A Brief Introduction to Hyperparameter Optimization

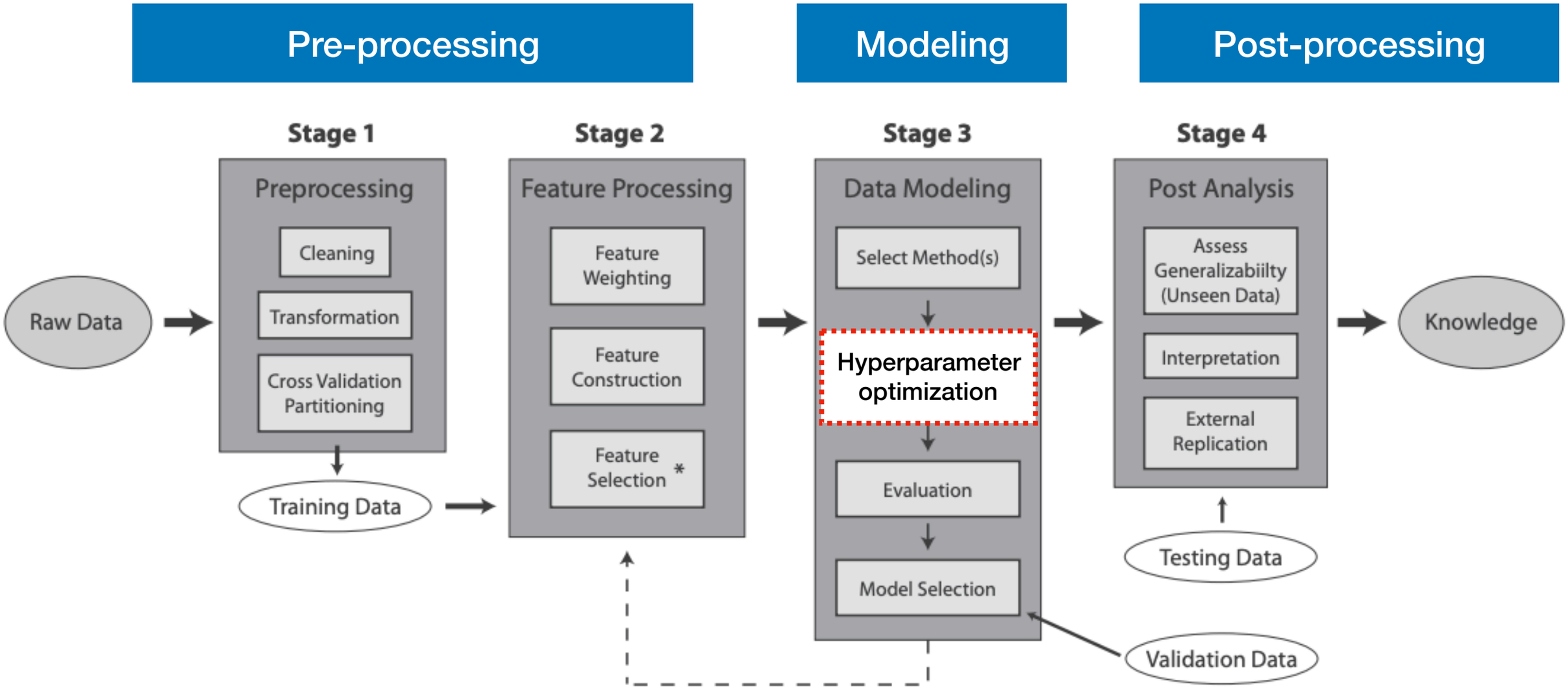
* with a focus on medical data

Jill Cates

PyDataDC

November 18, 2018

A Typical ML Pipeline



R.J. Urbanowicz et al. 2018

Bad Hyperparameters = Bad Model = Bad Predictions

Case Study

Sepsis Prediction

Defining Sepsis

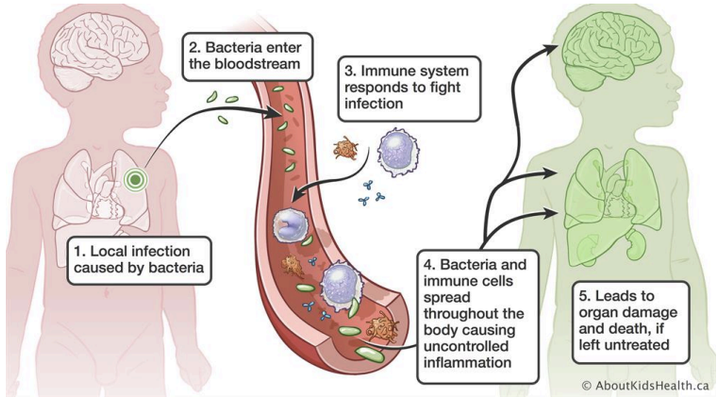
What is sepsis?

“life-threatening condition that arises when the body's response to infection causes injury to its own tissues and organs” [1]

750,000 patients are diagnosed with severe sepsis in the United States each year with a 30% mortality rate [2]

costs $20.3 billion each year ($55.6 million per day) in U.S. hospitals [3]

every hour that passes before treatment begins, a patients’ risk of death from sepsis increases by 8% [4]



