Target Leakage in ML

Yuriy Guts

AI UKRAINE 18
V INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND DATA SCIENCE APPLICATIONS

DataRobot

Machine Learning Engineer

Automated ML, time series forecasting, NLP

Ukrainian Catholic University
APPLIED SCIENCES FACULTY

Lecturer

AI, Machine Learning, Summer/Winter ML Schools

kaggle™

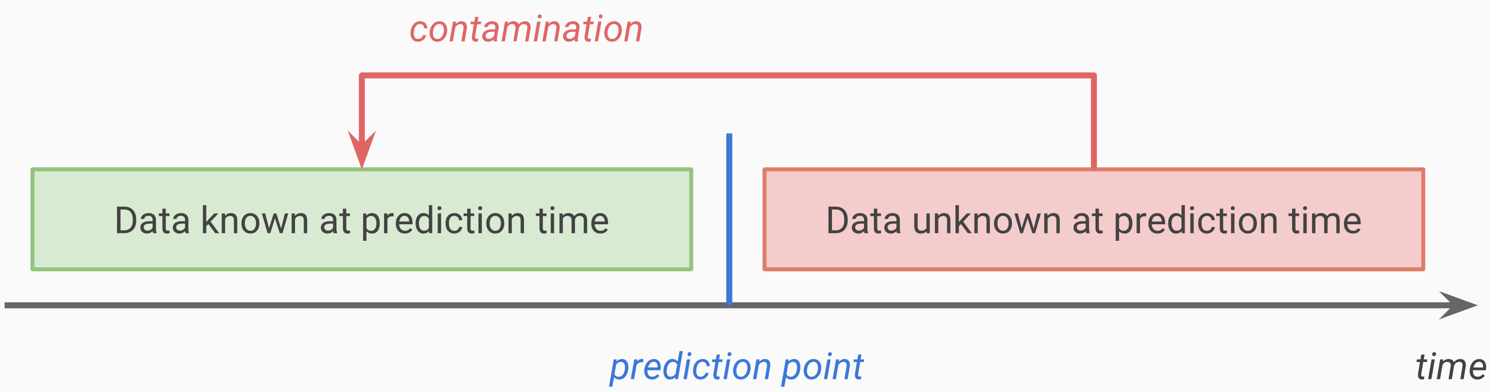
Compete sometimes

Currently hold an Expert rank, top 2% worldwide

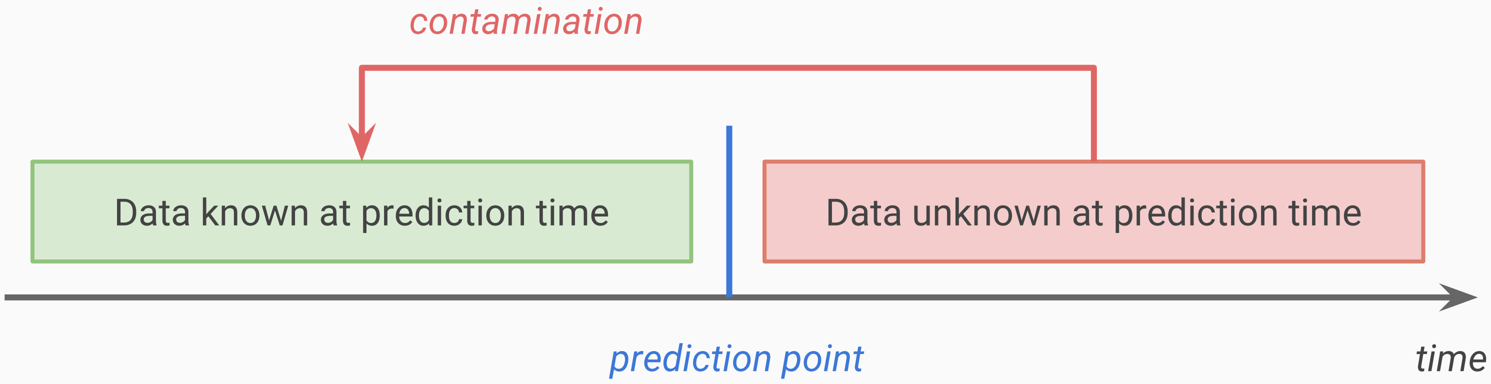
Follow my presentation and code at:

https://github.com/YuriyGuts/odsc-target-leakage-workshop

Leakage in a Nutshell



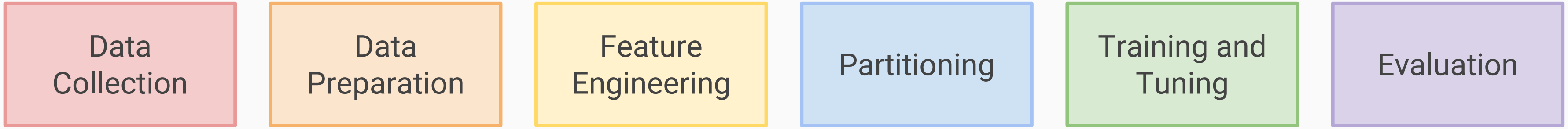
Leakage in a Nutshell



Training on contaminated data leads to overly optimistic expectations about model performance in production

"But I always validate on random K-fold CV. I should be fine, right?"

They suspect nothing



Leakage can happen anywhere during the project lifecycle

Leakage in Data Collection

Where is the leakage?

EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

Target is a function of another column

EmployeeID	Title	ExperienceYears	MonthlySalaryGBP	AnnualIncomeUSD
315981	Data Scientist	3	5,000.00	78,895.44
4691	Data Scientist	4	5,500.00	86,784.98
23598	Data Scientist	5	6,200.00	97,830.35

The target can have different formatting or measurement units in different columns.

Forgetting to remove the copies will introduce target leakage.

Check out the example: **example-01-data-collection.ipynb**

Where is the leakage?

SubscriberID	Group	DailyVoiceUsage	DailySMSUsage	DailyDataUsage	Gender
24092091	M18-25	15.31	25	135.10	0
4092034091	F40-60	35.81	3	5.01	1
329815	F25-40	13.09	32	128.52	1
94721835	M25-40	18.52	21	259.34	0

Feature is an aggregate of the target

SubscriberID	Group	DailyVoiceUsage	DailySMSUsage	DailyDataUsage	Gender
24092091	M18-25	15.31	25	135.10	0
4092034091	F40-60	35.81	3	5.01	1
329815	F25-40	13.09	32	128.52	1
94721835	M25-40	18.52	21	259.34	0

E.g., the data can have derived columns created after the fact for reporting purposes

Where is the leakage?

