A Few ML/AI Basics

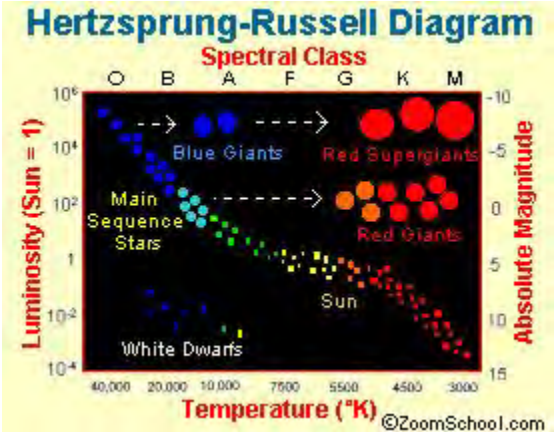
Caltech

ZTF

Ashish Mahabal
AY 119, 2025

Basic concepts

- **Data and Splitting**
- **Evaluation Techniques**
- **Performance Metrics and Measures**
- **Model Concepts**
- **Visualization and Interpretation**



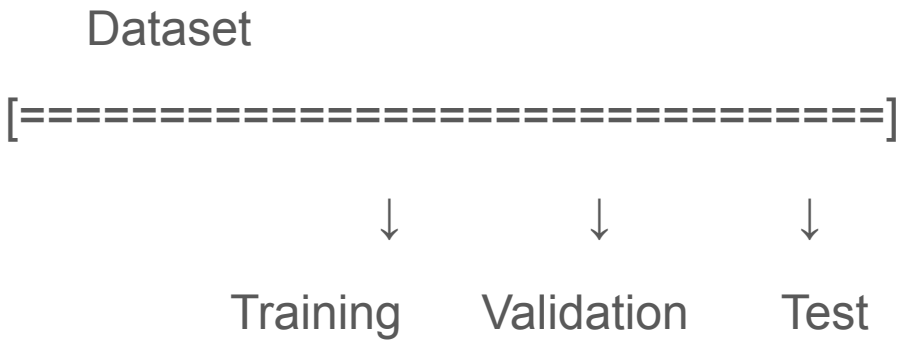
Dataset and Splitting

Dataset

- A dataset is a structured collection of data samples used to build and test models.
- **Structured Data:** tabular format (e.g., CSV files).
- **Unstructured Data:** images, text, audio, etc.

Training, Validation, and Test Sets

- **Training set:** Used by the algorithm to learn and fit parameters.
- **Validation set:** Used during model tuning to optimize hyperparameters.
- **Test set:** Held out entirely from training, used once at the end for unbiased evaluation of model performance.



Cross-Validation

- A method to assess how well your results generalize to an independent dataset.
- Popular method: **k-fold cross-validation** (typically k = 5 or 10).

Steps (example with k=5):

- Split data into 5 equal parts.
- Train model 5 times, each time holding out a different part for validation.
- Average results to get robust performance estimate.

Fold 1: [V] [T] [T] [T] [T]

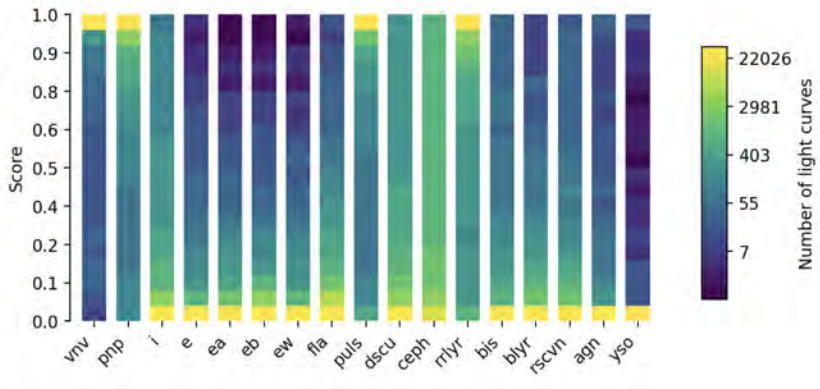
Fold 2: [T] [V] [T] [T] [T]