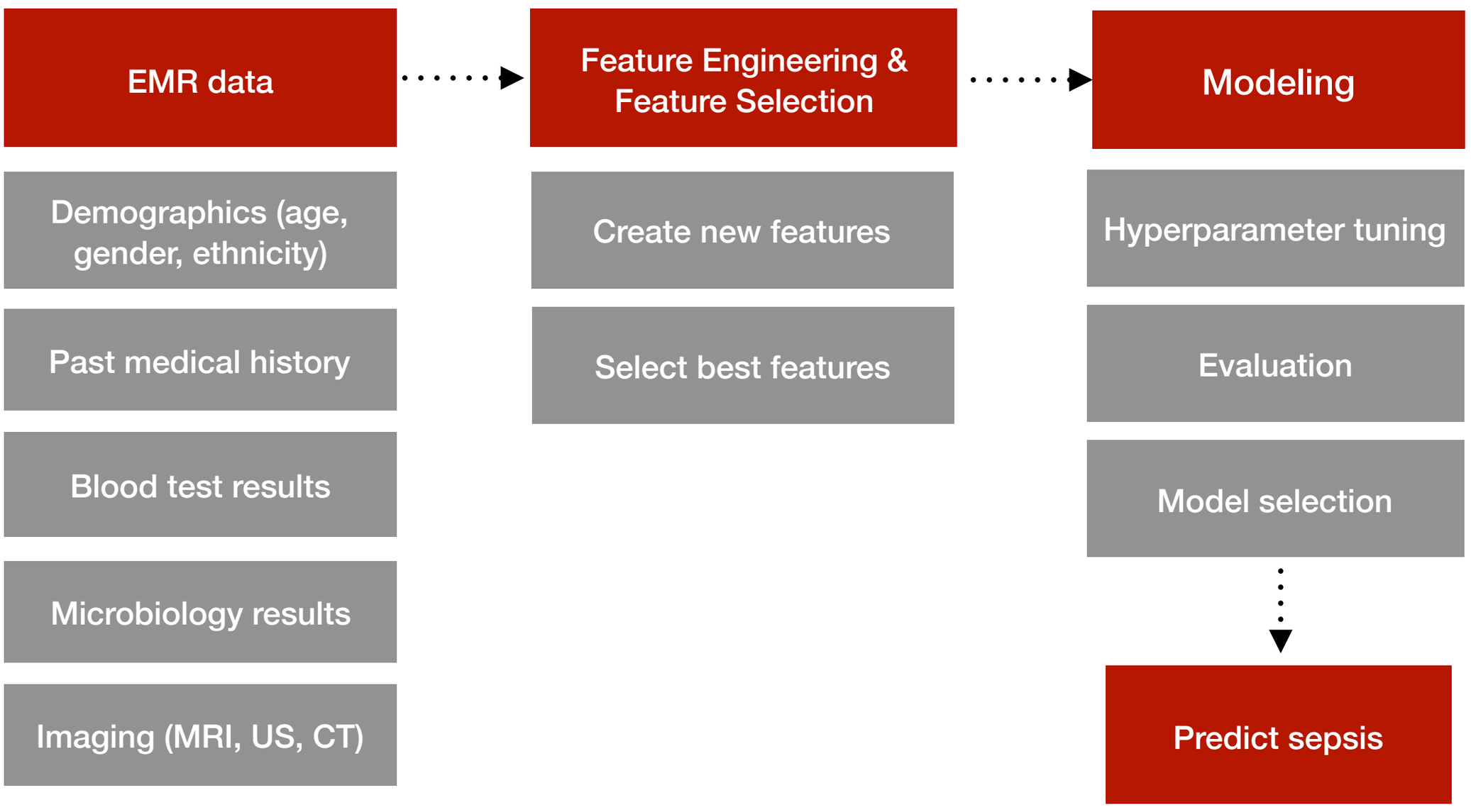
KILLER BUG Teenage boy nearly dies from popping a pimple – after it triggered killer sepsis infection

> Geraint squeezed a spot on his neck, not realising that he'd end up fighting for his life

Proposal

Build a model that predicts a patient's likelihood of getting sepsis

An Overview of Our Pipeline



Our Data

50,000 hospital admissions and 40,000 patients

Data	Description
Admissions information	Diagnosis upon admission, time of admission/discharge
Patient demographics	Age, gender, religion, marital status
Prescriptions	Which drugs were they prescribed and when?
Unit transfers	Did they move from the medical ward to ICU?
Vital signs	Heart rate, blood pressure, respiratory rate, spO2
Lab results	Blood tests, urine tests
Diagnoses	ICD-10 codes
Chest X-ray images	DICOM format

Data Pre-processing

Clean up inconsistencies in medical terms

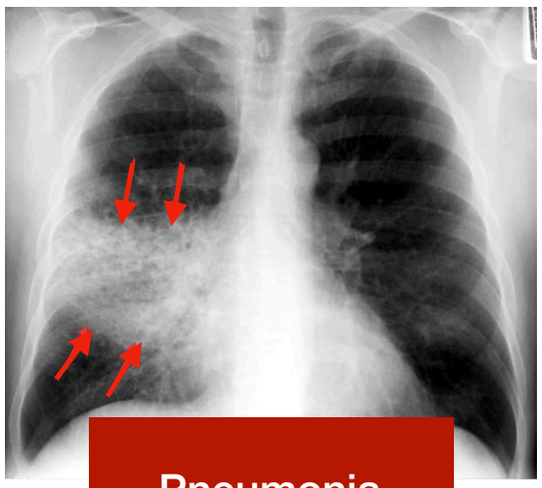
- Aspirin vs. ASA (acetylsalicylic acid)
- NS (normal saline) vs. 0.9% sodium chloride

Unified Medical Language System

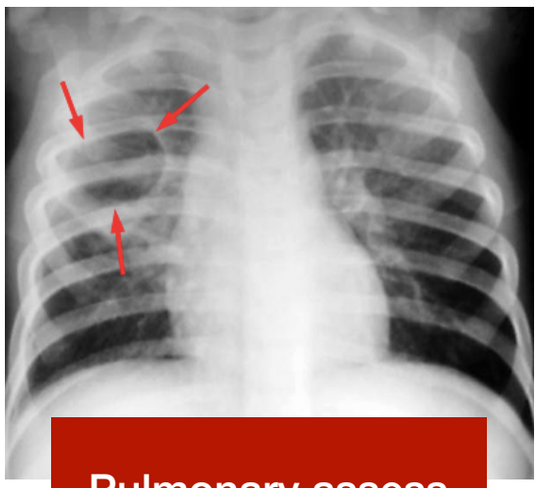
Generate new features from imaging data

- identify lung opacities in X-ray image
- lung_abnormality = (0,1)
- infection_size = [x,y,width,height]

This is a separate model in itself!



Pneumonia



Pulmonary assess

National Institutes of Health

NIH CXR dataset contains +100,000 annotated X-ray images

Creating a sepsis score

How do we identify sepsis in a patient?

- ICD-10 codes [4], [5]:
  - Bacteremia - `R78.81`
  - Sepsis unspecific - `A41.9`
  - Acute hepatic failure without coma - `K72.00`

* International Statistical Classification of Diseases and Related Health Problems (ICD), 10th revision, developed by the World Health Organization (WHO)
* ICD codes are listed for billing patients at end of stay

icd9_code	short_title
9694	Pois-benzodiazepine tran
72888	Rhabdomyolysis
4139	Angina pectoris NEC/NOS
30560	Cocaine abuse-unspec
E8532	Acc poisn-benzdiaz tranq
4660	Acute bronchitis
2967	Bipolor I current NOS
49390	Asthma NOS
412	Old myocardial infarct
4019	Hypertension NOS

- Severity scores based on lab results and vitals:
  - SOFA: Sequential Organ Failure Assessment [6]
  - SIRS: Systemic Inflammatory Response Syndrome [7]
  - LODS: Logistic Organ Dysfunction System [8]

Creating a sepsis score

SOFA: Sequential Organ Failure Assessment

mortality prediction score that is based on the degree of dysfunction of six organ systems