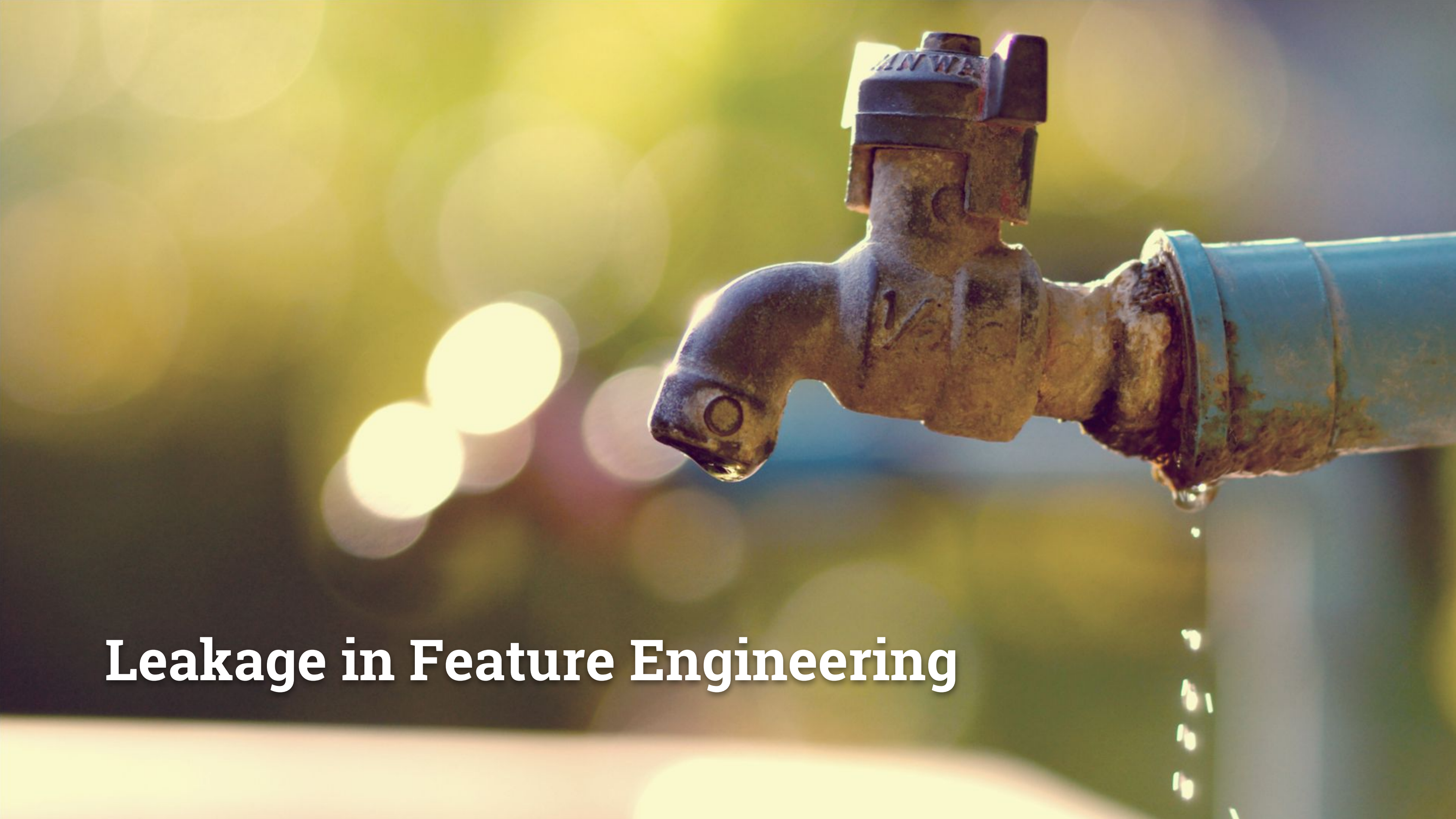
Education	Married	AnnualIncome	Purpose	LatePaymentReminders	IsBadLoan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0

Mutable data due to lack of snapshot-ability

Education	Married	AnnualIncome	Purpose	LatePaymentReminders	IsBadLoan
1	Y	80k	Car Purchase	0	0
3	N	120k	Small Business	3	1
1	Y	85k	House Purchase	5	1
2	N	72k	Marriage	1	0

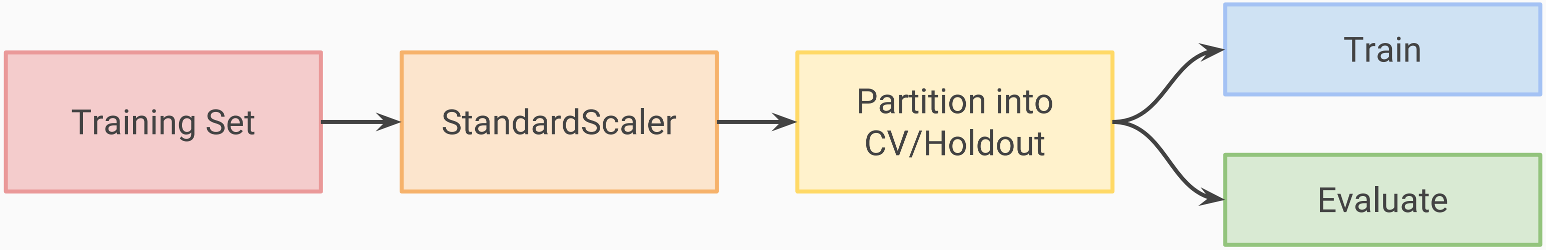
Database records get overwritten as more facts become available.

But these later facts won't be available at prediction time.

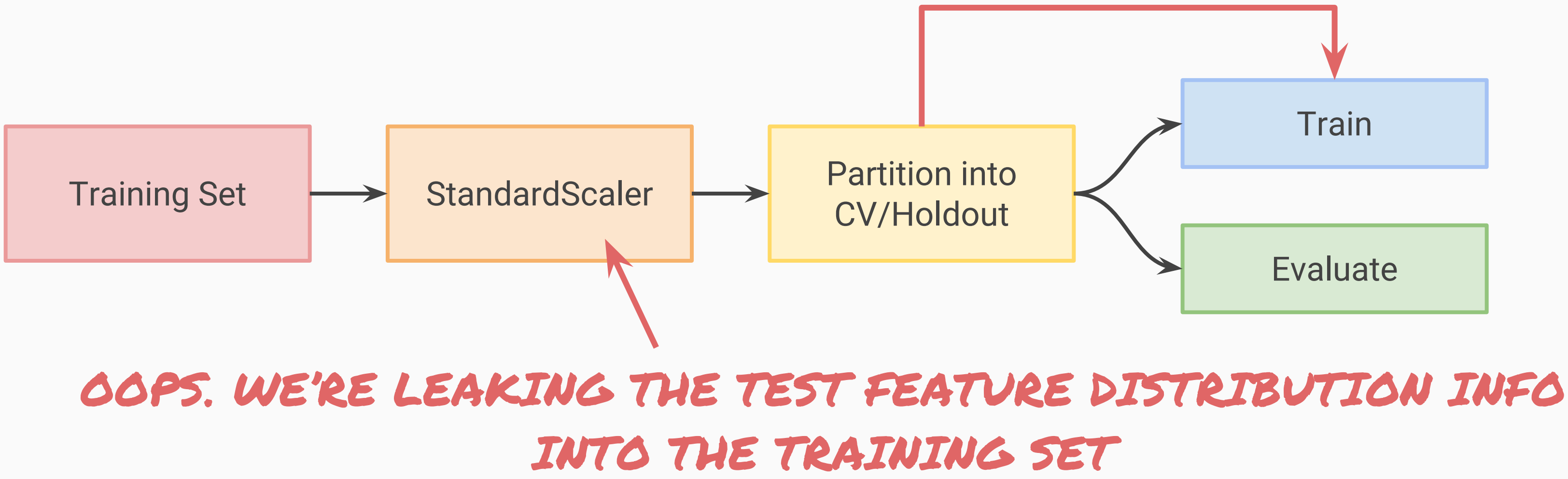


Leakage in Feature Engineering

My model is sensitive to feature scaling...