Fold 3: [T] [T] [V] [T] [T]

Fold 4: [T] [T] [T] [V] [T]

Fold 5: [T] [T] [T] [T] [V]

(V = Validation, T = Training)

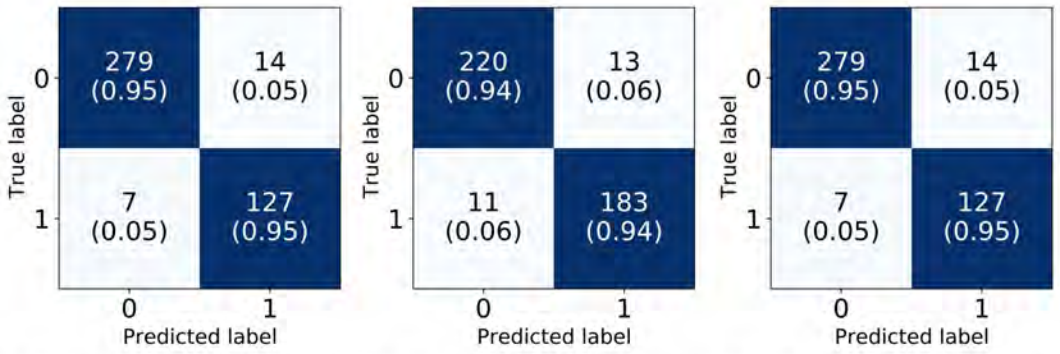


ZTF SCOPE multiple binary classifiers

Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

- Used for evaluating classification problems.
- Shows true vs. predicted class labels clearly.
- Metrics derived from Confusion Matrix:
  - **Accuracy**: (TP + TN) / Total
  - **Precision**: TP / (TP + FP)
  - **Recall (Sensitivity)**: TP / (TP + FN)
  - **F1 Score**: harmonic mean of Precision and Recall.



(a) Glitch-versus-all confusion matrix. (b) NS-versus-all confusion matrix. (c) BBH-versus-all confusion matrix.

With Chen, Nayyer (GW collab.)

Loss Function

- Measures how well the model's prediction matches the true data.
- Common examples:
    - **Mean Squared Error (MSE)**: For regression.
    - **Cross-Entropy Loss**: For classification.

Examples:

- **MSE:** MSE= $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$
- **Cross-Entropy Loss:** $L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$

For binary Cross Entropy:

Suppose your model outputs y'=0.9 for a positive example (y=1).

CE Loss = $-[1 \cdot \log(0.9)+0 \cdot \log(0.1)]=0.105$

If instead, your model incorrectly predicts y'=0.1,

CE Loss = $-[1 \cdot \log(0.1)+0 \cdot \log(0.9)]=2.303$

Thus, incorrect and confident predictions incur much higher penalties.