Table 1
The Sequential Organ Failure Assessment (SOFA) score

SOFA score	1	2	3	4
Respiration[a]				
PaO_2/FIO_2 (mm Hg)	<400	<300	<220	<100
SaO_2/FIO_2	221-301	142-220	67-141	<67
Coagulation				
Platelets $\times 10^3/mm^3$	<150	<100	<50	<20
Liver				
Bilirubin (mg/dL)	1.2-1.9	2.0-5.9	6.0-11.9	>12.0
Cardiovascular[b]				
Hypotension	MAP <70	Dopamine ≤5 or dobutamine (any)	Dopamine >5 or norepinephrine ≤0.1	Dopamine >15 or norepinephrine >0.1
CNS				
Glasgow Coma Score	13-14	10-12	6-9	<6
Renal				
Creatinine (mg/dL) or urine output (mL/d)	1.2-1.9	2.0-3.4	3.5-4.9 or <500	>5.0 or <200

vitals → PaO_2/FIO_2, SaO_2/FIO_2

blood test results → Bilirubin

urine test results → Glasgow Coma Score

Jones et al. 2010. Crit Care Med.

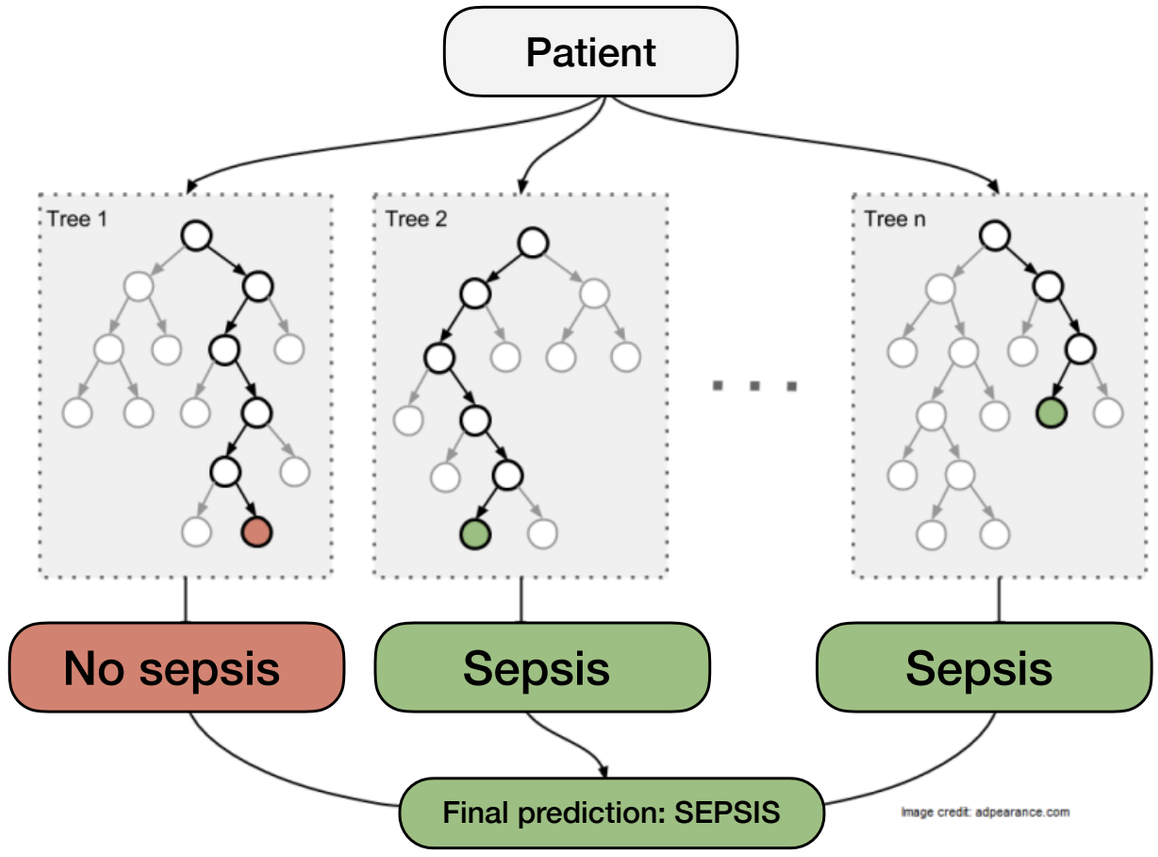
Sepsis = acute change in total SOFA score ≥ 2 points upon infection (regardless of baseline) [9]

Picking a Model

A binary classification problem

admission_id	sepsis
1001	0
1002	1
1003	0
1004	1

Random Forest Classifier



prob=0.667

Output

A probability score between 0 and 1
representing a patient's likelihood of sepsis

No Free Lunch Theorem

"all optimization problem strategies perform equally well when averaged over all possible problems"

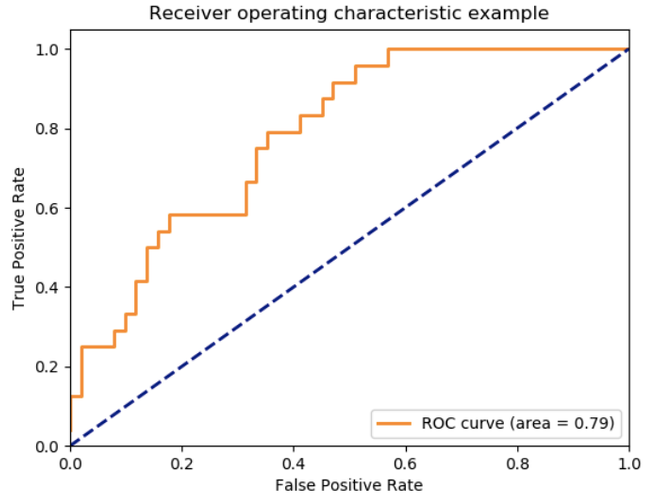
(See Seinfeld's Soup Nazi episode)

Evaluating the Quality of Our Model

$$RMSE = \sqrt{\frac{\Sigma_{i=1}^{N}(y - \hat{y})^2}{N}}$$

$$recall = \frac{TP}{TP + FN}$$

Area Under the Receiver Operating Curve (AUROC)

