Field-level inference

Extract all the information from cosmic fields.
Sampling. Machine learning. BAO reconstruction. Interpretability and Robustness.



LSS x CMB Simulations

Big box, full sky simulations for correlated analyses.

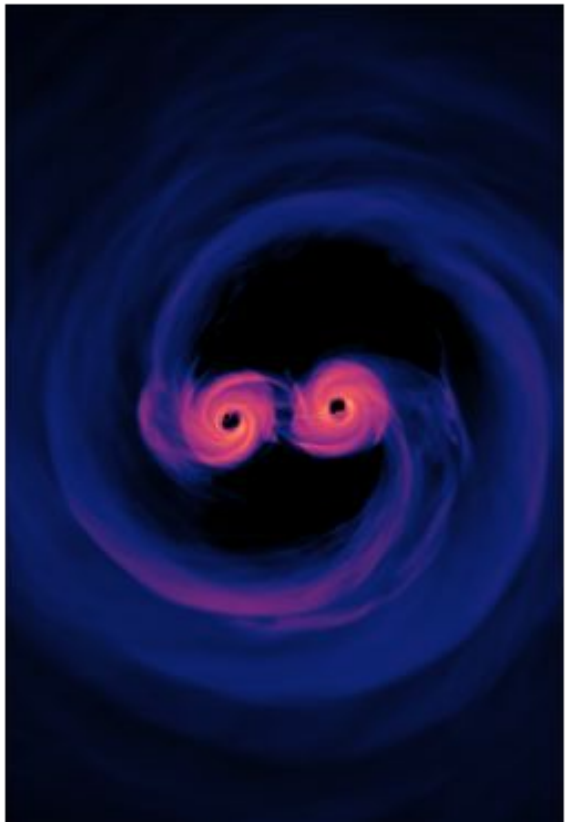


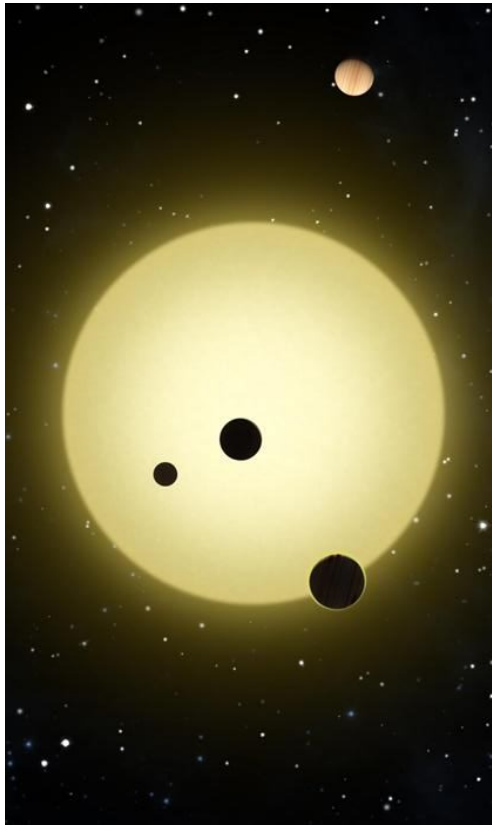
Look-Elsewhere Effect

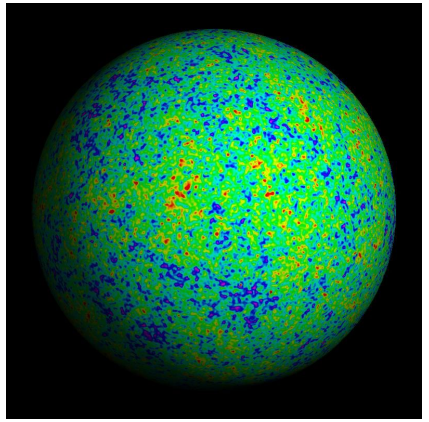
Large prior, small posterior? Unified Bayesian & frequentist approach to correct statistical significance.
Recently applied to SMBHB.

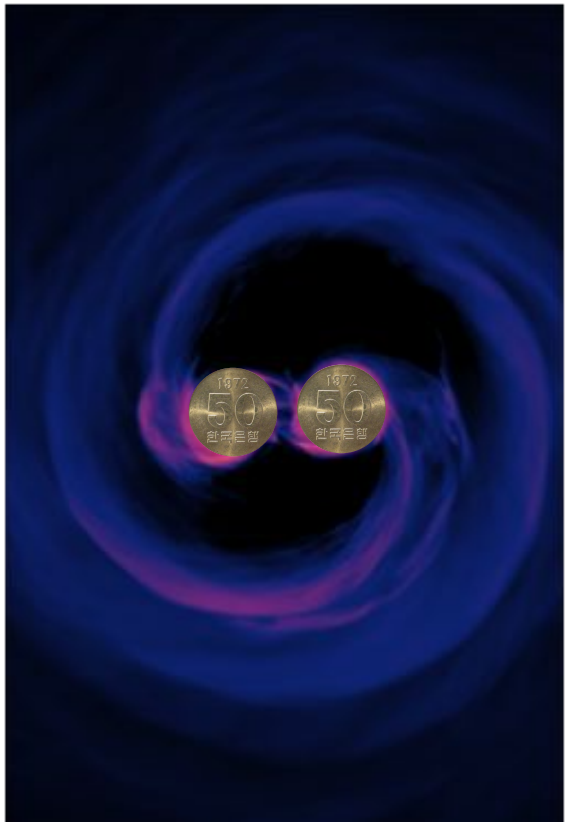

Overview
Frequentist x Bayes
Connection to ML
General ML Tricks of the Trade


Pay Attention
There are three images I made with ChatGPT
and it did something silly.
Shout (or raise your hand) if you spot them!





