YOU ARE WHAT YOU EAT!
ML.NET with Jupyter Notebooks

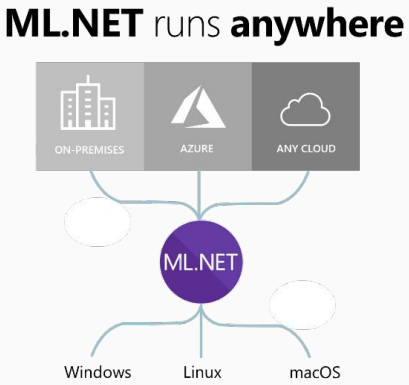
Microsoft® Most Valuable Professional

Speaker: Daniel Costea

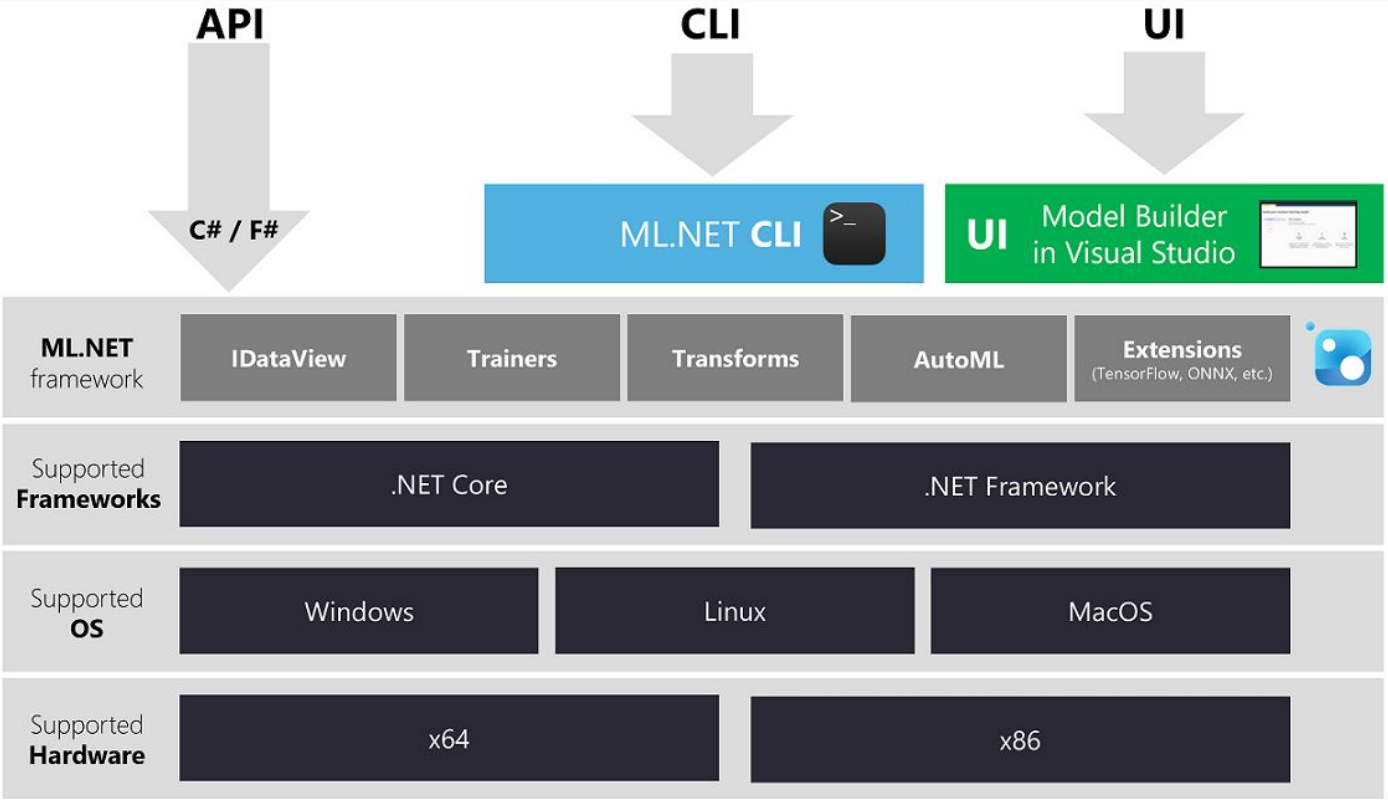
MICROSOFT ML.NET

WHAT IS ML.NET?