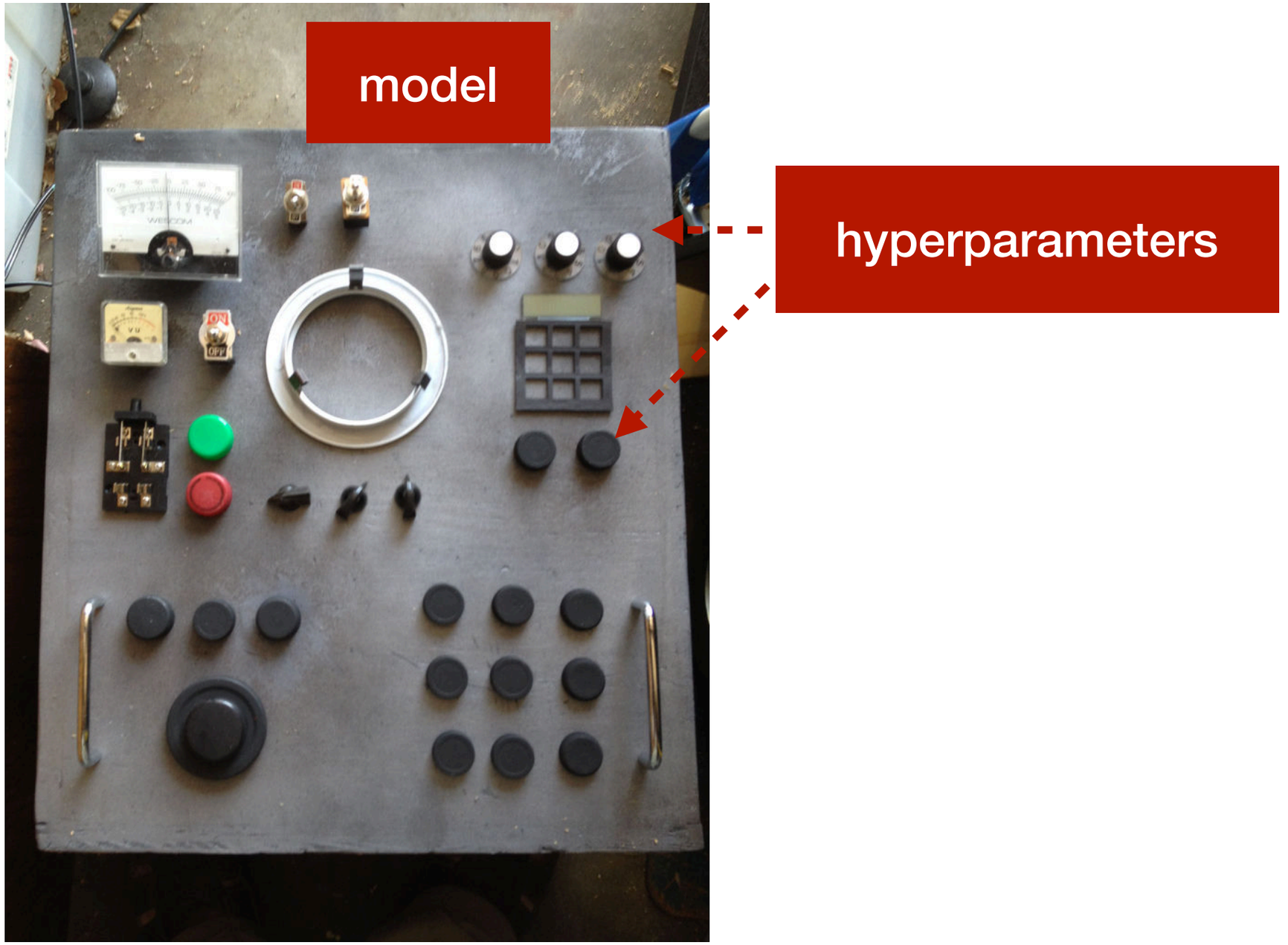


$$precision = \frac{TP}{TP + FP}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Hyperparameter Tuning

What is a hyperparameter?



Configuration that is external to the model
Set to a pre-determined value before model training

What is a hyperparameter?

Example: clinical trials

active ingredients

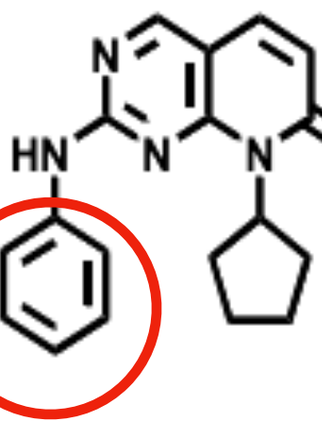
concentrations

Did it cure the patient?

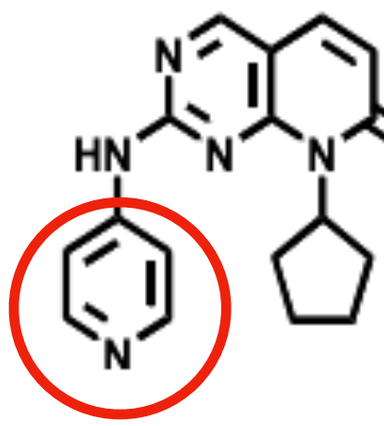
goal: maximize drug effectiveness

What is a hyperparameter?