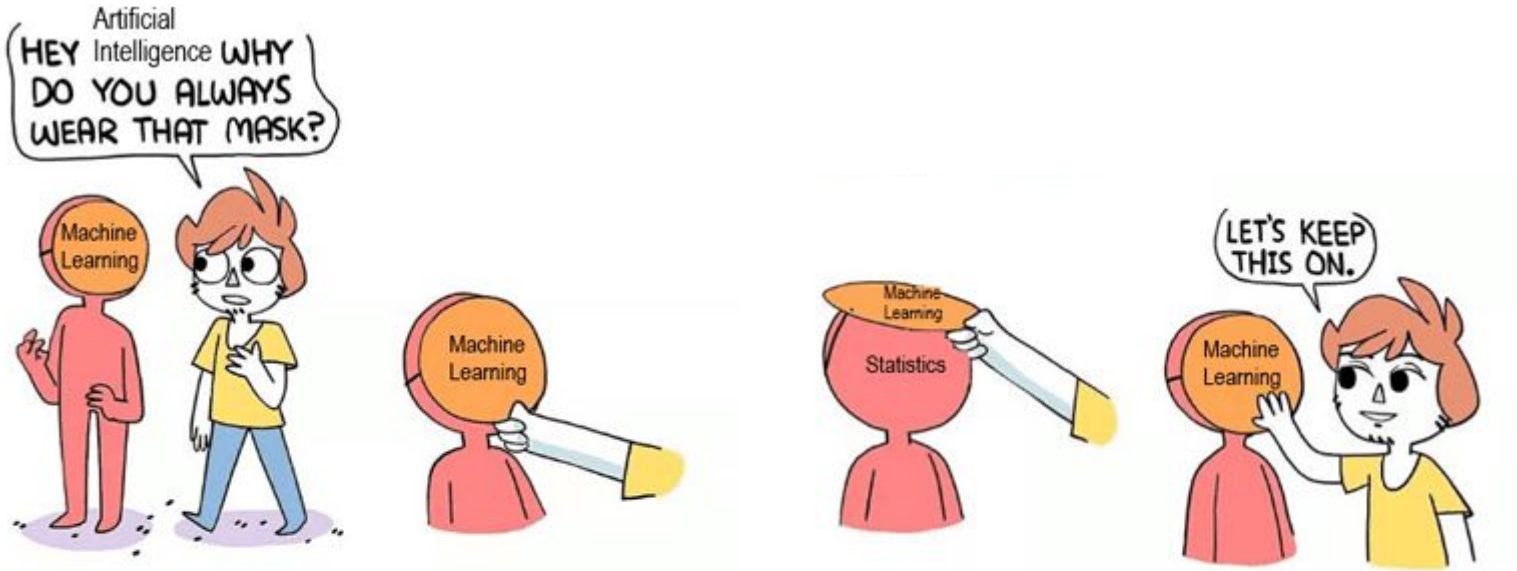



1972
50
한국은행
1972
50
한국은행

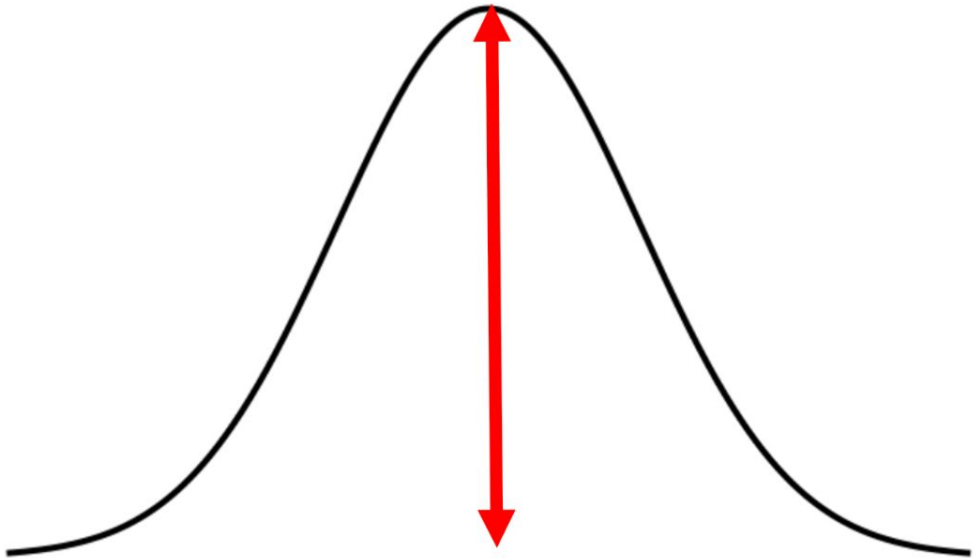

한국은행

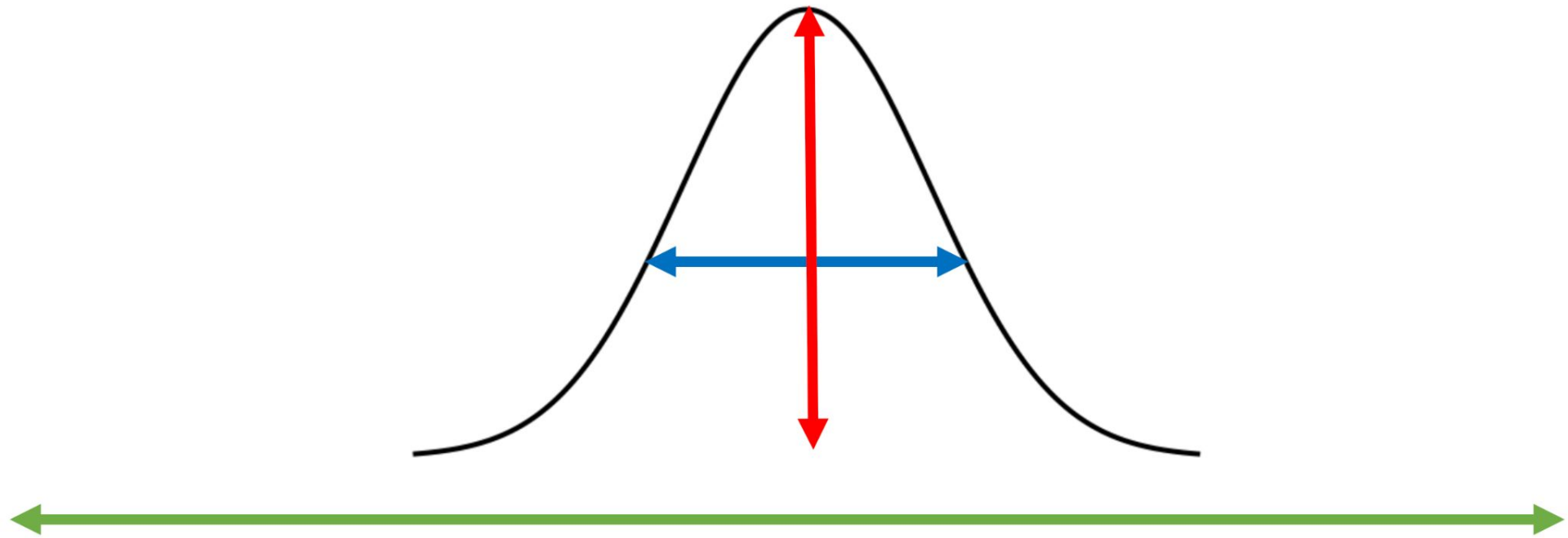

오백원


십원
한국은행

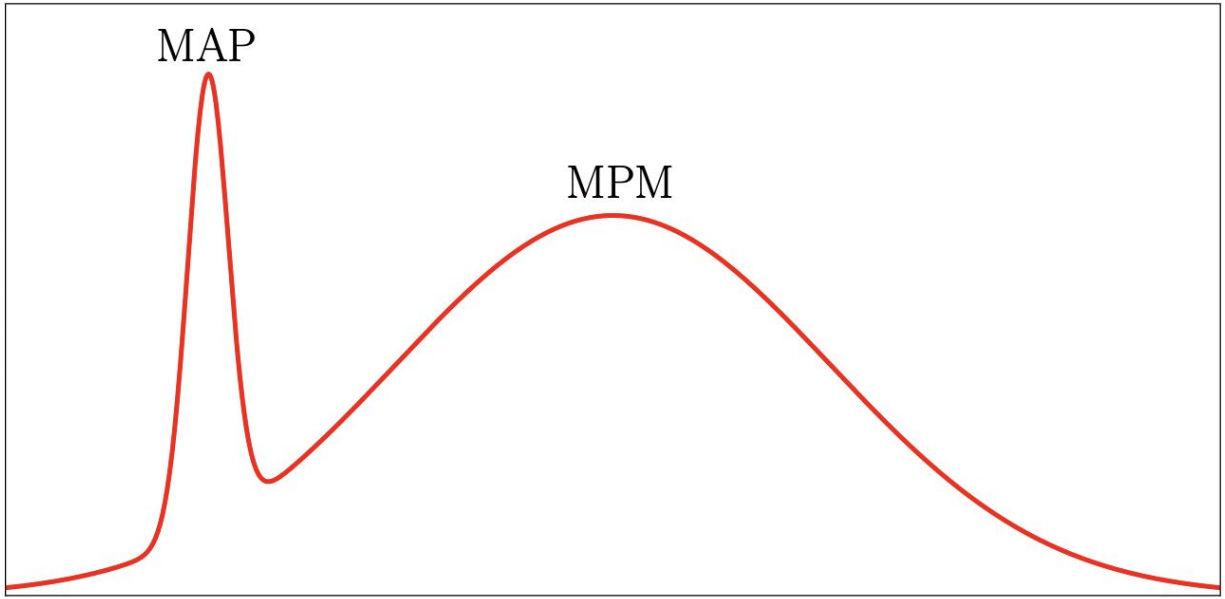

Artificial Intelligence
HEY WHY DO YOU ALWAYS WEAR THAT MASK?
Machine Learning
Machine Learning
Machine Learning
Statistics
LET'S KEEP THIS ON.
Machine Learning