- ML.NET = machine learning framework for .NET developers
- Open-source
- Cross-platform
- On-premise
- In-process



ML.NET ARCHITECTURE



MACHINE LEARNING DATA FLOW

- ❑ Data Loading
- ❑ Data Preparation
- ❑ Data Visualization
- ❑ Data Preparation
- ❑ Model Validation
- ❑ Model Evaluation
- ❑ Data Prediction

DATA LOADING

DATASET ANATOMY

OBSERVATIONS

index	Temperature	Luminosity	Infrared	Distance	CreatedAt	Source
0	24.82	50.49	0	8.84	01/03/2020 18:22:56	FlashLight
1	23.96	4.1	0	154.09	06/03/2020 21:31:55	Day
2	32.5	72.27	0	66.87	05/03/2020 11:29:22	FlashLight
3	47.32	100	94.34	41.55	04/03/2020 9:28:24	Lighter
4	25.09	12.11	0	400	04/03/2020 10:26:18	Day
5	56.03	100	92.29	400	04/03/2020 8:27:51	Lighter
6	32.18	11.72	15.43	60.87	05/03/2020 12:30:21	Infrared
7	23.28	45.51	0	12.08	02/03/2020 10:23:51	FlashLight
8	27.34	70.31	0	194.26	03/03/2020 19:24:48	FlashLight
9	24.77	18.55	0	400	03/03/2020 18:25:00	Day

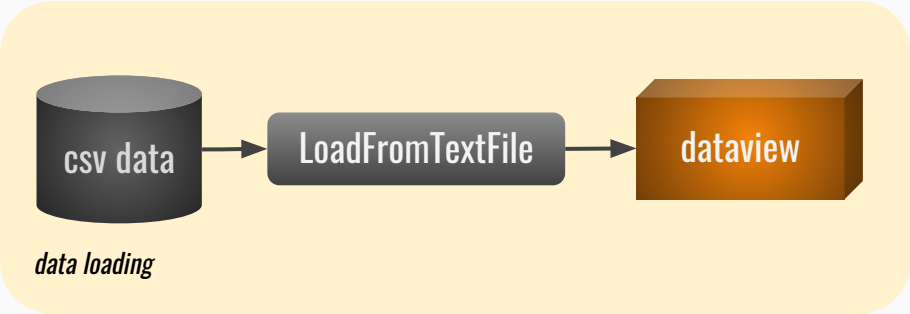
FEATURES

LABEL

- Each set of data consists of features used to make prediction and expected outcome is called label (target feature).
- The word “supervised” comes from a fact that labels need to be assigned to data by the human supervisor (this is the case for supervised learning).

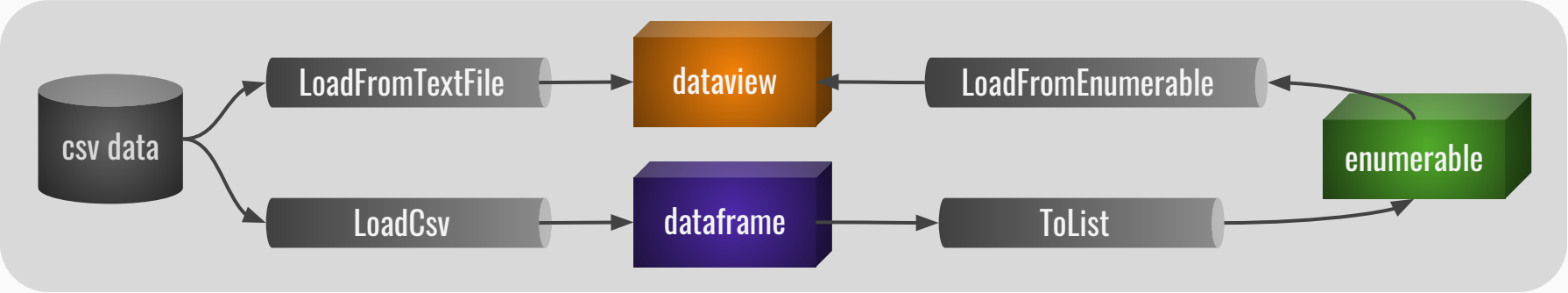
DATA LOADER AND DATAVIEW

- **Data Loader** - So far, data can be read from databases (SQL Server, Oracle, PostgreSQL, MySQL, etc.), text files (csv, tsv, txt), binary files, image files and from memory (IEnumerable collections)
- **DataView** - In ML.NET, data is similar to a SQL view: It's a lazily evaluated, in-memory, immutable, cursorable, heterogenous, schematized dataset (each column has name, type, metadata), composable (new views are formed by applying transformations on other views)



FROM DATAFRAME TO DATAVIEW

- Inspect your data in various ways
  - Collections (IEnumerable)
  - DataFrame (similar to Pandas from Python)
  - Preview extracts data from a DataView (do not use Preview in production!)
    - DataView is lazy evaluated, but you can take peek at any data view object by calling the Preview method



DATAFRAME

Basic questions about the dataset:

- ❑ How many observations?
- ❑ How many features?
- ❑ Data types of my features? Are they numeric? Categorical?
- ❑ Which is the label feature?