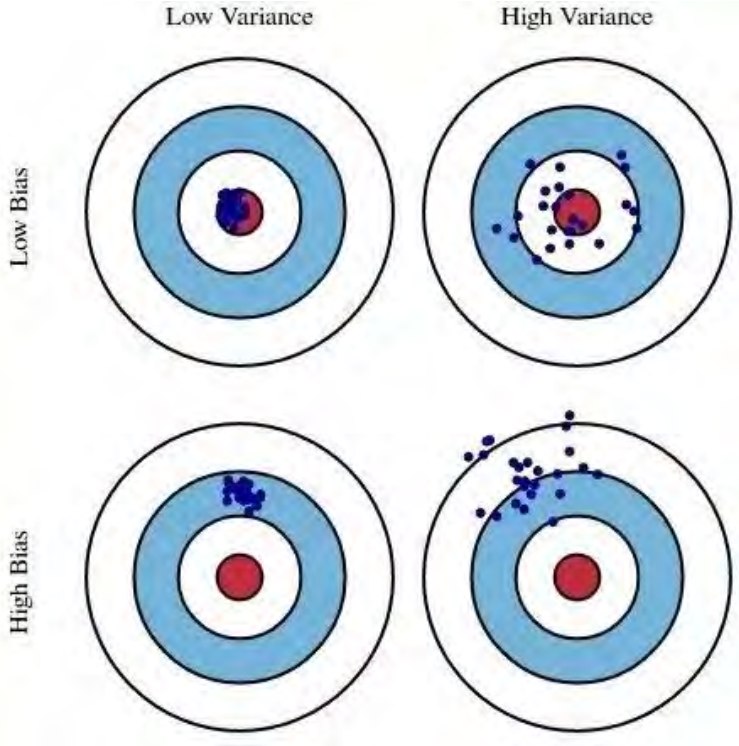
Bias and Variance

- **Bias:** Error due to overly simplistic models (underfitting).
- **Variance:** Error due to overly complex models sensitive to fluctuations in training data (overfitting).

Goal: Achieve balance (Bias-Variance Trade-off).

Model Complexity →

High Bias (underfit) Optimal Balance High Variance (overfit)



https://www.quora.com/What-does-Bagging-reduces-the-variance-while-retaining-the-bias-mean

Regularization

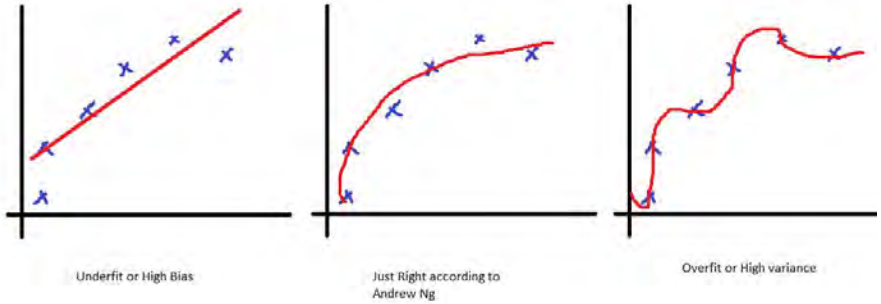
- Technique to reduce overfitting.
- Penalizes model complexity by adding a regularization term.

Common examples:

- **L1 (Lasso)** penalizes absolute magnitude of coefficients.
- **L2 (Ridge)** penalizes squared magnitude of coefficients.

Regularization Example (Linear Regression):

Loss (L2)=MSE + $\lambda \sum_{i=1}^{n} w_i^2$



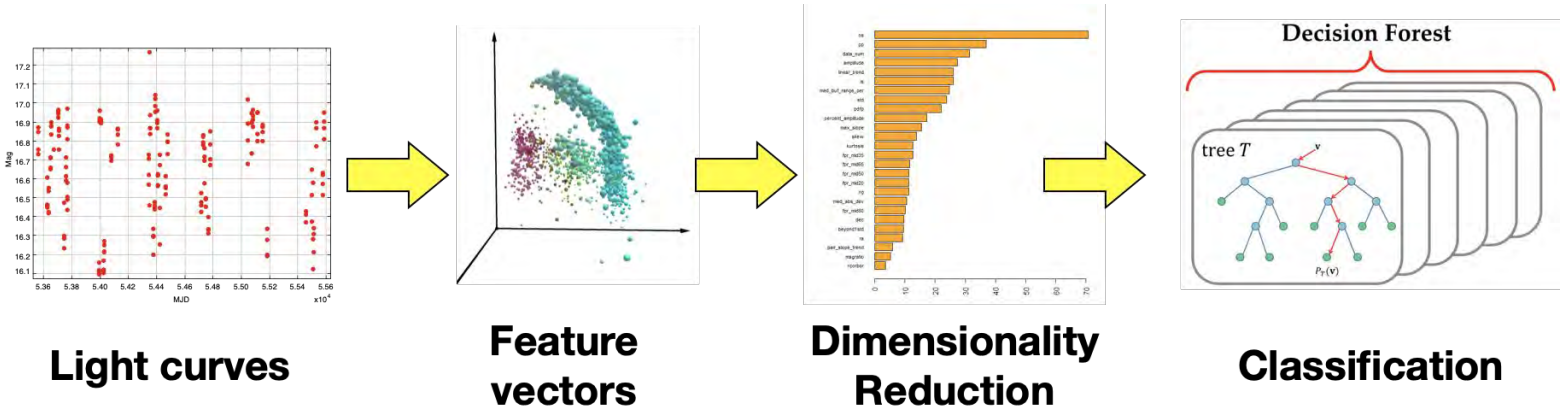
https://towardsdatascience.com/regularization-what-why-when-and-how-d4a329b6b27f/