LOOK AT THE BOARD
THAT'S WHERE I'LL LEADING THIS PART OF THE LESSON



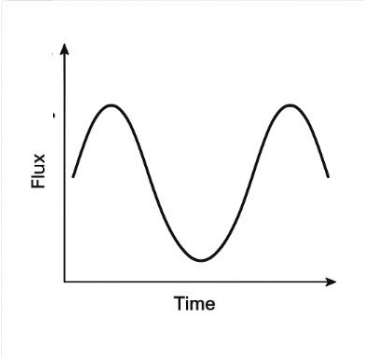


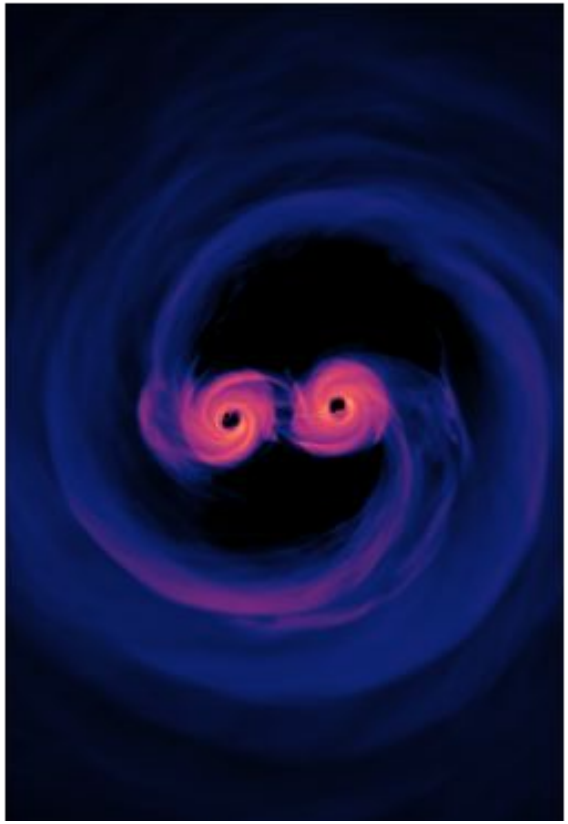
Maximize height or mass?

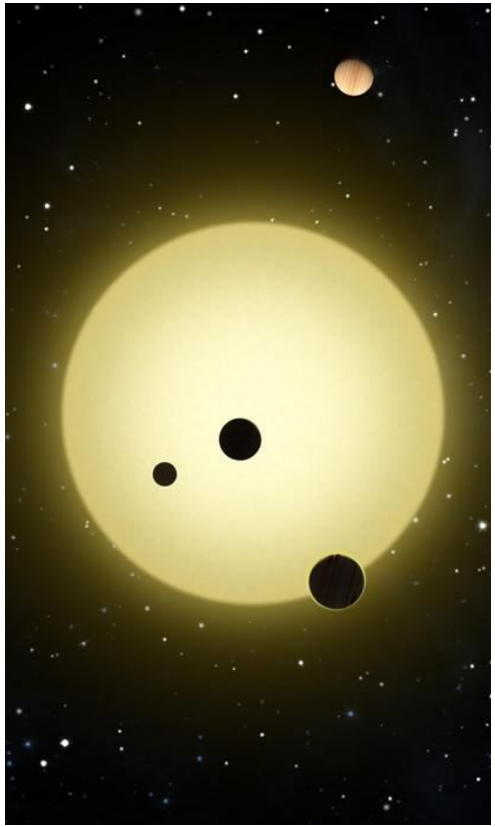


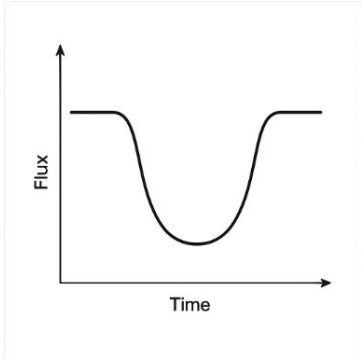
Is dark energy evolving?

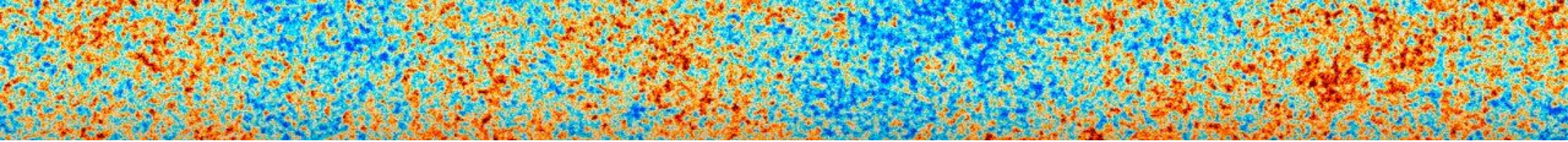
DESI DR2 2503.14738
Volume effects? 2505.02658


Flux
Time





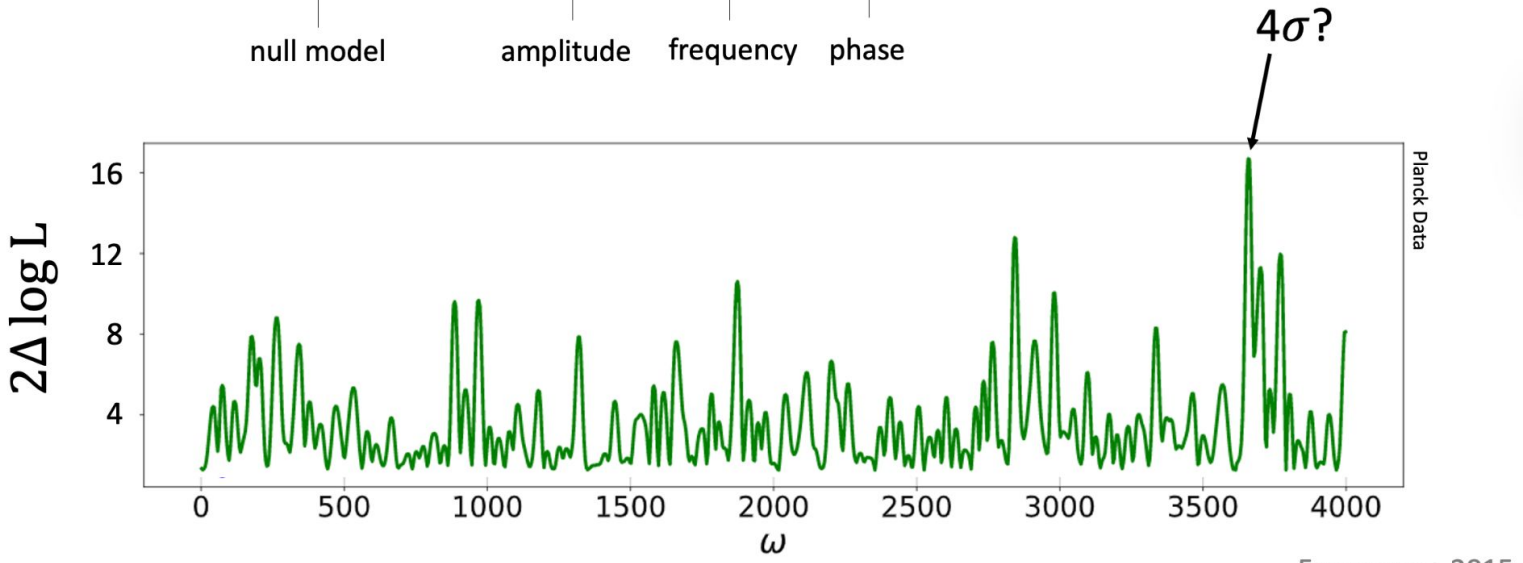

Flux
Time



The L👀k-Elsewhere Effect *Cosmic Inflation Edition*

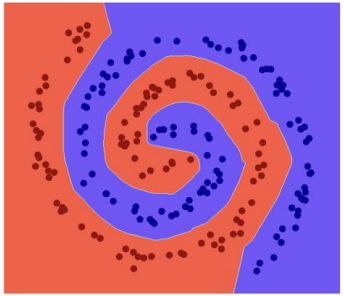
$$P(k) = P_{\Lambda\mathrm{CDM}}(k)[1 + A\sin(2\omega k + \phi)]$$