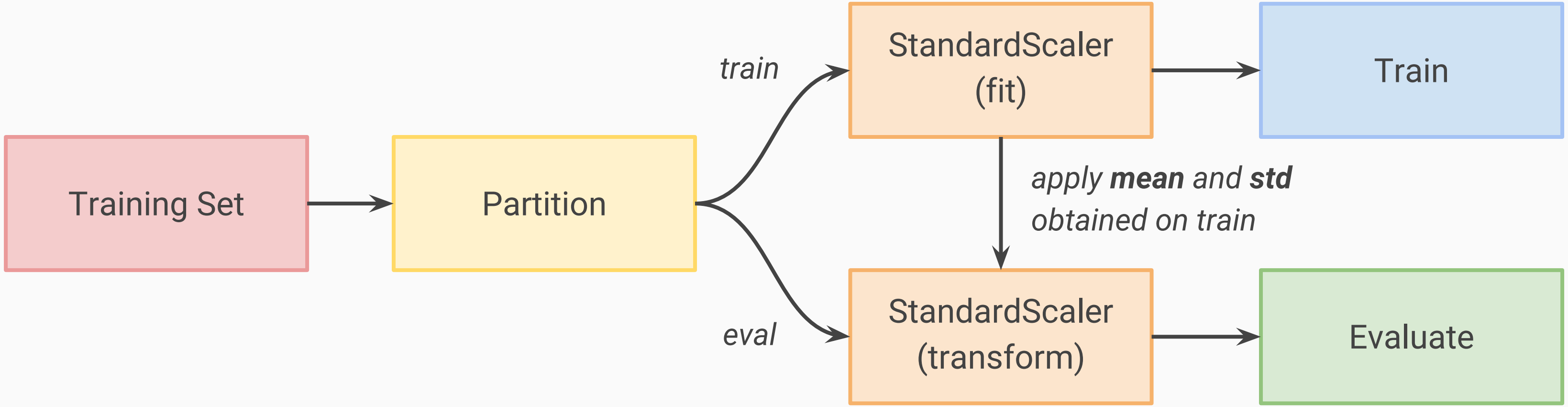


My model is sensitive to feature scaling...



Check out the example: **example-02-data-prep.ipynb**

Removing leakage in feature engineering



Obtain feature engineering/transformation parameters only on the training set

Apply them to transform the evaluation sets (CV, holdout, backtests, …)

Encoding of different variable types

Text:

Learn DTM columns from the **training set only**, then transform the evaluation sets

(avoid leaking possible out-of-vocabulary words into the training pipeline)

Categoricals:

Create mappings on the **training set only**, then transform the evaluation sets

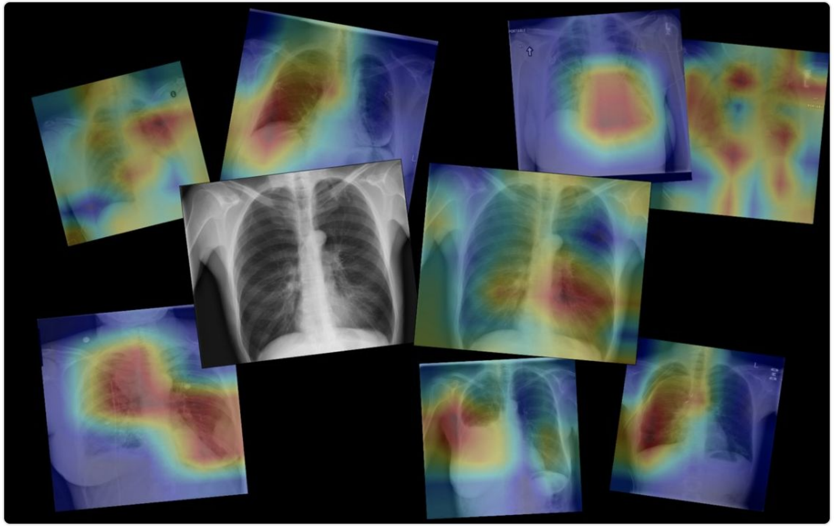
(avoid leaking cardinality/frequency info into the training pipeline)

Leakage in Partitioning

Andrew Ng
@AndrewYNg

Follow

Our full paper on Deep Learning for pneumonia detection on Chest X-Rays. @pranavrajpurkar @jeremy_irvin16 @mattlungrenMD arxiv.org/abs/1711.05225



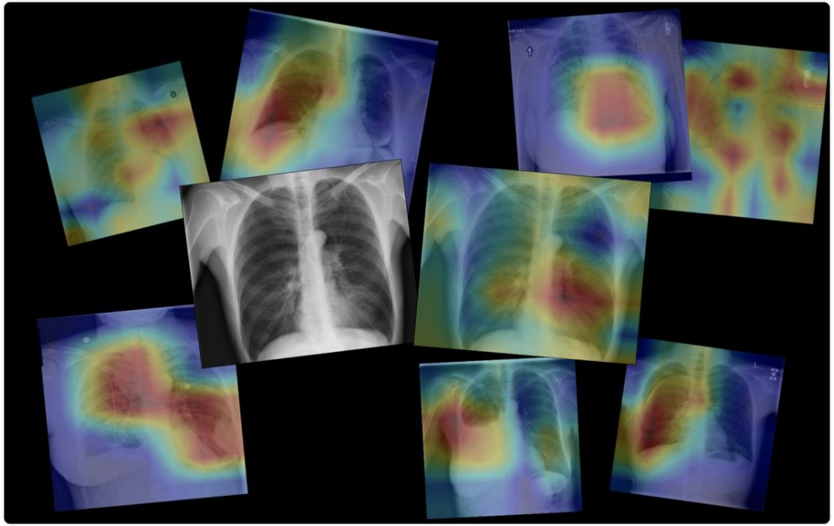
9:09 PM - 15 Nov 2017 from Mountain View, CA

665 Retweets **1,352** Likes

Andrew Ng @AndrewYNg

Follow

Our full paper on Deep Learning for pneumonia detection on Chest X-Rays. @pranavrajpurkar @jeremy_irvin16 @mattlungrenMD arxiv.org/abs/1711.05225



9:09 PM - 15 Nov 2017 from Mountain View, CA

665 Retweets 1,352 Likes

3. Data

3.1. Training

We use the ChestX-ray14 dataset released by Wang et al. (2017) which contains 112,120 frontal-view X-ray images of 30,805 unique patients. Wang et al. (2017) annotate each image with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. We label images that have pneumonia as one of the annotated pathologies as positive examples and label all other images as negative examples for the pneumonia detection task. We randomly split the entire dataset into 80% training, and 20% validation.

Before inputting the images into the network, we downscale the images to 224×224 and normalize based on the mean and standard deviation of images in the ImageNet training set. We also augment the training data with random horizontal flipping.

https://twitter.com/AndrewYNg/status/931026446717296640

Group Leakage

Nick Roberts
@nizkroberts

Follow

Replying to @AndrewYNg @pranavrajpurkar and 2 others

Were you concerned that the network could memorize patient anatomy since patients cross train and validation?
"ChestX-ray14 dataset contains 112,120 frontal-view X-ray images of 30,805 unique patients. We randomly split the entire dataset into 80% training, and 20% validation."

OOPS. THERE ARE FOUR TIMES MORE UNIQUE IMAGES THAN PATIENTS ples. For the pneumonia detection task, we randomly split the dataset into training (28744 patients, 98637 images), validation (1672 patients, 6351 images), and test (389 patients, 420 images). There is no patient overlap between the sets.

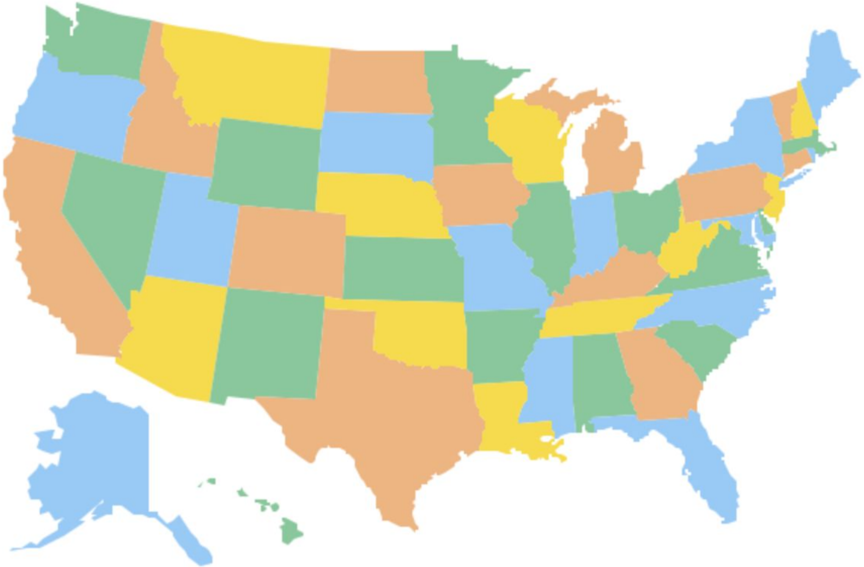
Pathology	Wang et al. (2017)	Yao et al. (2017)	CheXNet (ours)
Atelectasis	0.716	0.772	**0.8209**
Cardiomegaly	0.807	0.904	**0.9048**
Effusion	0.784	0.859	**0.8831**
Infiltration	0.609	0.695	**0.7204**
Mass	0.706	0.792	**0.8618**
Nodule	0.671	0.717	**0.7766**
Pneumonia	0.633	0.713	**0.7632**
Pneumothorax	0.806	0.841	**0.8932**
Consolidation	0.708	0.788	**0.7939**
Edema	0.835	0.882	**0.8932**
Emphysema	0.815	0.829	**0.9260**
Fibrosis	0.769	0.767	**0.8044**
Pleural Thickening	0.708	0.765	**0.8138**
Hernia	0.767	0.914	**0.9387**