Feature importance and selection

- Identifying and using only the most informative variables.
- Reduces overfitting, improves interpretability.

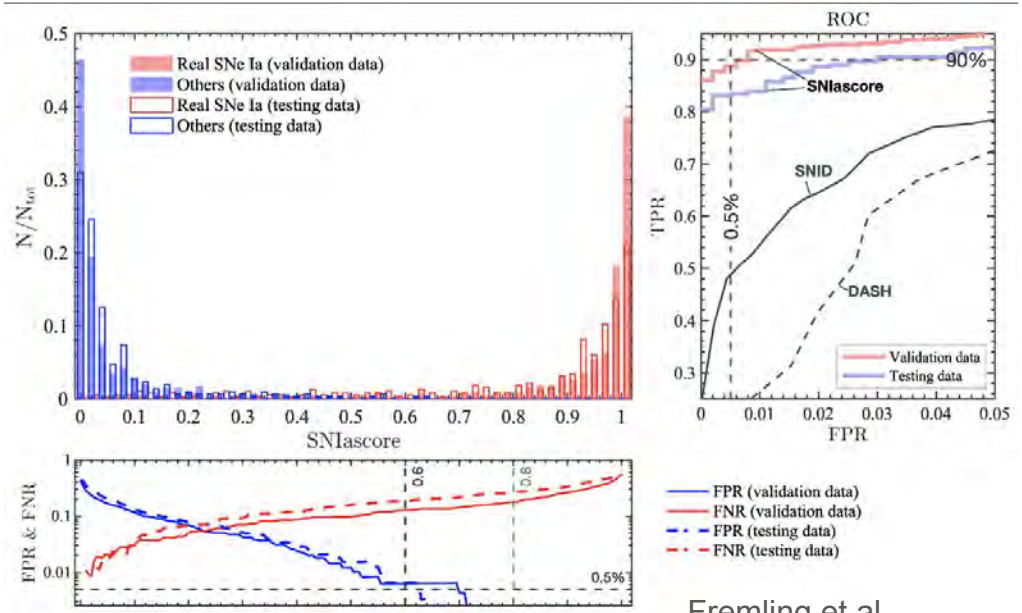
Common approaches:

- Statistical tests (Chi-square, t-tests)
- Model-based importance (e.g., Random Forest)



ROC Curve and AUC

- **Receiver Operating Characteristic (ROC)**: Graph of True Positive Rate (Recall) vs. False Positive Rate.
- **Area Under Curve (AUC)**: Summarizes ROC in single number (0.5 random guess, 1 perfect classifier).



Fremling et al.