null model amplitude frequency phase

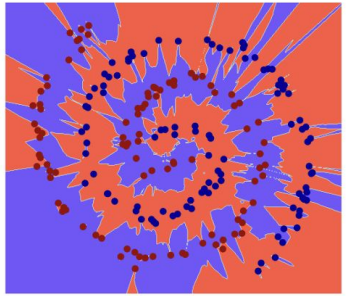


Fergusson+ 2015

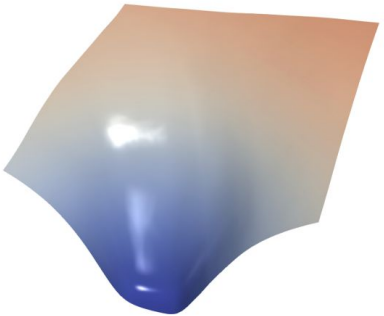
Neural Network Generalization



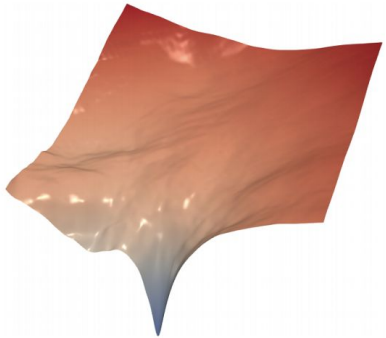
(a) 100% train, 100% test



(b) 100% train, 7% test



(c) Minimizer of network in (a) above



(d) Minimizer of network in (b) above

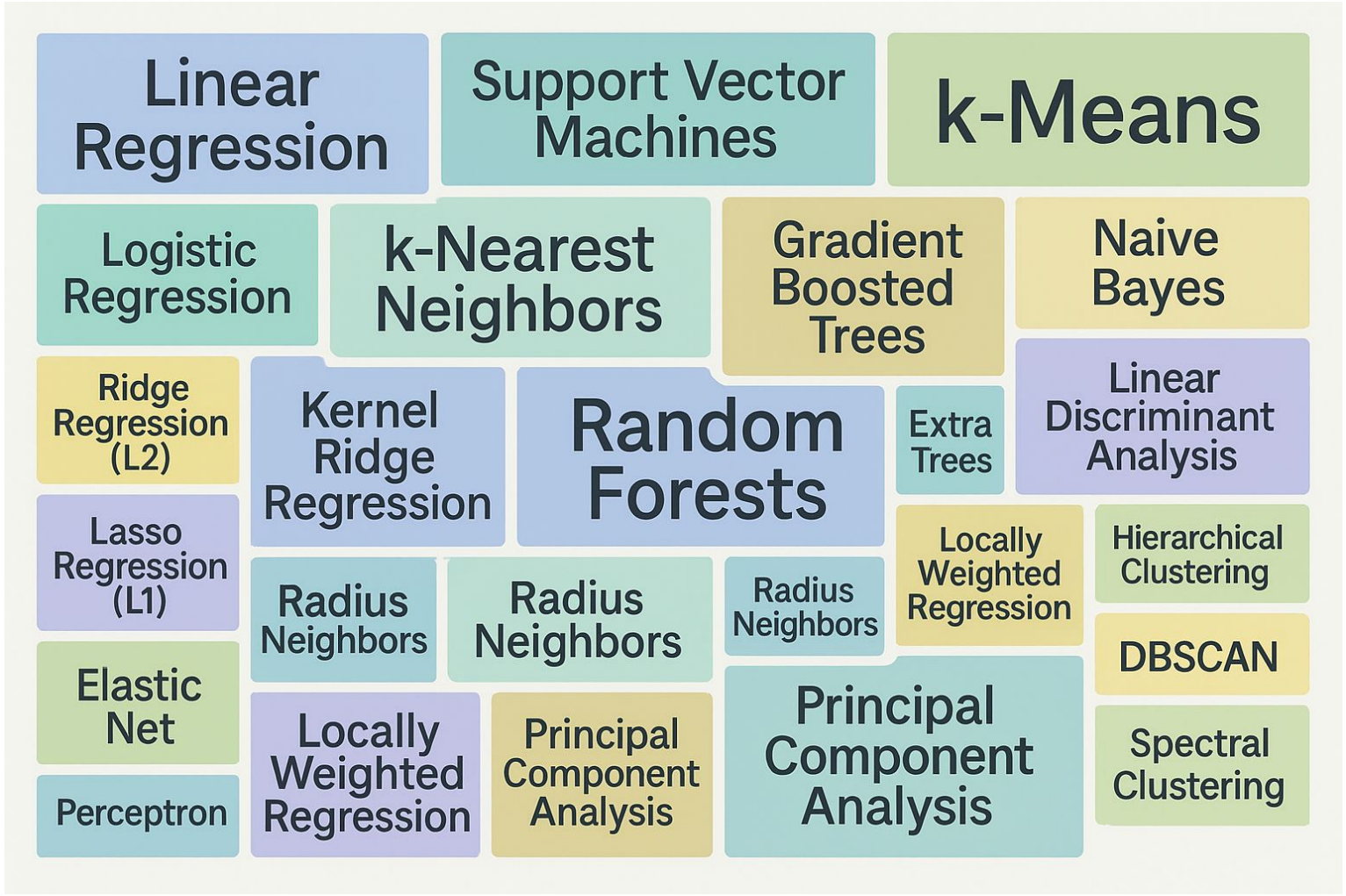
Huang et al. (2020)

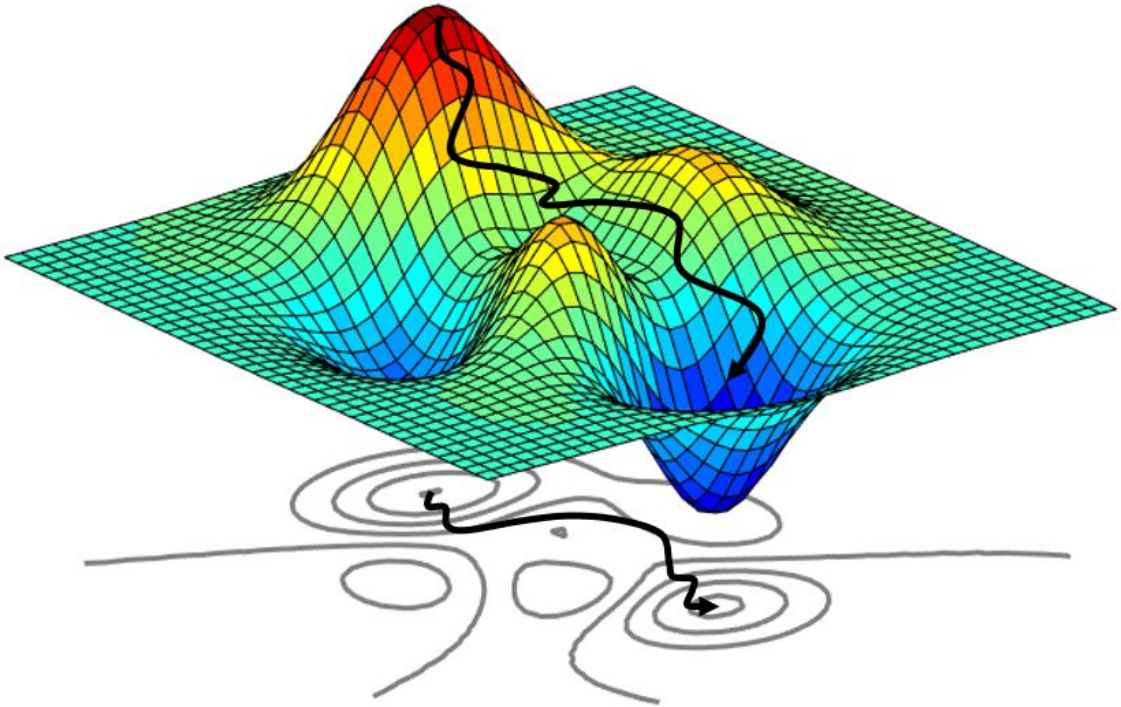
Overview

Frequentist x Bayes

Connection to ML

General ML Tricks of the Trade


Linear Regression
Support Vector Machines
k-Means
Logistic Regression
k-Nearest Neighbors
Gradient Boosted Trees
Naive Bayes
Ridge Regression (L2)
Kernel Ridge Regression
Random Forests
Extra Trees
Linear Discriminant Analysis
Lasso Regression (L1)
Radius Neighbors
Radius Neighbors
Radius Neighbors
Locally Weighted Regression
Hierarchical Clustering
DBSCAN
Elastic Net
Locally Weighted Regression
Principal Component Analysis
Principal Component Analysis
Spectral Clustering
Perceptron