Example: drug discovery

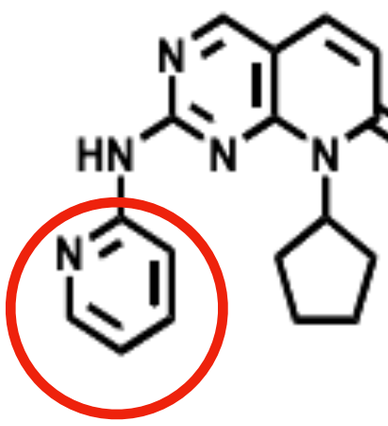


0174413
Cdk4/D: 0.210 μM
Cdk2/A: 0.012 μM



0204661
Cdk4/D: 0.092 μM
Cdk2/A: 0.002 μM

Toxic



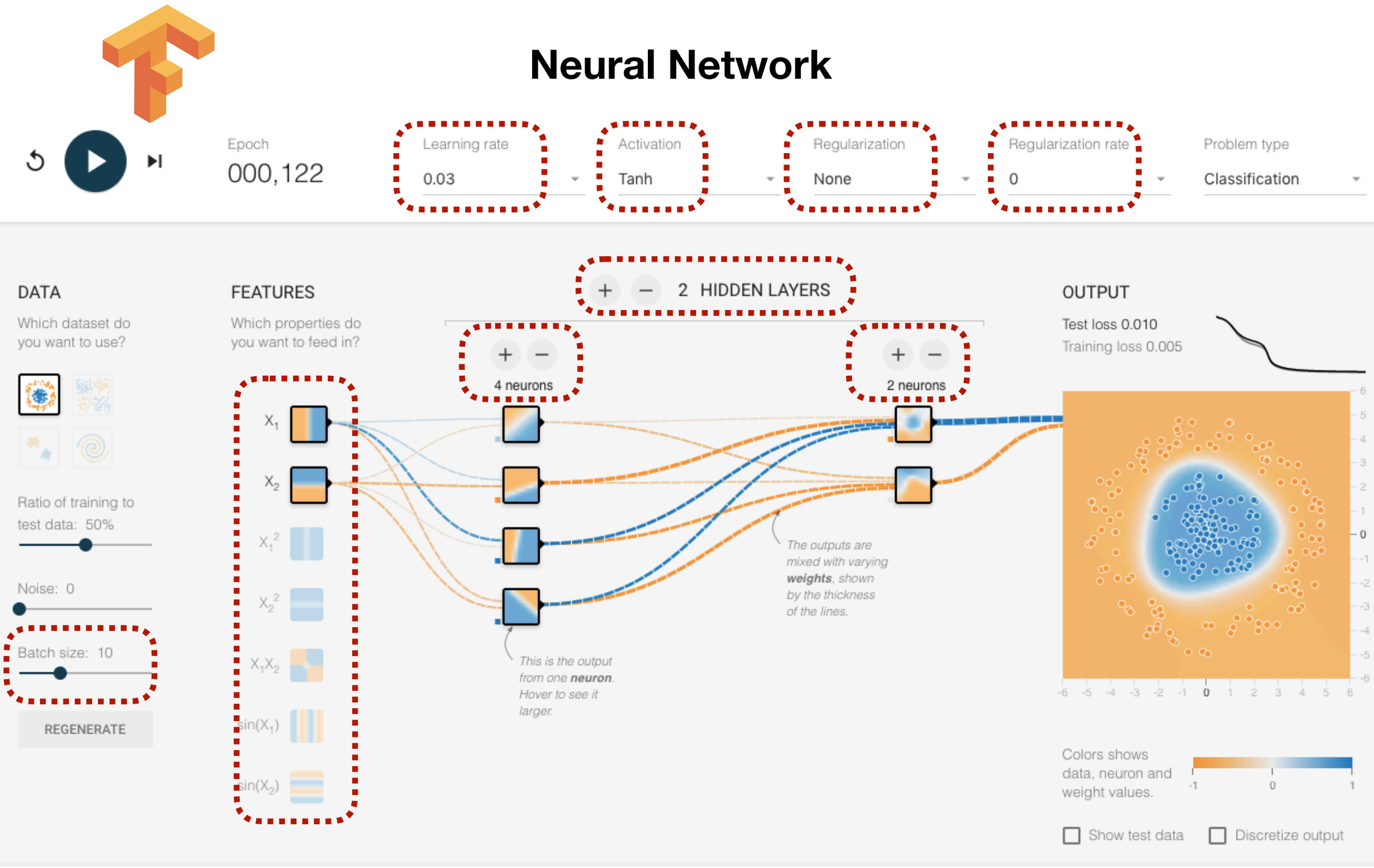
0205783
Cdk4/D: 0.145 μM
Cdk2/A: 5.010 μM

Therapeutic

Hyperparameter Examples

- **Random Forest Classifier**
  - n_estimators (# of decision trees)
  - max_depth
- **Singular Value Decomposition**
  - n_components (# latent factors)
- **Support Vector Machine**
  - Regularization (C)
  - Tolerance threshold (ℰ)
  - Kernel
- **Gradient descent**
  - Learning rate
  - Regularization (λ)
- **K-means clustering**
  - K clusters

Hyperparameter Examples



https://playground.tensorflow.org

Our Hyperparameters

Random Forest Classifier

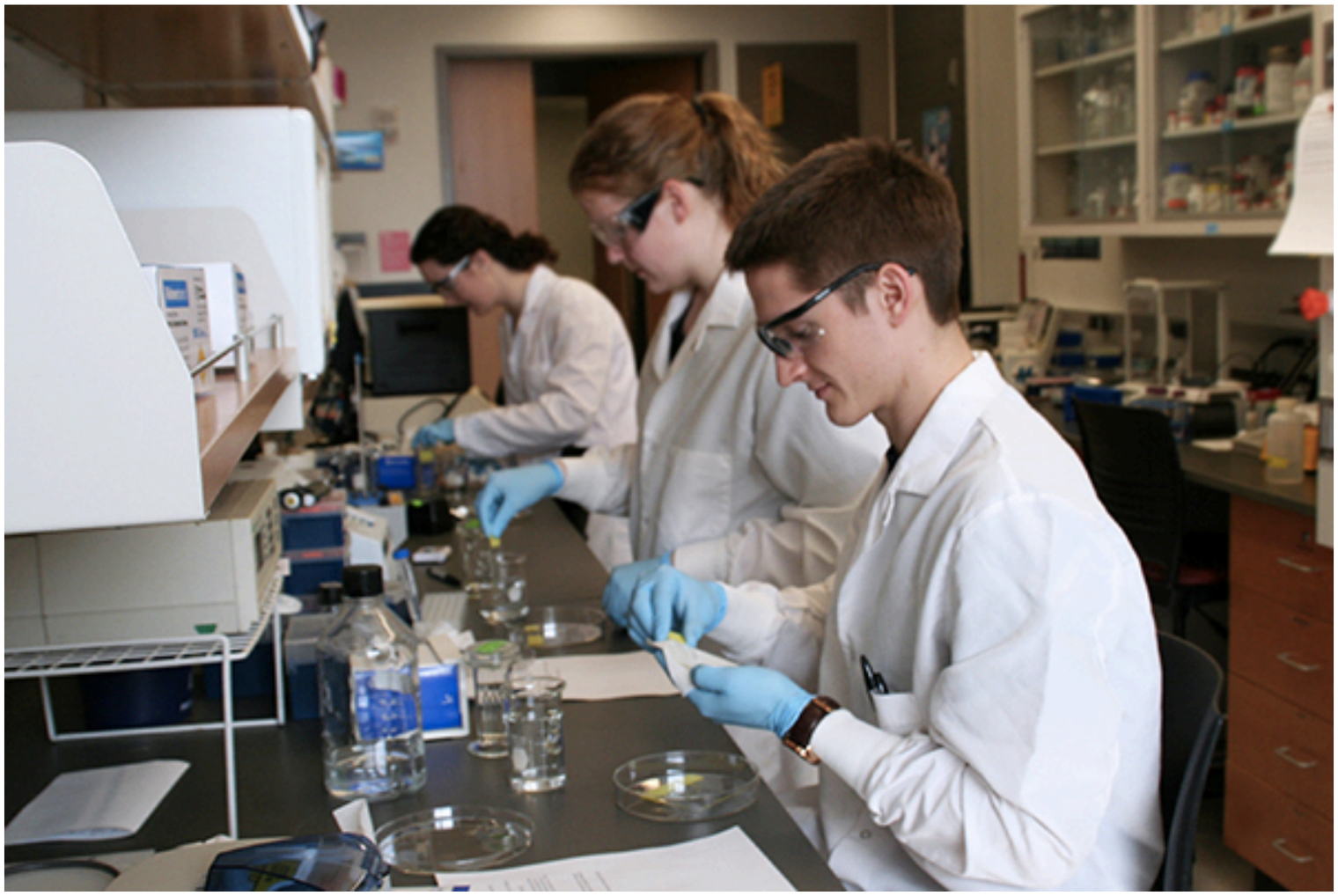
- n_estimators (number of decision trees)
- max_depth (maximum tree depth)