Paper v1 (AUC)

CheXNet (ours)
0.8094
0.9248
0.8638
0.7345
0.8676
0.7802
0.7680
0.8887
0.7901
0.8878
0.9371
0.8047
0.8062
0.9164

Paper v3 (AUC)

The Cold Start Problem



Observe:

Illinois

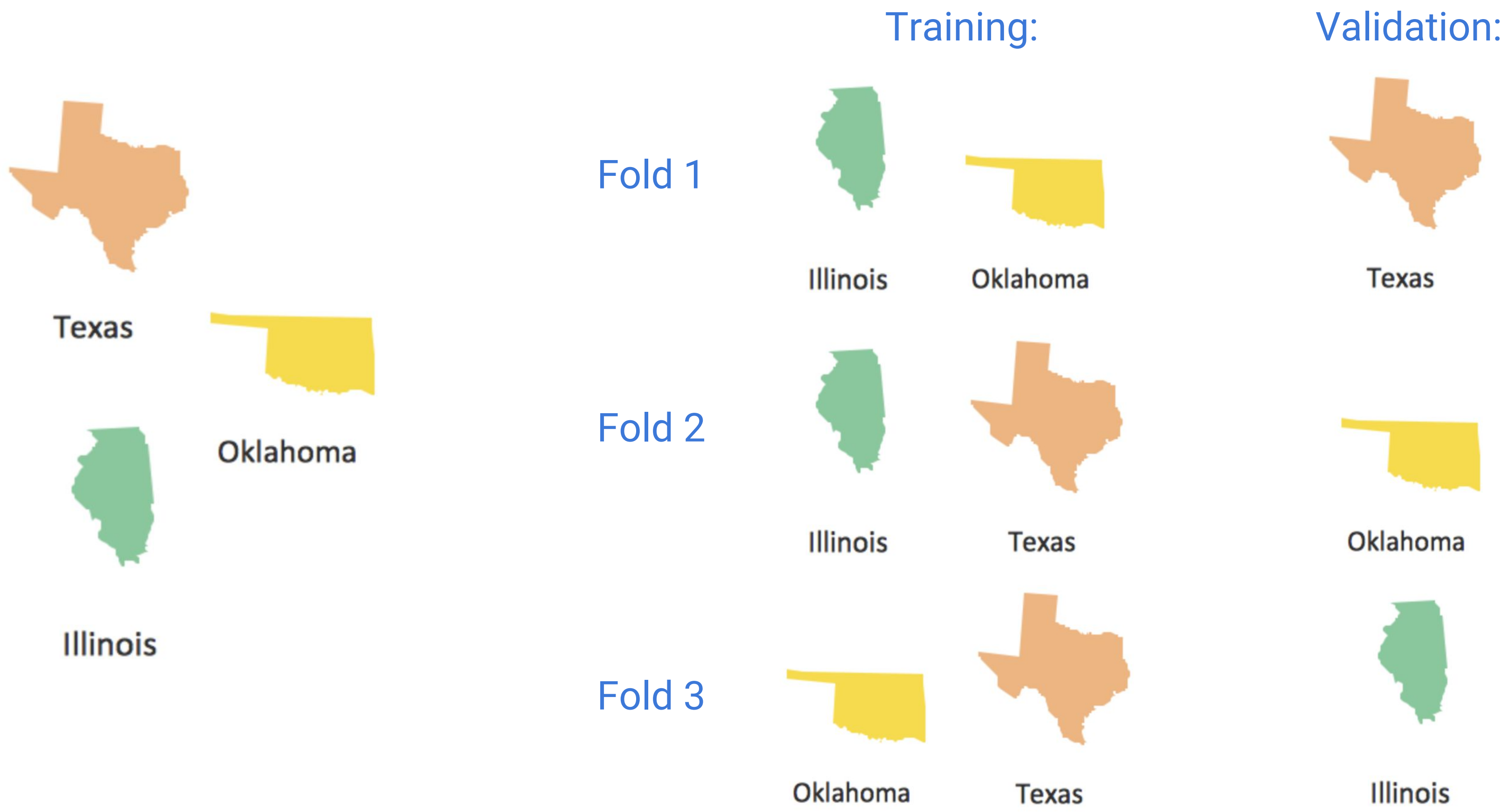
Oklahoma

Texas

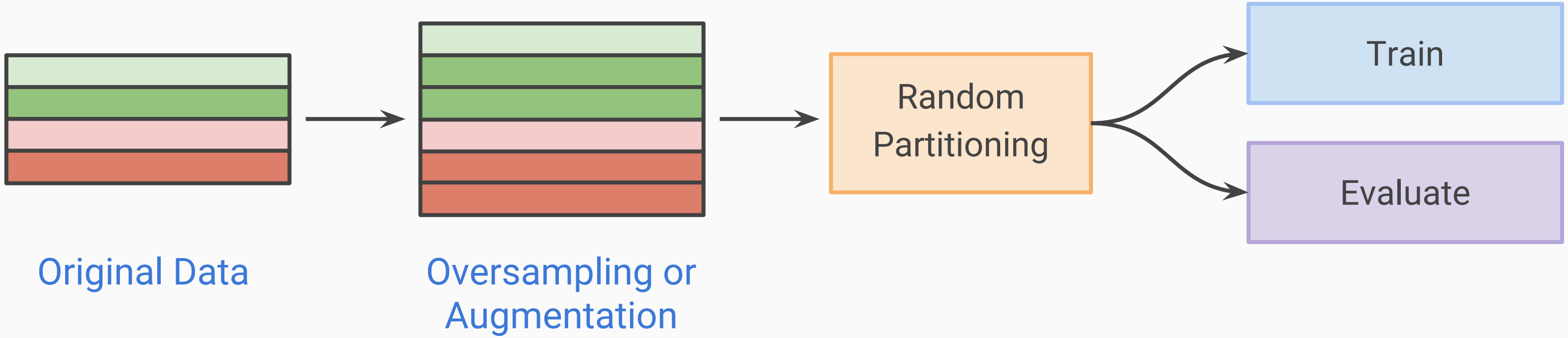
Predict:

North Carolina

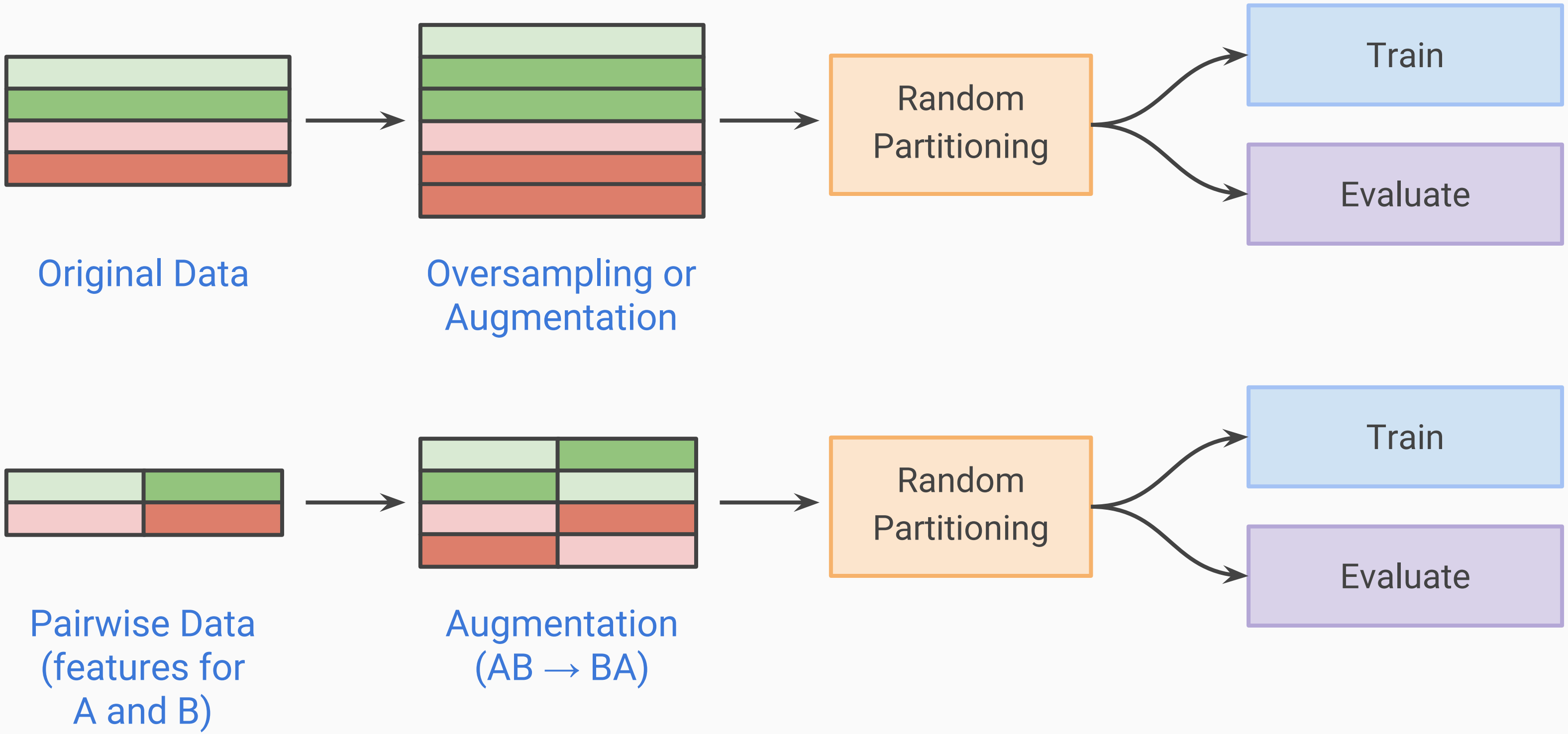
Group Partitioning, Out-of-Group Validation