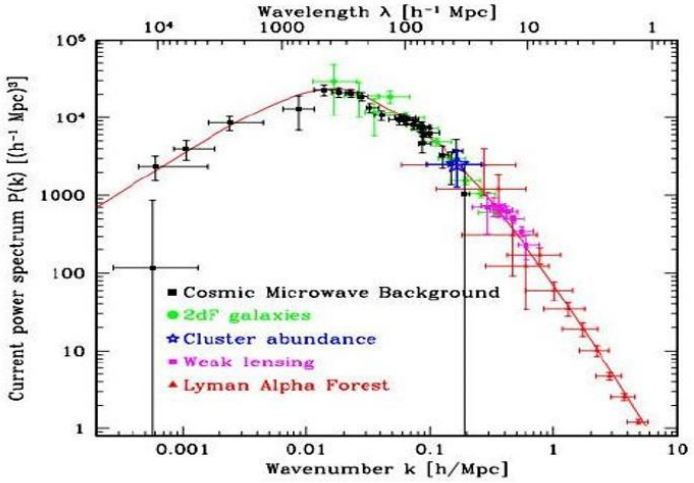

Overview
Frequentist x Bayes
Connection to ML
General ML Tricks of the Trade

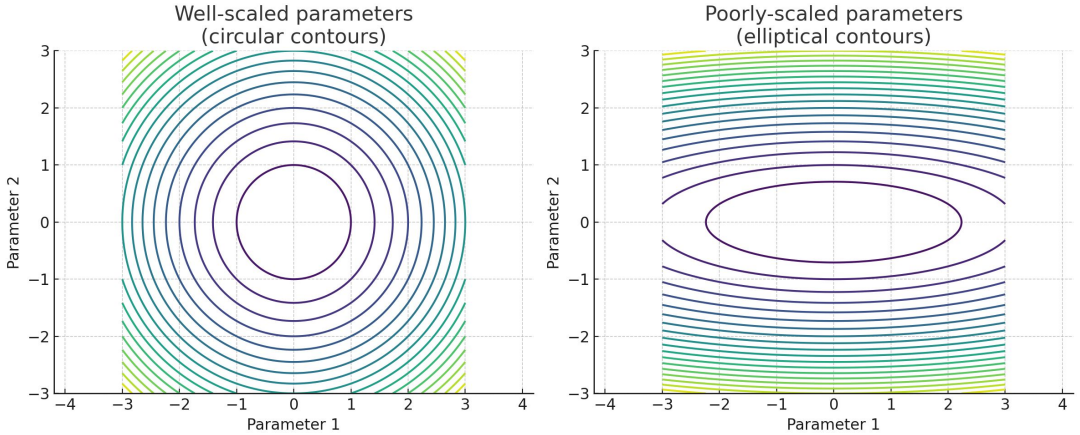
Scaling & Preprocessing

Scaling & Preprocessing

Why: Many ML algorithms need scaled inputs for stability (e.g., gradient descent).

Types: Standardization (z-score), normalization ([0,1] range), log transforms.





Bias–Variance Tradeoff

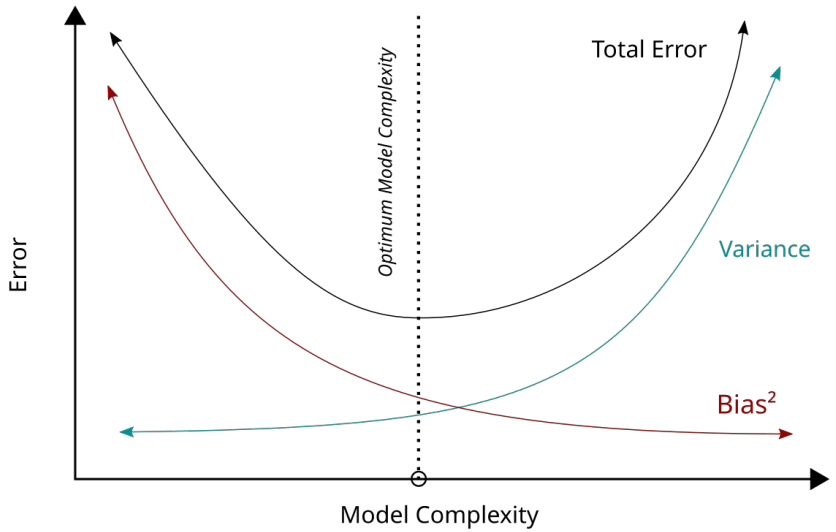
Bias–Variance Tradeoff

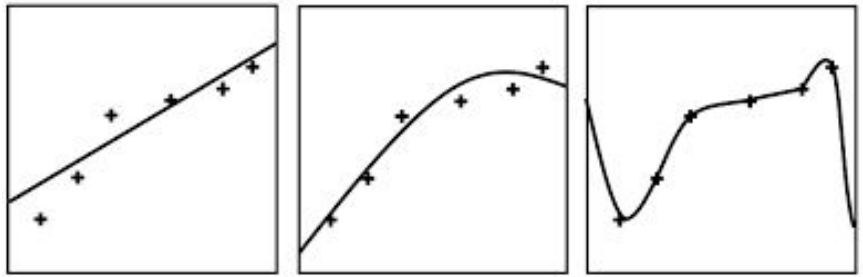
Bias: Error from oversimplifying model (e.g., linear model for curved relation).

Variance: Error from fitting noise (too complex model).

Tradeoff:
Increasing complexity ↓ bias but ↑ variance.

Example: Low-order polynomial underfits a supernova light curve; high-order overfits noise.





Split into Training / Validation / Test Data

Split into Training / Validation / Test Data

Train: Fit the model.

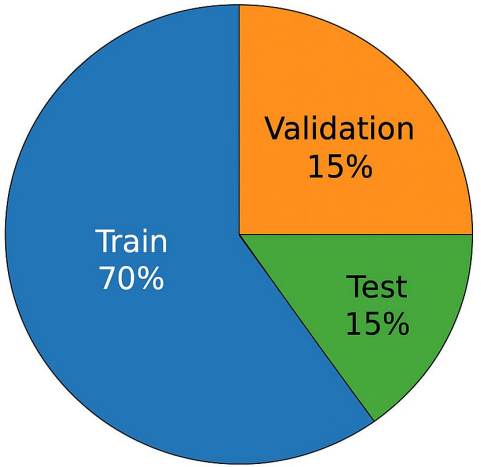
Validation: Tune hyperparameters, pick model complexity.

Test: Final unbiased performance check.

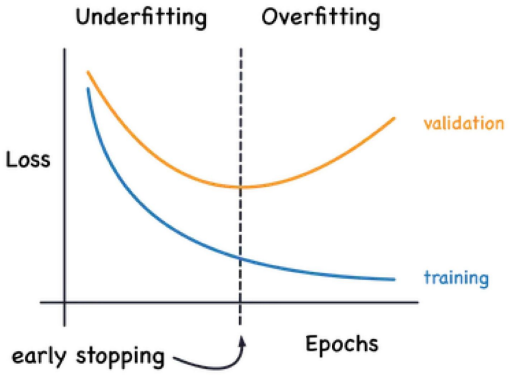
Why: Prevents overfitting and gives a true estimate of generalization.

Beware:

- Data leakage — don't peak at the test data!
- Overfitting on training data!
- Overfitting on validation data!



Credit: ChatGPT



K-fold Cross Validation

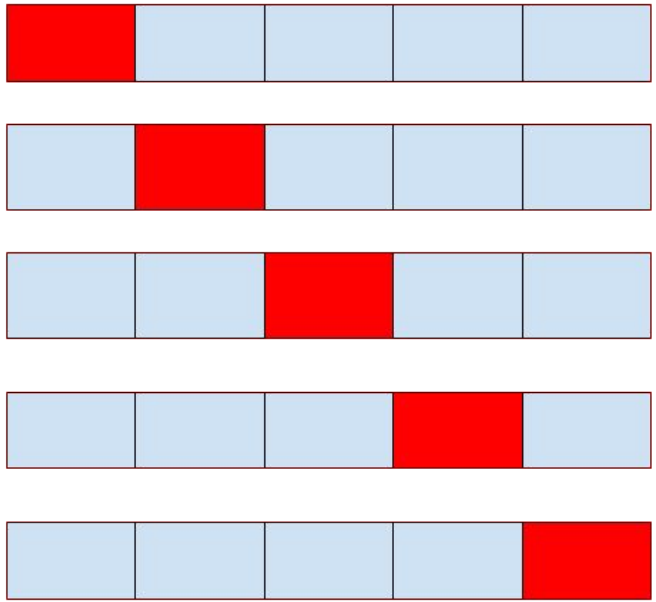
K-fold Cross Validation

Train/Validation/Test split:
Simple but wastes data.

For smaller datasets:

- Split data into k pieces and do k trainings each time leaving out one piece for validation.
- Take average of validation scores to understand generalizability.