Sampling Techniques

1. Grad Student Descent
2. Grid Search
3. Random Search
4. Sequential Model-based Optimization

“Grad Student” Descent

a.k.a. tinkering until you get decent results



Grid Search

`skelarn.ensemble.RandomForestClassifier()`

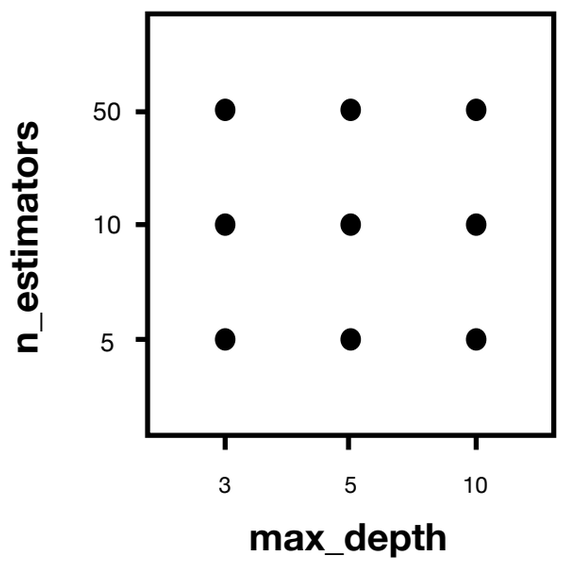
Search Space

- `n_estimators = [5,10,50]`
- `max_depth = [3,5]`

Provide discrete set of hyperparamter values

Models

1) `n_estimators=5, max_depth=3`
2) `n_estimators=5, max_depth=5`
3) `n_estimators=10, max_depth=3`
4) `n_estimators=10, max_depth=5`
5) `n_estimators=50, max_depth=3`
6) `n_estimators=50, max_depth=5`



Random Search

Random Search for Hyper-Parameter Optimization

James Bergstra JAMES.BERGSTRA@UMONTREAL.CA
Yoshua Bengio YOSHUA.BENGIO@UMONTREAL.CA
Département d'Informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC, H3C 3J7, Canada

"for most data sets only a few of the hyper-parameters really matter…"

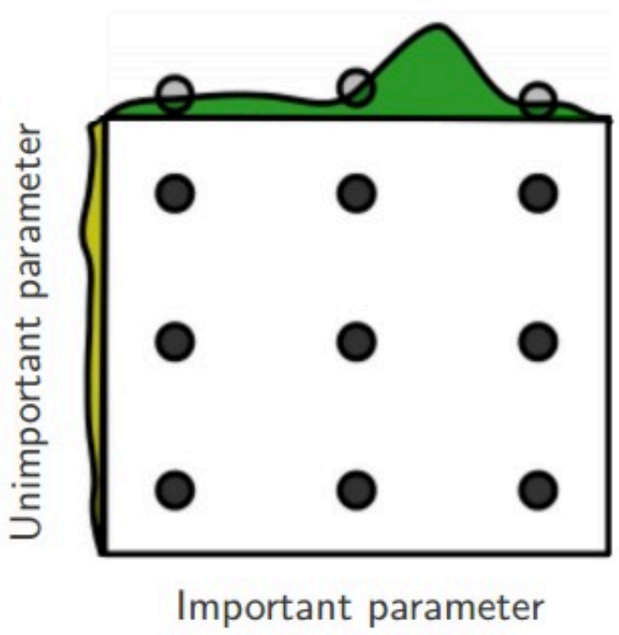
"…different hyper-parameters are important on different data sets"

- Based on assumption that not all hyperparameters are equally important
- Works by sampling hyperparamater values from a distribution