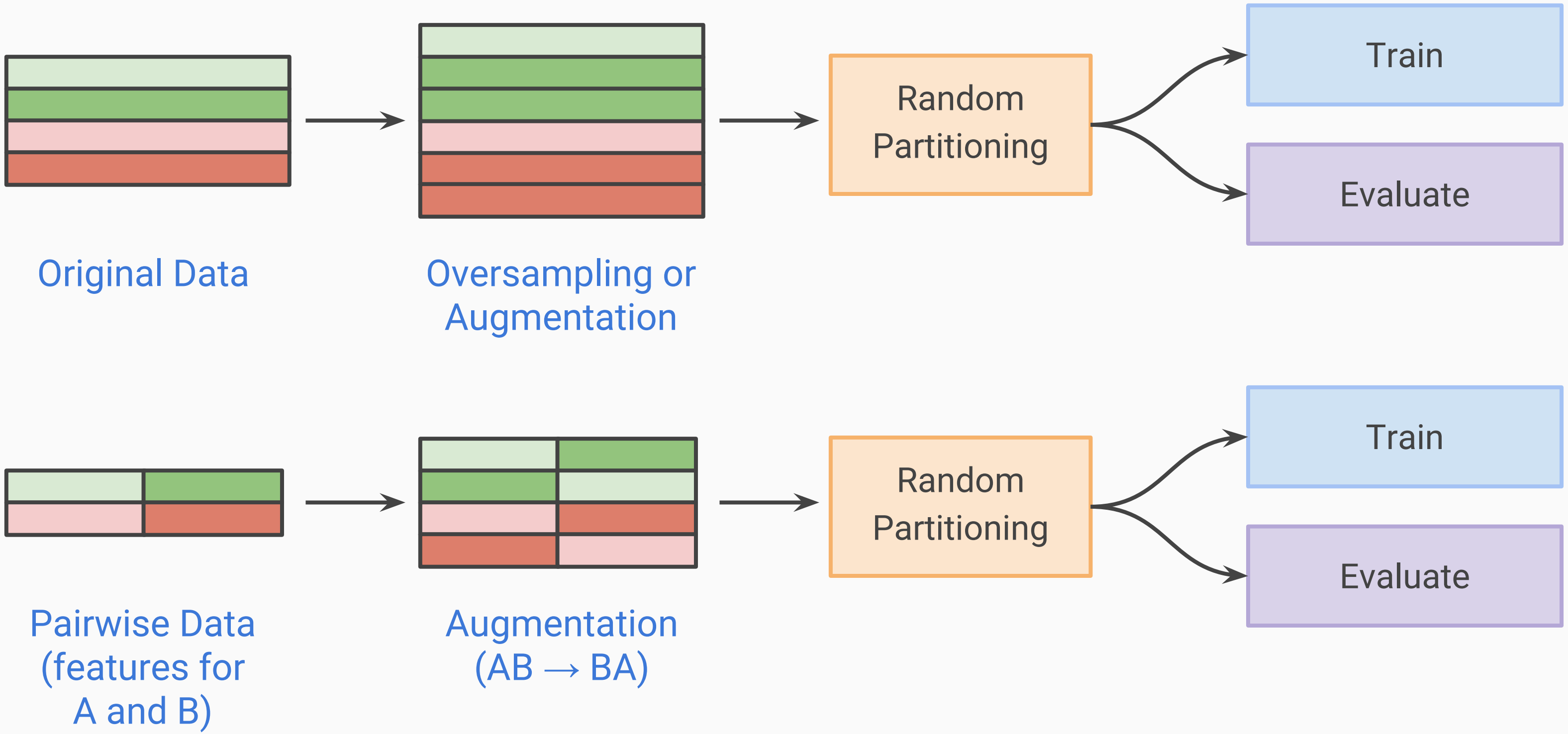
Leakage in oversampling / augmentation



Leakage in oversampling / augmentation

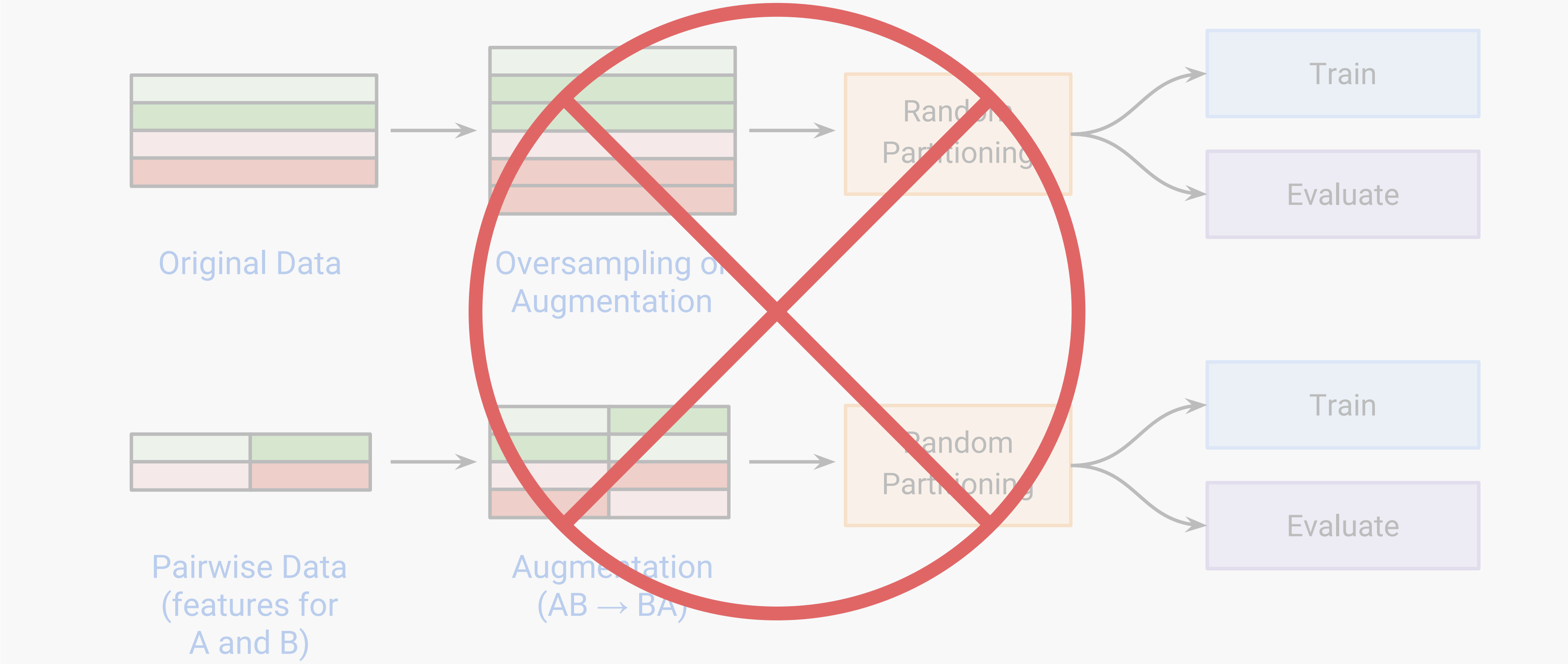


Leakage in oversampling / augmentation



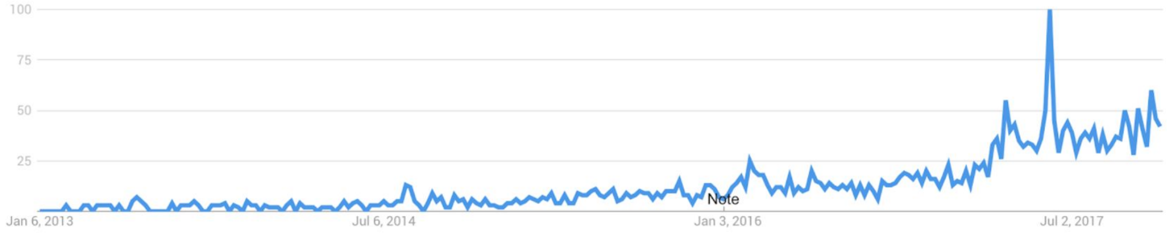
OOPS. WE MAY GET COPIES SPLIT BETWEEN TRAINING AND EVALUATION

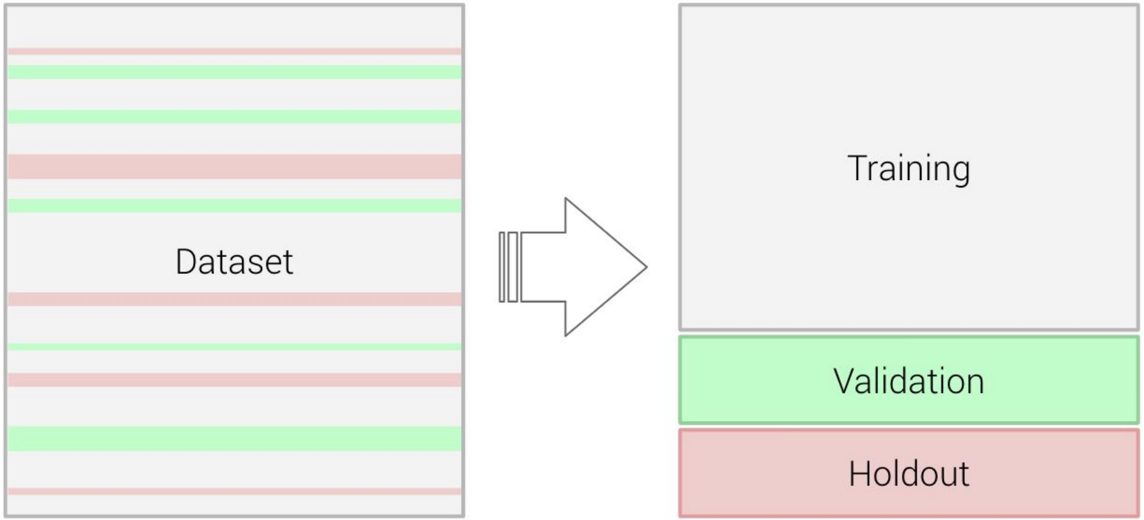
Leakage in oversampling / augmentation



First partition, then augment the training data.

Random Partitioning for Time-Aware Models

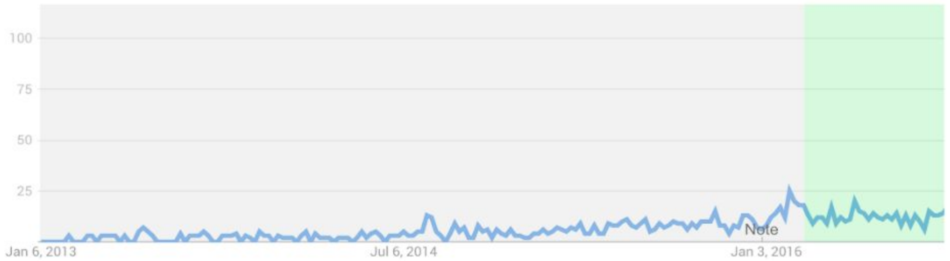




Random Partitioning for Time-Aware Models

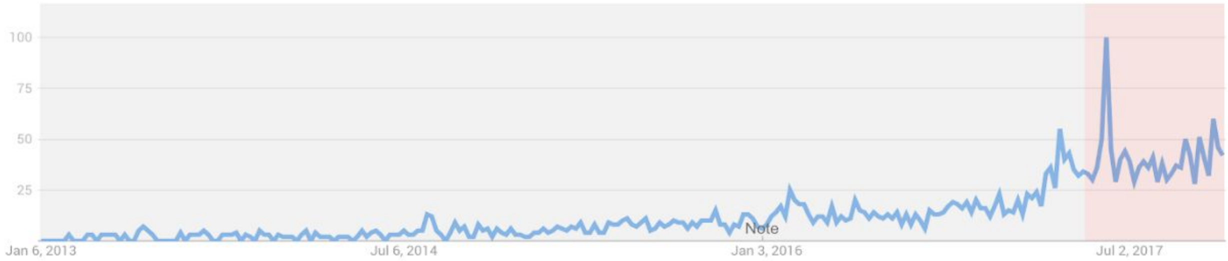
Out-of-Time Validation (OTV)