This is typically not used for data-rich deep learning.



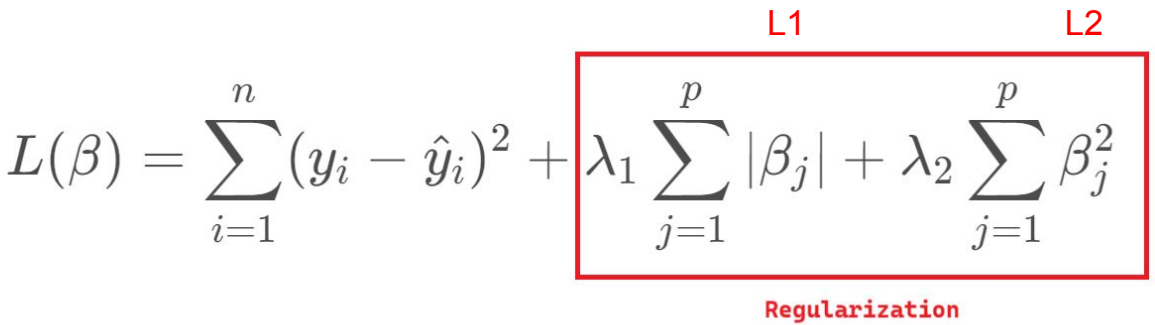
Regularization

Regularization

What: Penalize large weights/parameters to prevent overfitting.

- **L2 (Ridge):** Shrinks coefficients; like a Gaussian prior.
- **L1 (Lasso):** Drives some coefficients to zero; like a Laplace prior.

Bayesian connection: Like putting a prior on the coeffs.



Feature engineering

Feature engineering

Definition: Creating or transforming input variables to improve model performance.

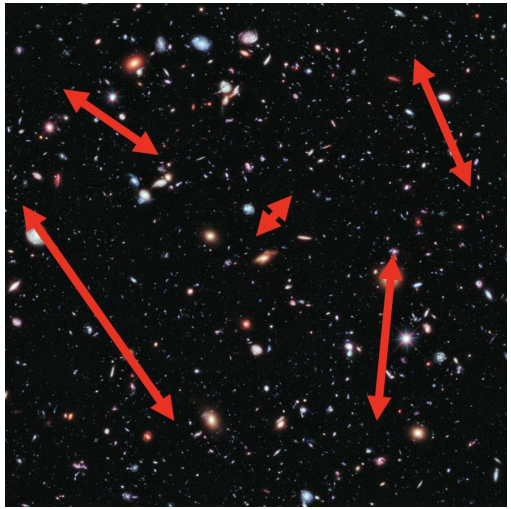
Classical ML: Good features often more important than fancy models.

- E.g.: fluxes → color indices (g–r) for galaxy classification
- Power spectrum in cosmology

Modern ML: Sometimes best to **use raw data directly** (images, spectra, light curves).

- Let the model *learn* the right representations.

Astro lesson: Hand-crafted features can boost performance with small data, but with large surveys and deep nets, raw can win.

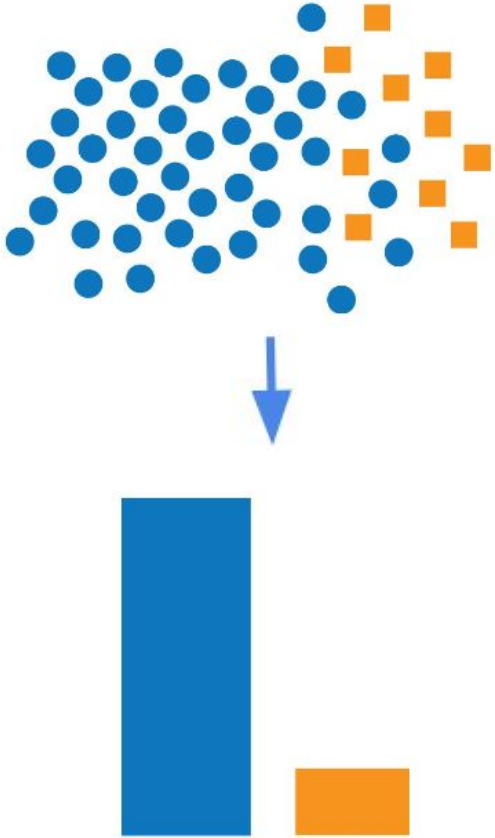


2-pt correlation function
power spectrum

Imbalanced Data

Imbalanced Data

Astro example: Accuracy in a rare-quasar classification (95% negative class). “Good” accuracy can come from predicting all negatives.



Solutions:

Oversample the data to balance.

Use metrics aligned with science goal (e.g. ROC-AUC).

Error Estimation & Robustness

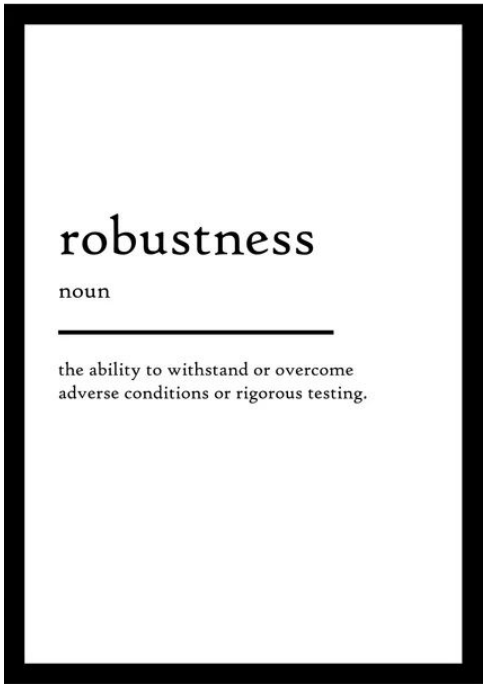
Error Estimation & Robustness

Error bars:

- *Bootstrap:* Resample *with replacement* from the dataset, compute the statistic many times, look at the distribution.
- Jackknife: Leave-one-out (or leave-a-block-out), recompute statistic, and estimate variability.
- *Bayesian Neural Nets*

Robustness tests: Hyperparameter/seed/metric sensitivity, train on different subsets, add noise, withhold features.

Astro example: Null tests in CMB maps; shuffling removes cosmological signal.



Selection Effects / Covariate Shift

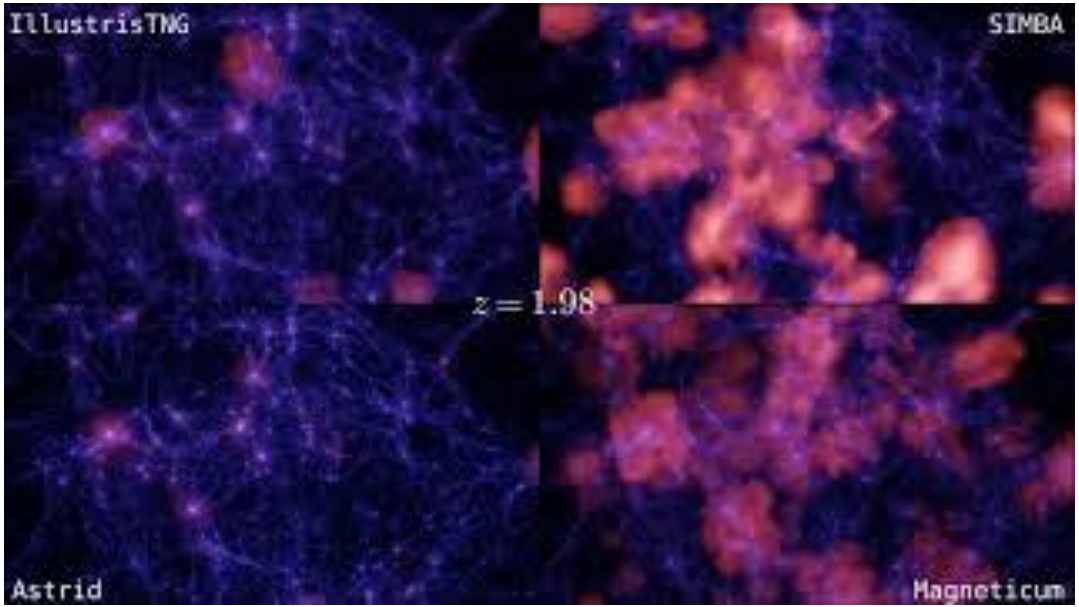
Selection Effects / Covariate Shift

What: Training and target populations differ.