The purpose of displaying examples from the dataset is not to perform rigorous analysis. Instead, it's to get a qualitative "feel" for the dataset.

- ❑ Do the columns make sense?
- ❑ Do the values make sense?
- ❑ Are the values on the right scale?
- ❑ Is missing data going to be a big problem?

JUPYTER NOTEBOOKS

WHAT IS JUPYTER NOTEBOOKS?

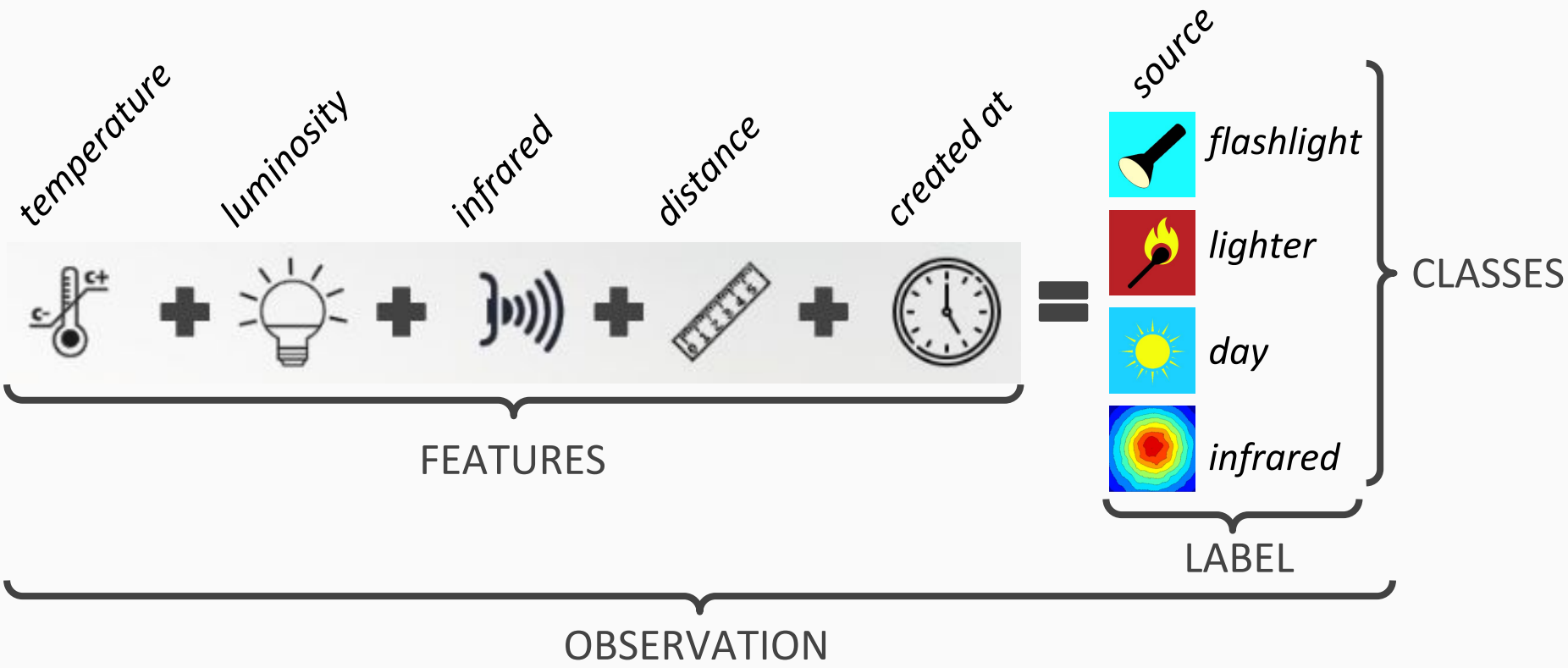
- Kernels (Python, R, .NET (C#), .NET (F#), PowerShell and more)
- Commands (type #!lsmagic to find more commands)
- PocketView (not documented yet!)
  - Interrogate supported tags:

```
var pocketViewTagMethods = typeof(PocketViewTags)
    .GetProperties()
    .Select(m => m.Name);
```