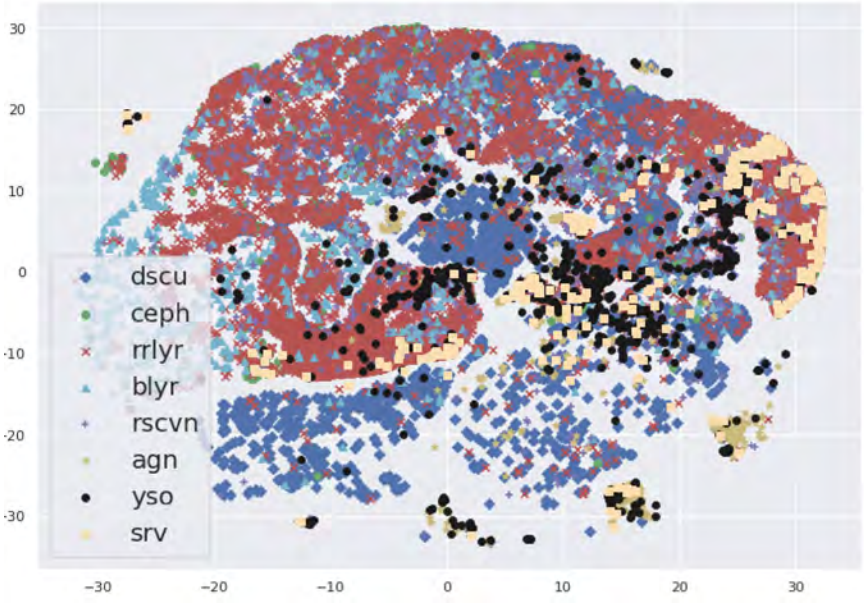
Feature Space Visualization (Dimensionality Reduction)

Visualizing high-dimensional data in lower dimensions for intuitive understanding.

Principal Component Analysis (PCA) commonly used method.

t-SNE

UMAP



t-NSE SCOPE ZTF variables

Practicalities

GitHub

Editor (VSCode?)

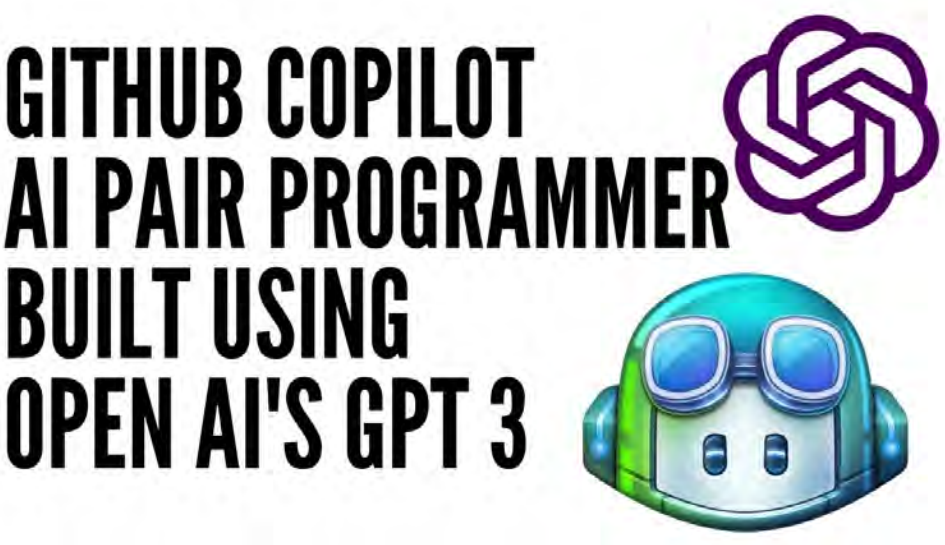
Editor + GitHub

(Co-Pilot)

Create a private GitHub repo

Explore some of the concepts discussed

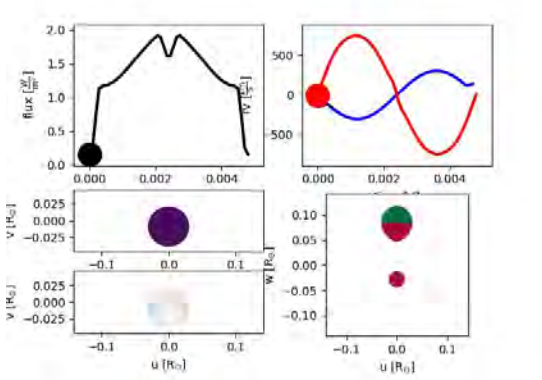
Share it with us on GitHub (AshishMahabal)



In a galaxy long long ago

Next two talks

Supervised and Unsupervised Classification



Credit: Kevin Burge