Backtest 2



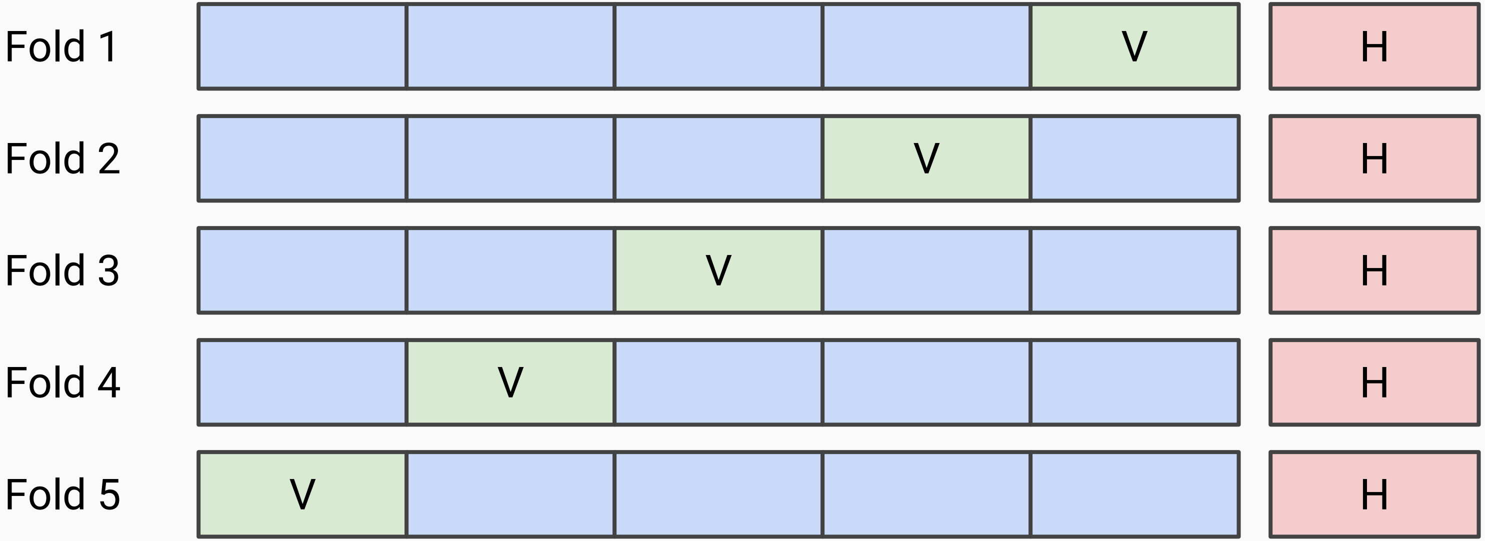
Backtest 1

Holdout



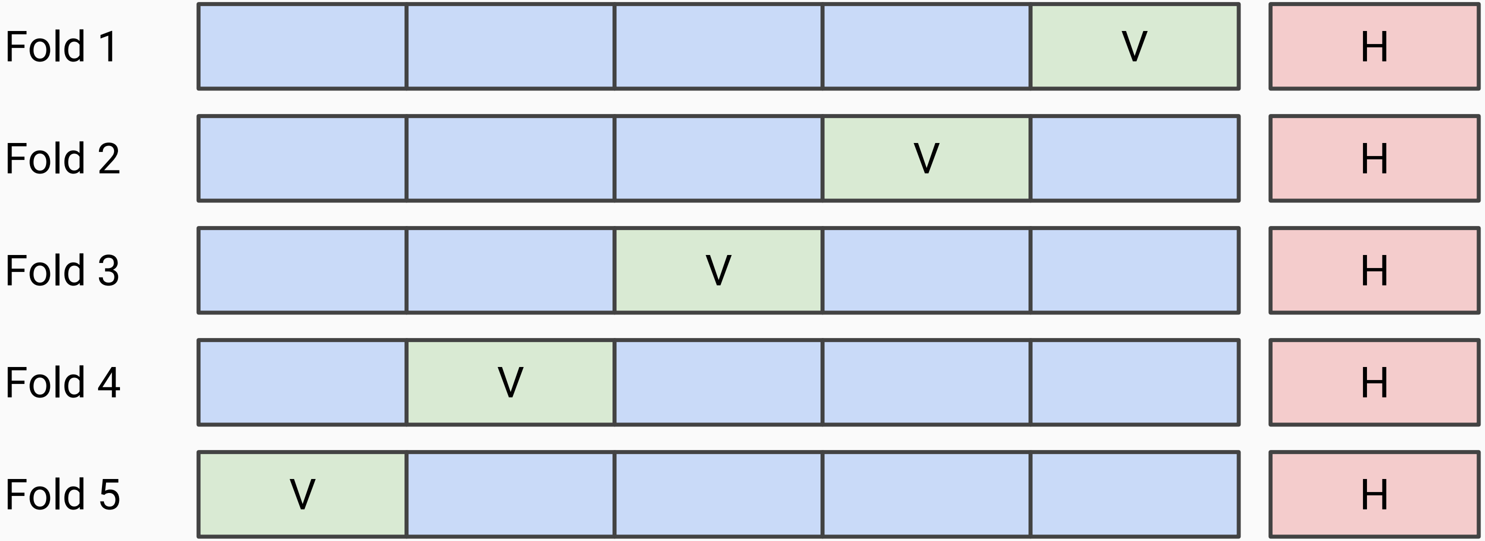
Leakage in Training & Tuning

Reusing a CV split for multiple tasks



Feature selection, hyperparameter tuning, model selection...