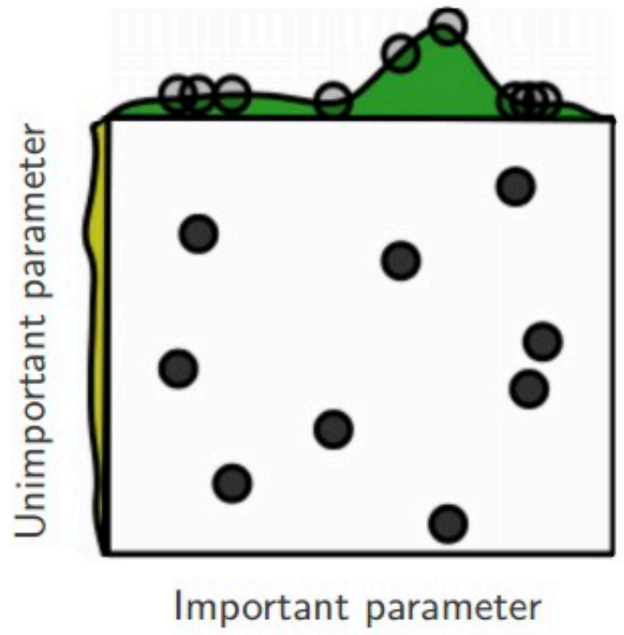
Random Search

A visual explanation of why random search can be better

Grid Search



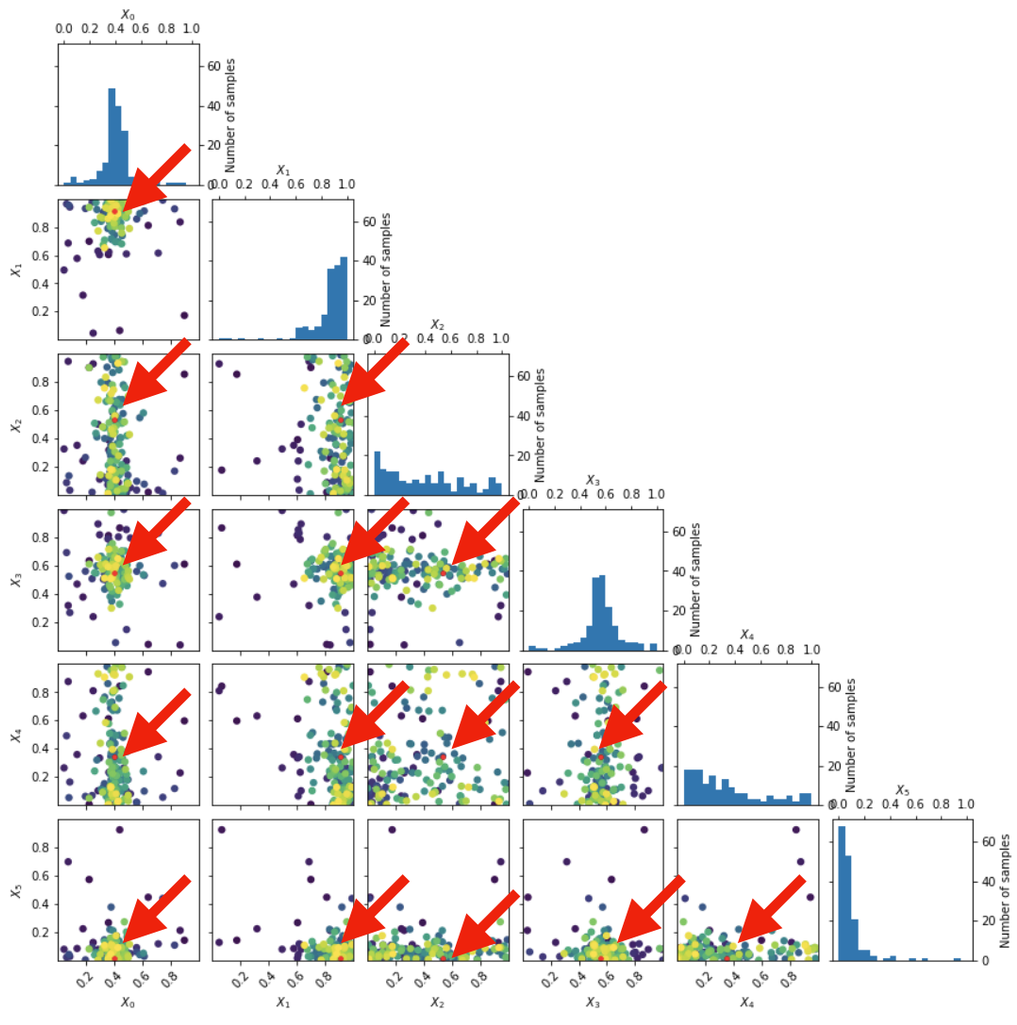
Random Search



Sequential Model-Based Optimization

Keeps track of previous iteration results

scikit-optimize (skopt)
hyperopt
Metric Optimization Engine (MOE)

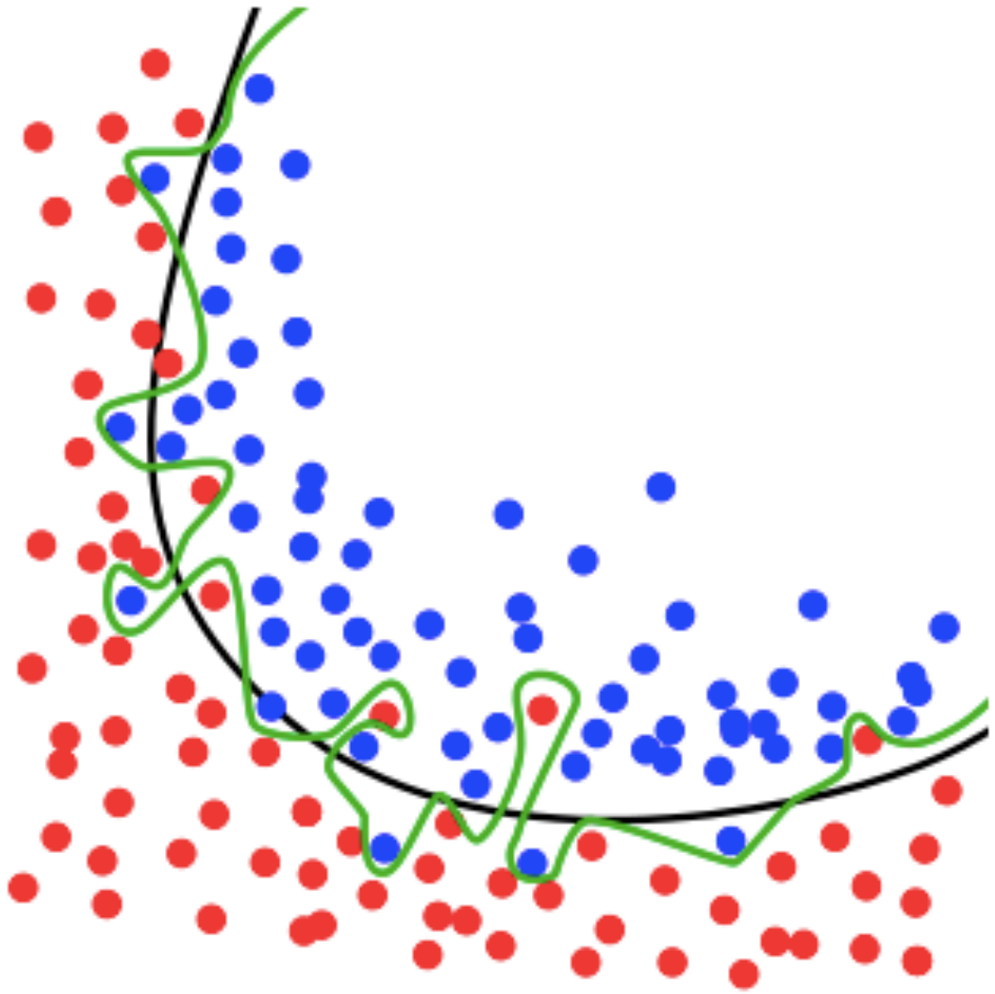


Which sampling technique is best?

No Free Lunch Theorem

“all optimization problem strategies perform equally well when averaged over all possible problems”

What is Overfitting?!