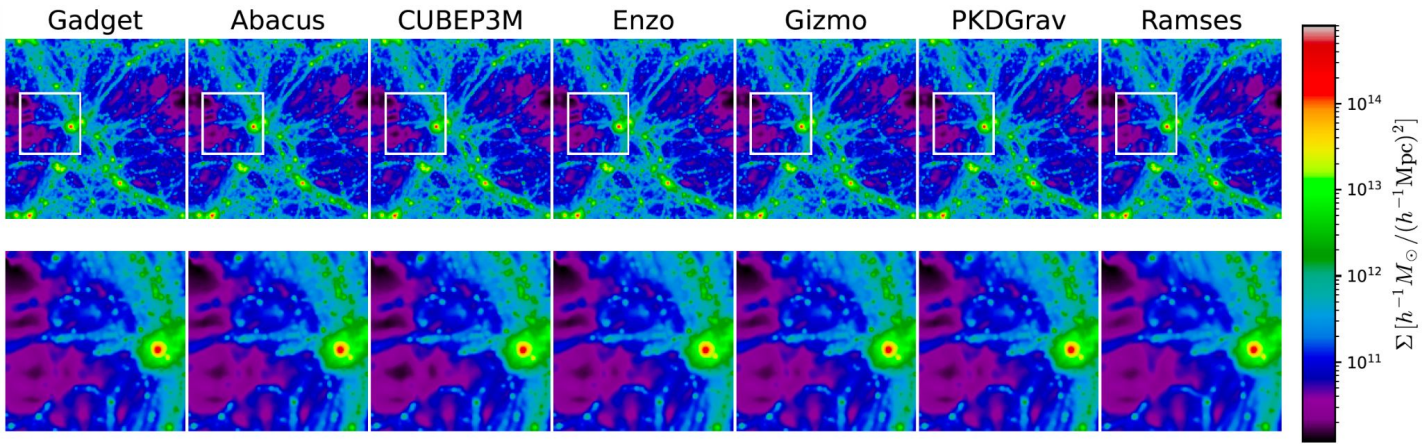
Astro example: Train on bright DESI galaxies, apply to faint LSST galaxies.

Why bad: Model learns bright-galaxy relations that fail on faint galaxies.

Avoid: Ensure representative splits; reweight training set; model selection function.