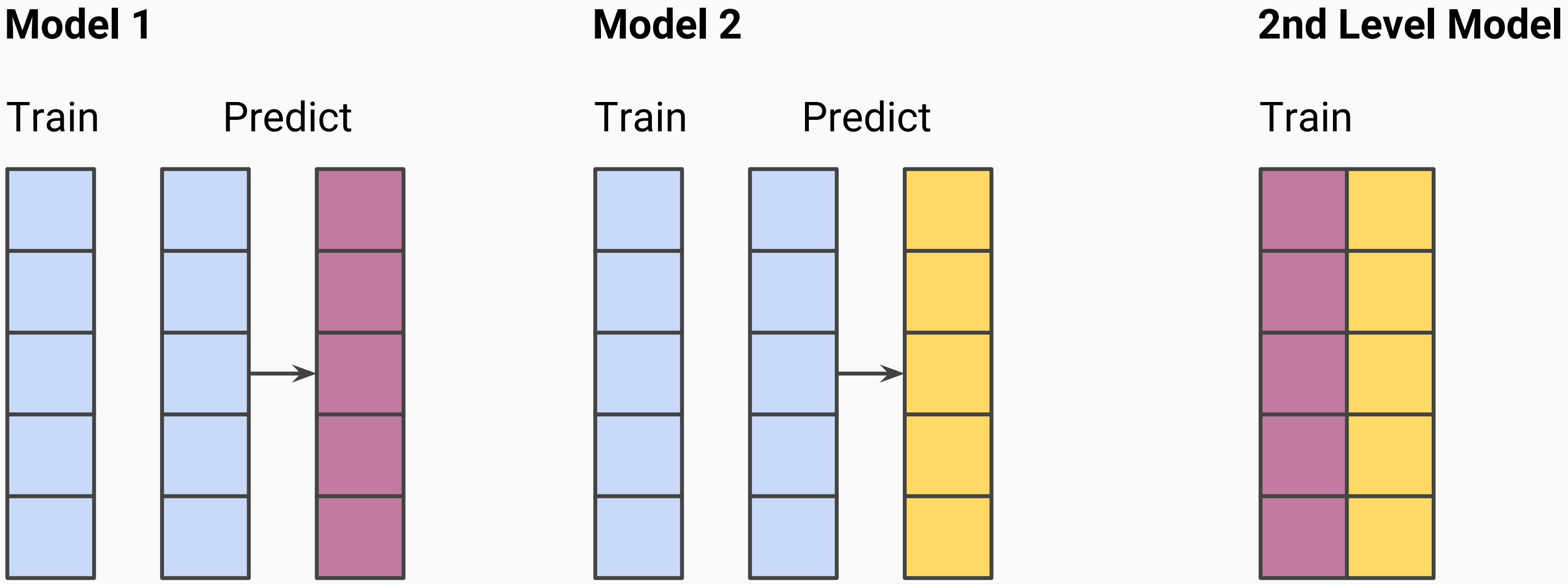
Reusing a CV split for multiple tasks



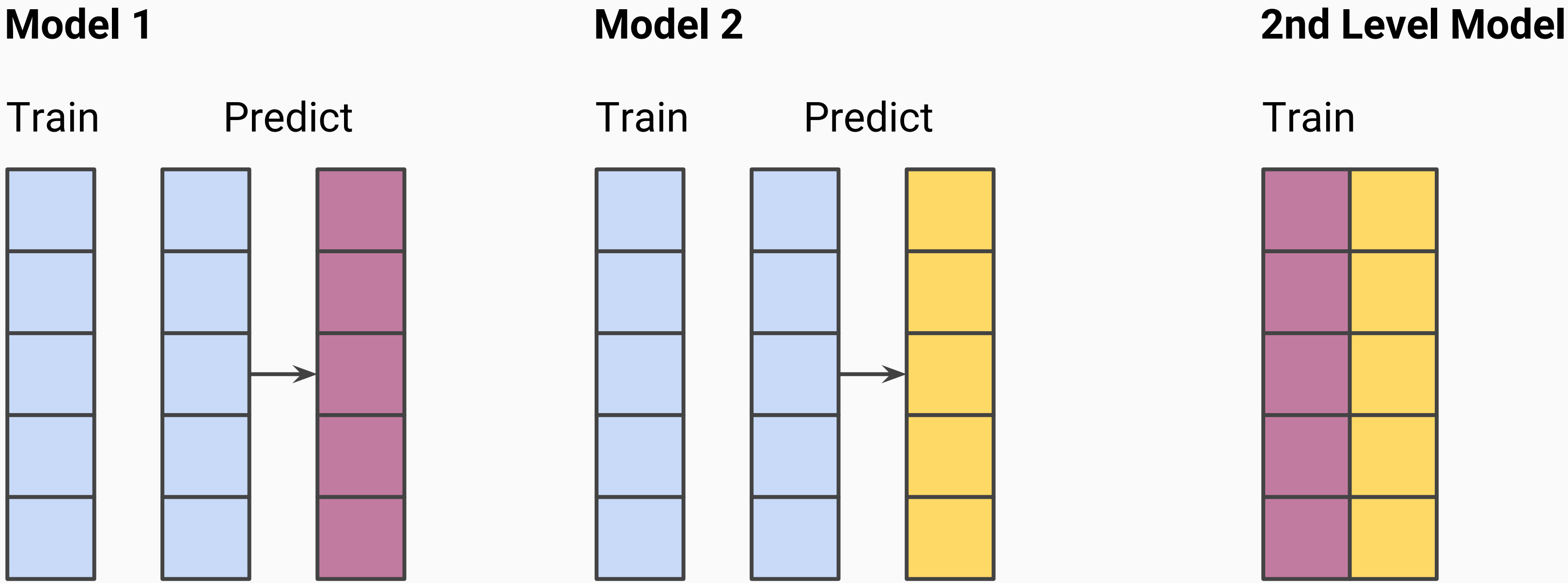
Feature selection, hyperparameter tuning, model selection...

OOPS. CAN OVERFIT VALIDATION FOLDS
BETTER USE DIFFERENT SPLITS FOR DIFFERENT TASKS

Model stacking on in-sample predictions

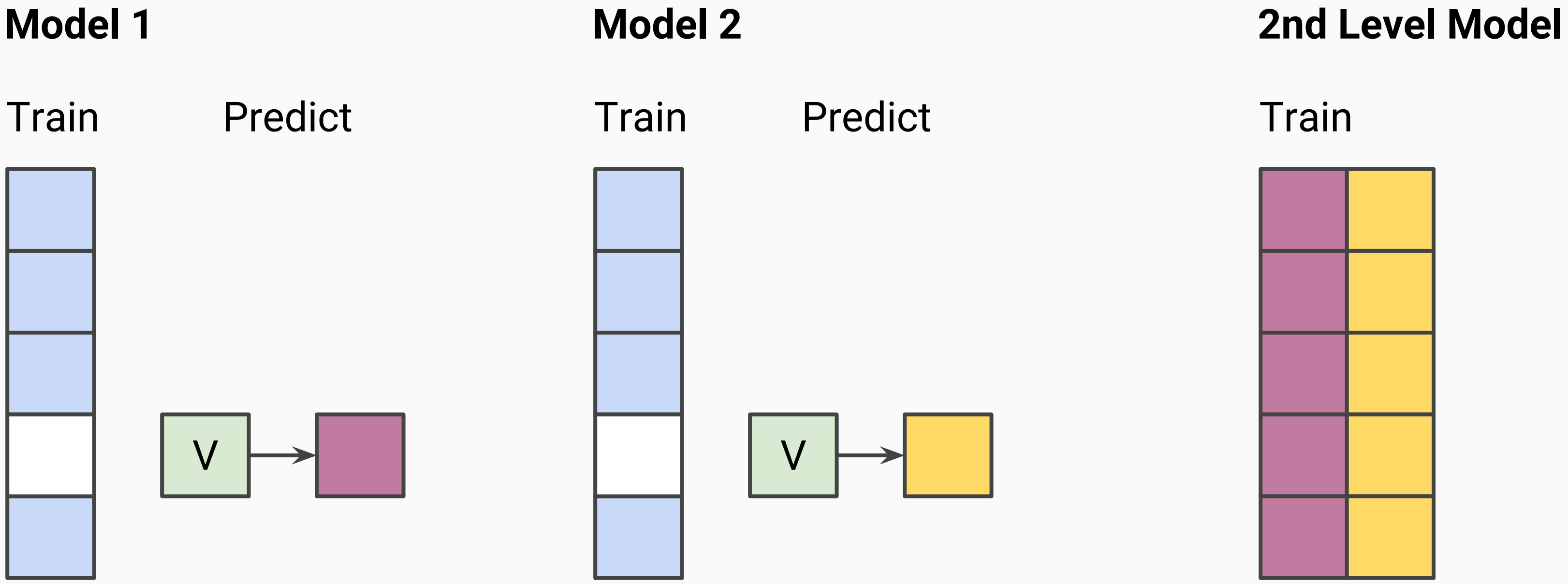


Model stacking on in-sample predictions



OOPS. WILL LEAK THE TARGET IN THE META-FEATURES

Better way to stack



Compute all meta-features only out-of-fold


Leakage in Competitive ML

Case Studies

- Removing customer/user IDs does not necessarily mean data anonymization
  (Kaggle: Wikipedia Participation Challenge, 2011)
- Anonymizing feature names does not mean anonymization either
  (Kaggle: Santander Value Prediction competition, 2018)
- Target can sometimes be recovered using side channels or external datasets
  (Kaggle: Dato “Truly Native?” competition, 2015)
- Overrepresented minority class opens possibilities for reverse engineering
  (Kaggle: Quora Question Pairs competition, 2017)

Prevention Measures

Leakage prevention checklist (not exhaustive!)

- Split the holdout away immediately and do not preprocess it in any way before final model evaluation.
- Make sure you have a data dictionary and understand the meaning of every column, as well as unusual values (e.g. negative sales) or outliers.
- For every column in the final feature set, try answering the question:
  “Will I have this feature at prediction time in my workflow? What values can it have?”
- Figure out preprocessing parameters on the training subset, freeze them elsewhere.
- Treat feature selection, model tuning, model selection as separate “machine learning models” that need to be validated separately.
- Make sure your validation setup represents the problem you need to solve with the model.
- Check feature importance and prediction explanations: do top features make sense?

Thank you!

yuriy.guts@gmail.com

linkedin.com/in/yuriyguts