The Bias-Variance Trade-off

Learning from noise vs. signal

Model is tightly bound to training set

How to detect overfitting

High performance on training set

Poor performance on test set

How to Prevent Overfitting

- Consider an ensemble model
- Regularization
- Cross-validation
- Occam’s Razor

Regularization

A penalty term

$$\lambda$$

L1 norm (Lasso Regression)

Good for feature selection
Sets weight of irrelevant features to 0

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

L2 norm (Ridge Regression)

Handles multicollinearity
Reduces weight of less important features

$$\mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$

ElasticNet

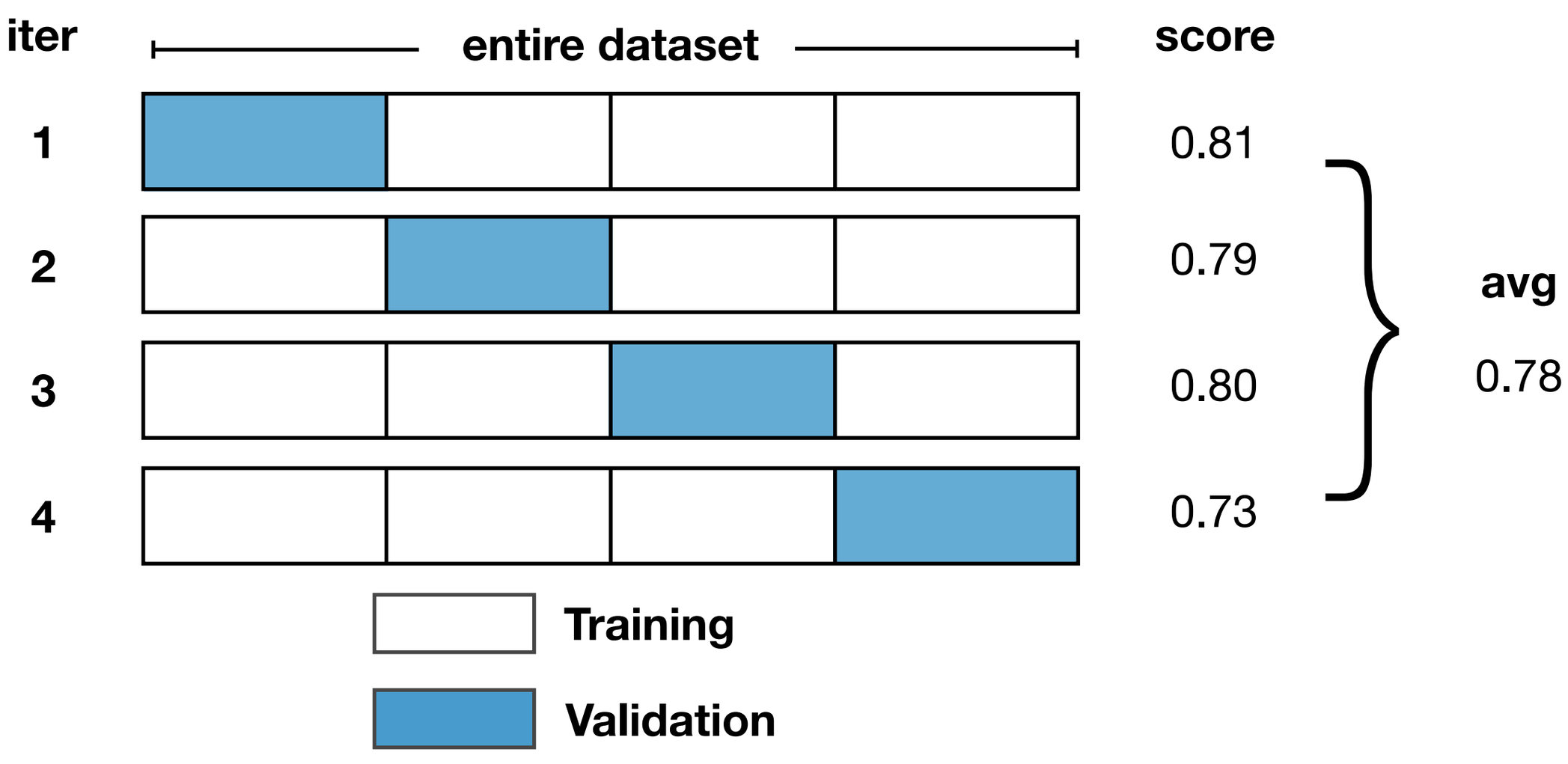
Combination of L1 and L2
Define “mixture ratio”

Cross-validation

Divide training data into k subsets (“folds”)

Train model on k-1 folds over k iterations

Calculate average score



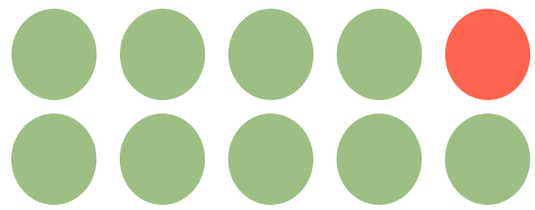
Occam's Razor

Pick the model with fewer assumptions!

Imbalanced Data

Inflated accuracy

- Example: 90% of patients did not have sepsis
- Predict that all patients did not have sepsis = 90% accuracy



How to overcome it

- Upsampling/downsampling
  - Bootstrapping
  - e.g. Synthetic Minority Over-sampling Technique (SMOTE)
- Use information retrieval metrics (recall, precision, F1, confusion matrix) rather than accuracy

A Word of Caution

Biased datasets

Gender Bias in Clinical Trials: Do Double Standards Still Apply?

K. Ramasubbu, H. Gurm, D. Litaker

Published Online: 7 Jul 2004 | https://doi.org/10.1089/15246090152636514

In Focus Blog – Published on: May 02, 2018

Women Are Still Underrepresented in Clinical Trials for Cardiovascular Disease Drugs

Kelly Davio

"Fluctuating hormones and differences between male and female study subjects could all complicate the design of the study"

Defining the "ground truth"

Error and discrepancy in radiology: inevitable or avoidable?

Adrian P. Brady [1]

Ann Surg Oncol. 2015 Jul;22(7):2359-64. doi: 10.1245/s10434-014-4205-5. Epub 2015 Jan 22.

Breast Imaging Second Opinions Impact Surgical Management.

Spivey TL[1], Carlson KA, Janssen I, Witt TR, Jokich P, Madrigrano A.

Is SOFA a reliable indicator of sepsis?

Selecting the appropriate evaluation metric

False positives vs. False negatives

Thank you!

Jill Cates
twitter: @jillacates
github: @topspinj
cates.jill@gmail.com

References

1) Sepsis article. Wikipedia.
2) Stevenson EK et al. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis. Crit Care Med 2014;42:625.
3) Cost H et al. In Healthcare Cost and Utilization Project (HCUP) Statistical Briefs: MDAgency for Healthcare Research and Quality USA, 2006.
4) Angus DC et al. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Criti Care Med. 2001;1303-10.
5) Martin GS et al. The Epidemiology of Sepsis in the United States from 1979 through 2000. N Engl J Med 2003; 348:1546-1554.
6) Vincent JL et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. Intensive Care Med. 1996;22:707–710.
7) Bone RC, Balk RA, Cerra FB, et al. Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis. Chest 1992;101:1644-55.
8) Le Gall JR. et al. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. JAMA. 1996;276(10):802–10.
9) Seymour CW, Rea TD, Kahn JM, Walkey AJ, Yealy DM, Angus DC. Severe sepsis in pre-hospital emergency care: analysis of incidence, care, and outcome. Am J Respir Crit Care Med. 2012;186(12):1264–1271.