- XPlot (data visualisation library)
  - Prints text, html, svn, charts

REAL WORLD SCENARIO

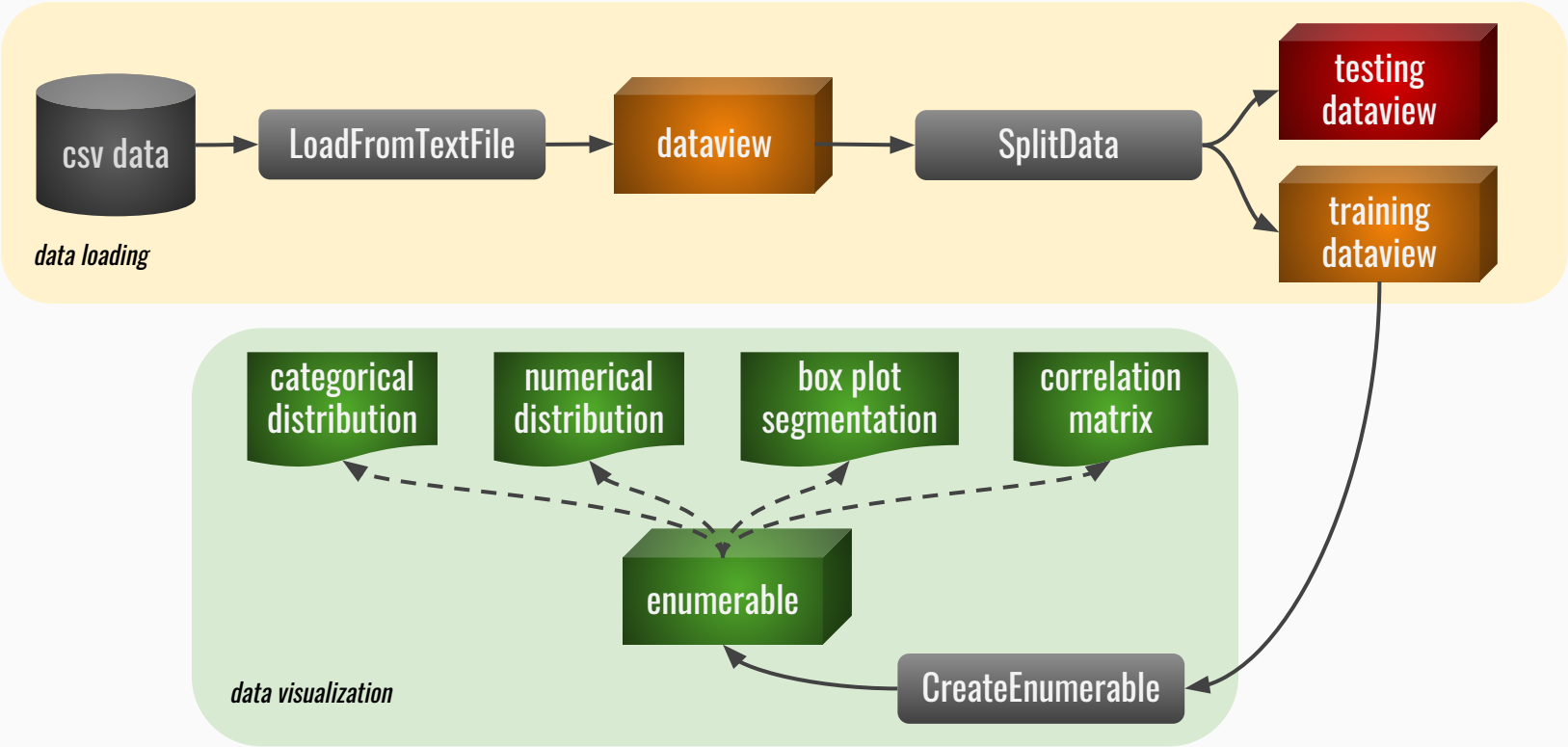
DEMO DATASET



DATAFRAME DEMO

DATA VISUALIZATION

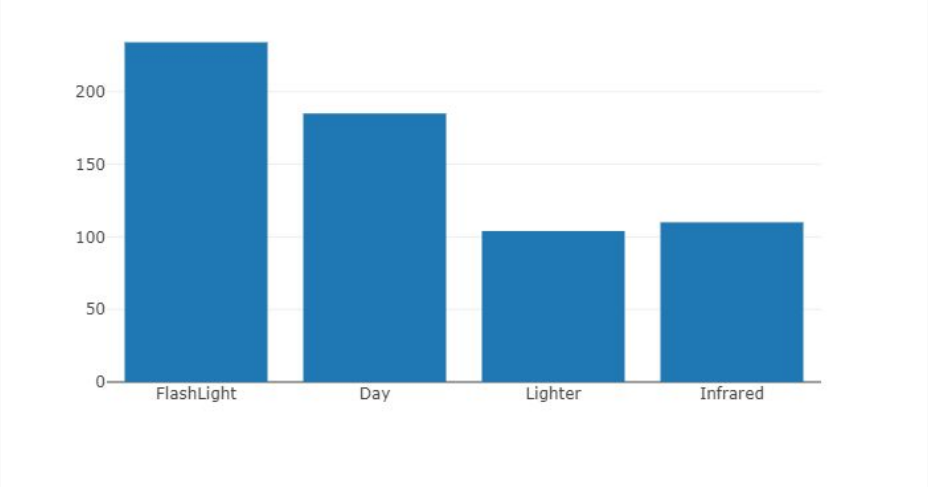
MACHINE LEARNING FLOW (DATA LOADING AND VISUALIZATION)



CATEGORICAL DISTRIBUTION

categorical distribution

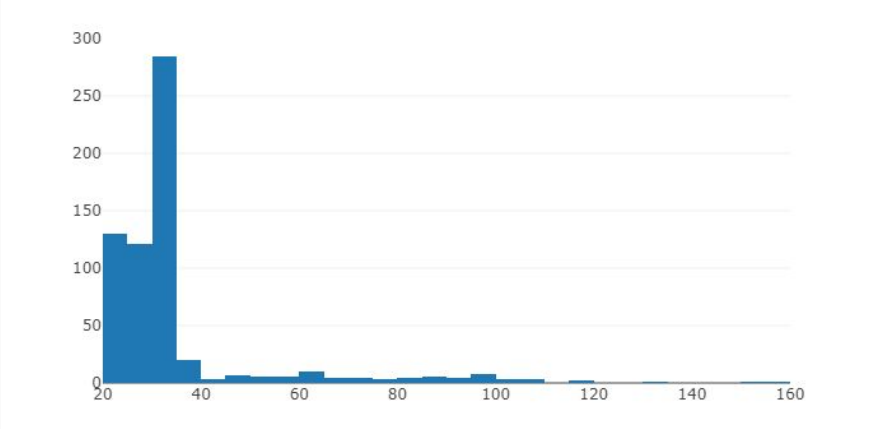
- Categorical features cannot be visualized through histograms. Instead, you can use bar plots.