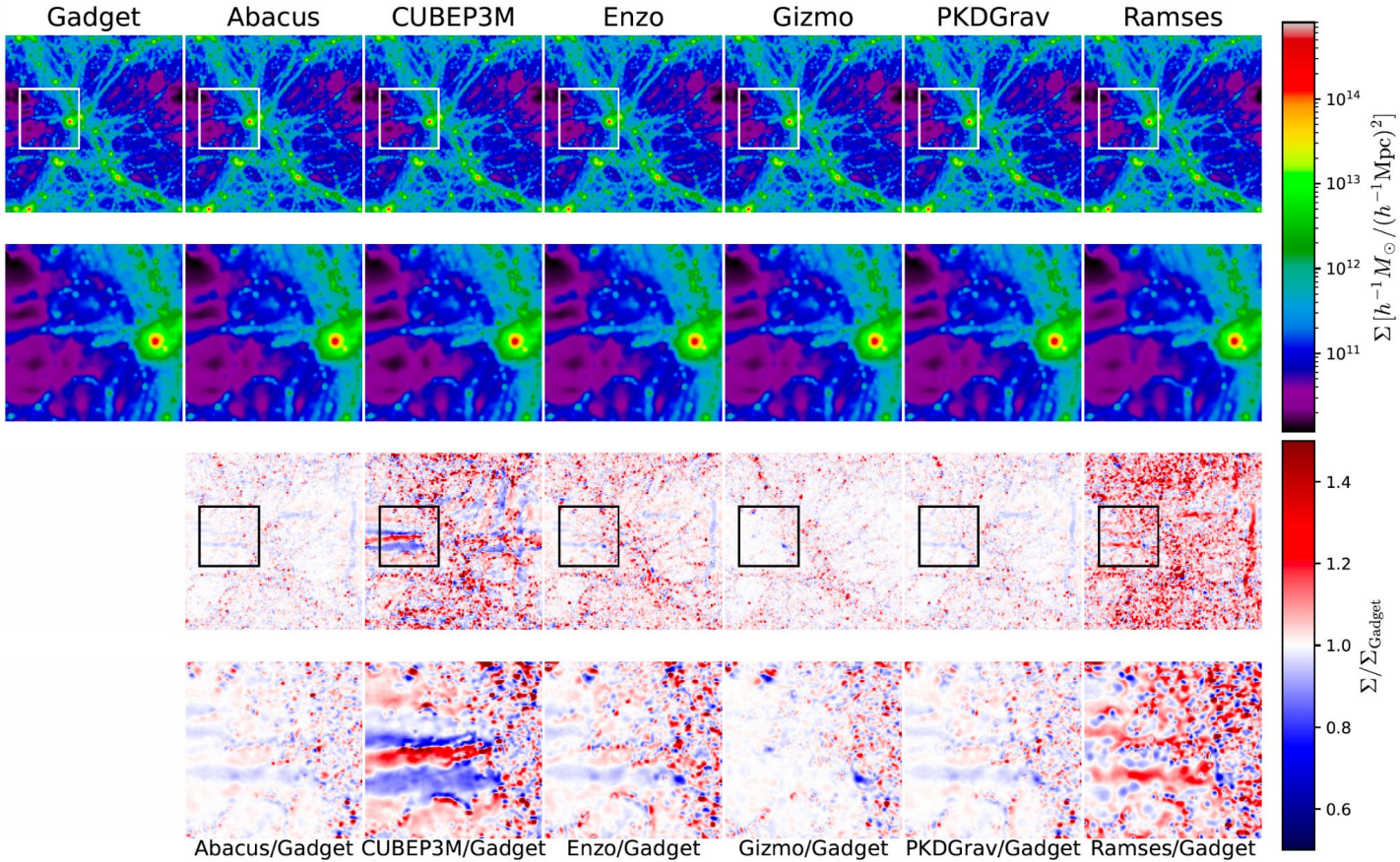


Example: Cosmological Simulations



Bayer+2025

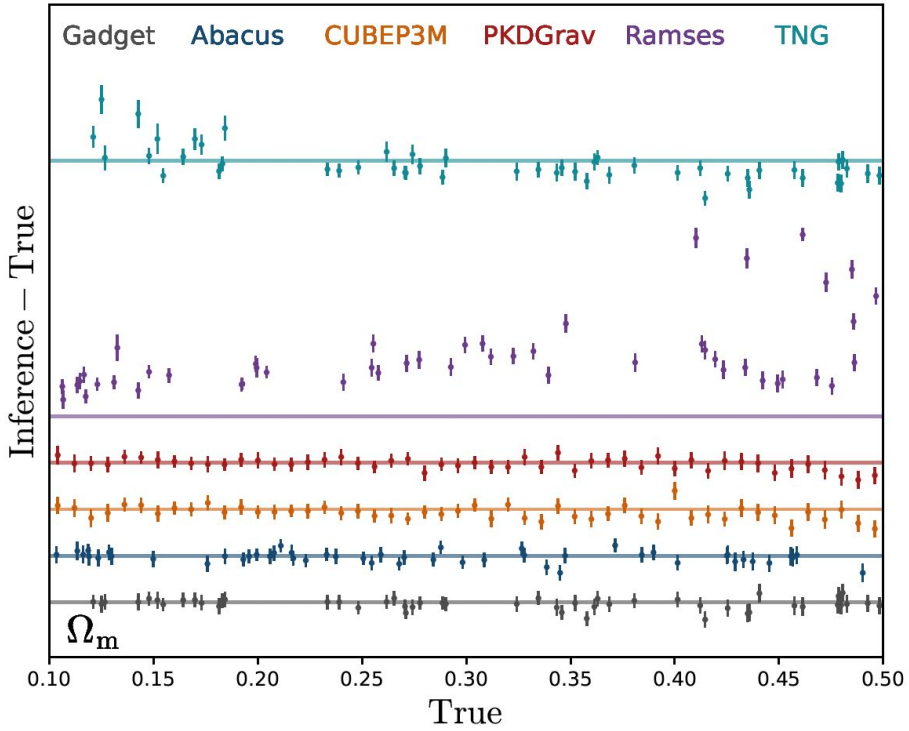
Example: Cosmological Simulations



Bayer+2025

Example: Cosmological Simulations

Train on Gadget, test on everything



Bayer+2025

Interpretability vs Performance

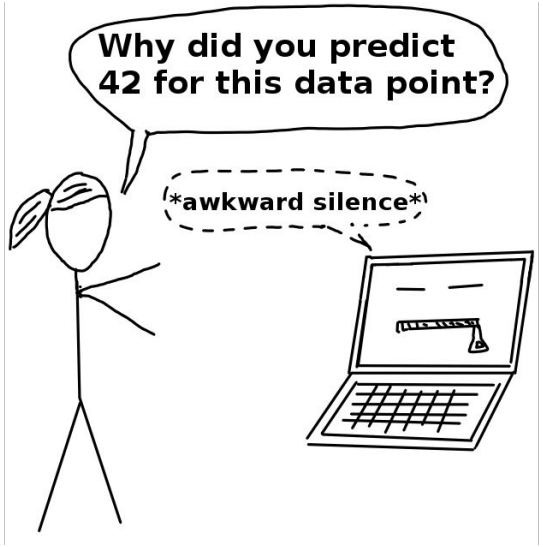
Interpretability vs Performance

Interpretable models: Linear regression — easy to explain, but limited expressiveness.

High-performance models: Deep nets — better accuracy, harder to interpret.

Tradeoff: What is the purpose of science? Interpretability or accuracy? Can we have one without the other?

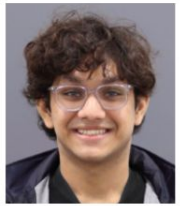
Methods:
Global: feature importance (ablation), coefficients.
Local: SHAP, LIME.
Visual: Saliency maps, integrated gradients, ...



Interpreting CNNs

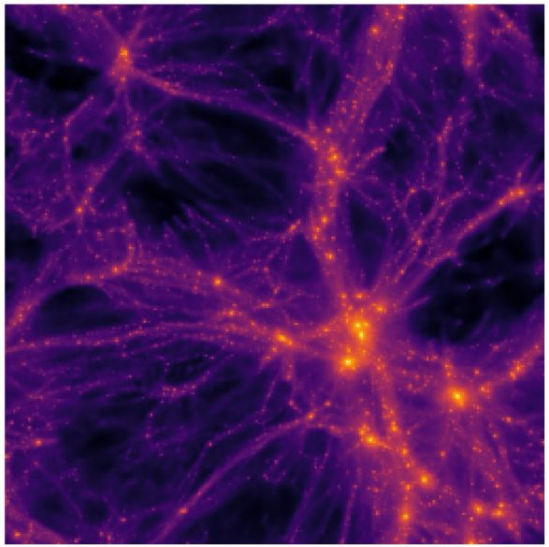
Measure response of network output with respect to input

Malika Golshan
Undergraduate
(Berkeley)



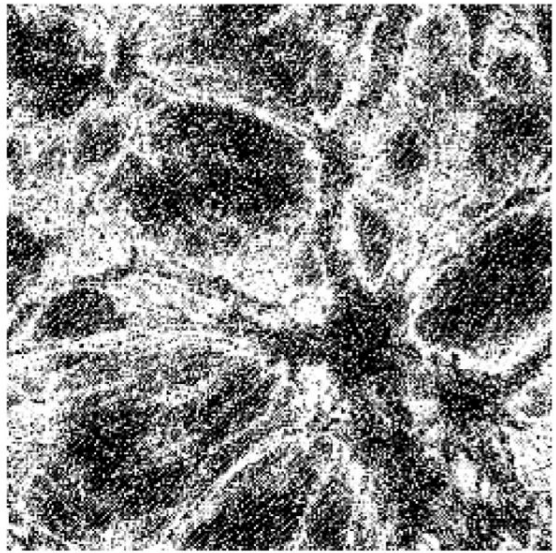
Arnab Lahiry
Graduate Student
(FORTH)

Density Field



Ω_{m}

Integrated Gradient



Golshan & Bayer (2024)
Lahiry, Bayer, Villaescusa-Navarro (2025)

Correlation ≠ Causation (Confounding)

Correlation ≠ Causation (Confounding)

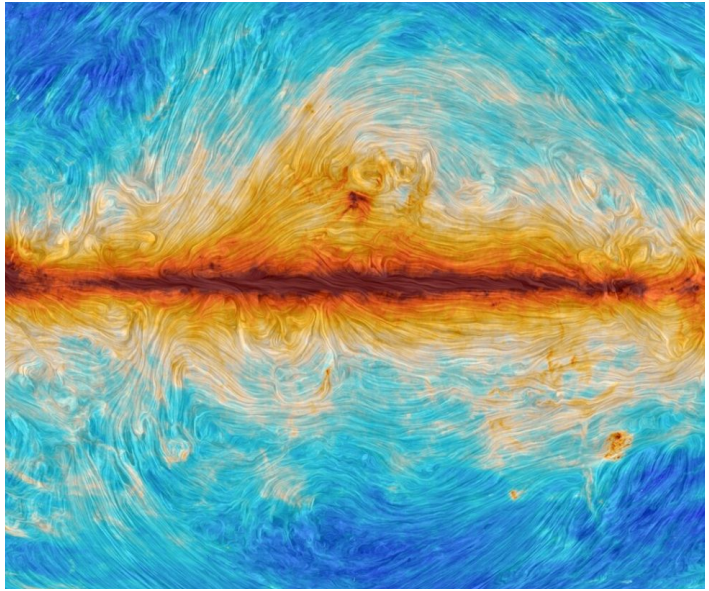
What: Spurious features drive predictions.

Astro Example: Photometric redshift from galaxy colors where model also takes galactic latitude as input:

> *Learns sky position → redshift (due to survey selection), not real colors → redshift.*

Why bad: Fails when applied to new sky regions with different latitude distributions

Avoid: Remove/adjust for confounders; test with **ablation**. Cross-validate on other surveys.



Planck polarization map
More dust in galactic plane, therefore harder to see

Reproducibility

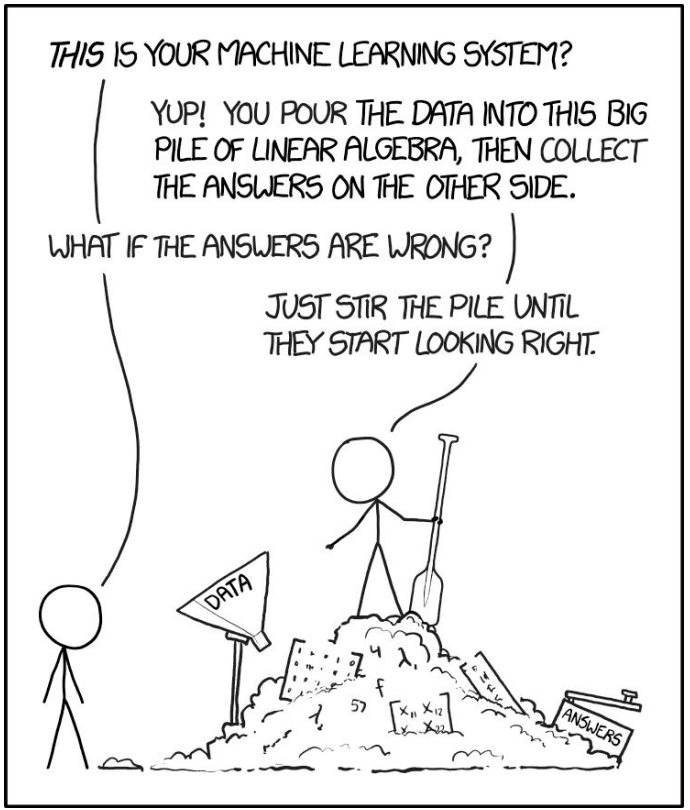
Reproducibility

What: Results change every run.

Astro example: No fixed seeds; library versions not recorded.

Why bad: Can’t trust or reproduce results – what if it was cherry picked.

Avoid: Fix random seeds, version-control code and data.



Spot the Oops Notebook

+

Do you have other *Tricks of the Trade*?

CosmoBench Hack

https://cosmobench.streamlit.app/

https://arxiv.org/abs/2507.03707