- In particular, you'll want to look out for sparse classes, which are classes that have a very small number of observations.



NUMERICAL DISTRIBUTION

numerical distribution

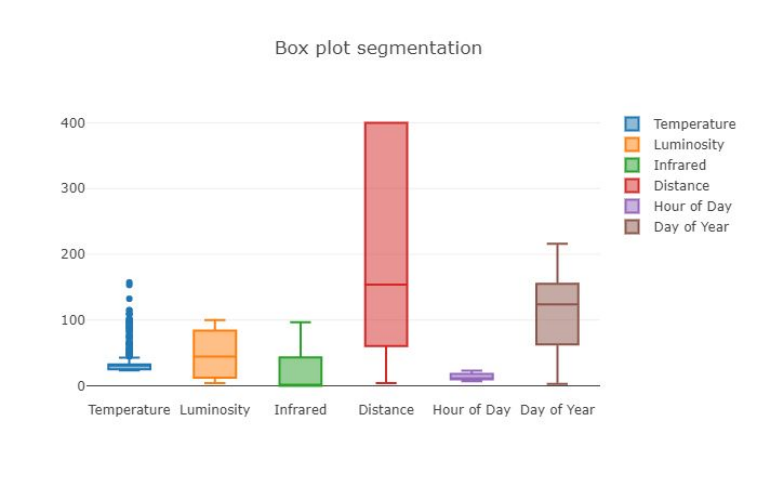
- Here are a few things to look out for:
  - Distributions that are unexpected
  - Potential outliers that don't make sense
  - Features that should be binary (i.e. "wannabe indicator variables")
- Boundaries that don't make sense
- Potential measurement errors



BOX PLOT SEGMENTATION

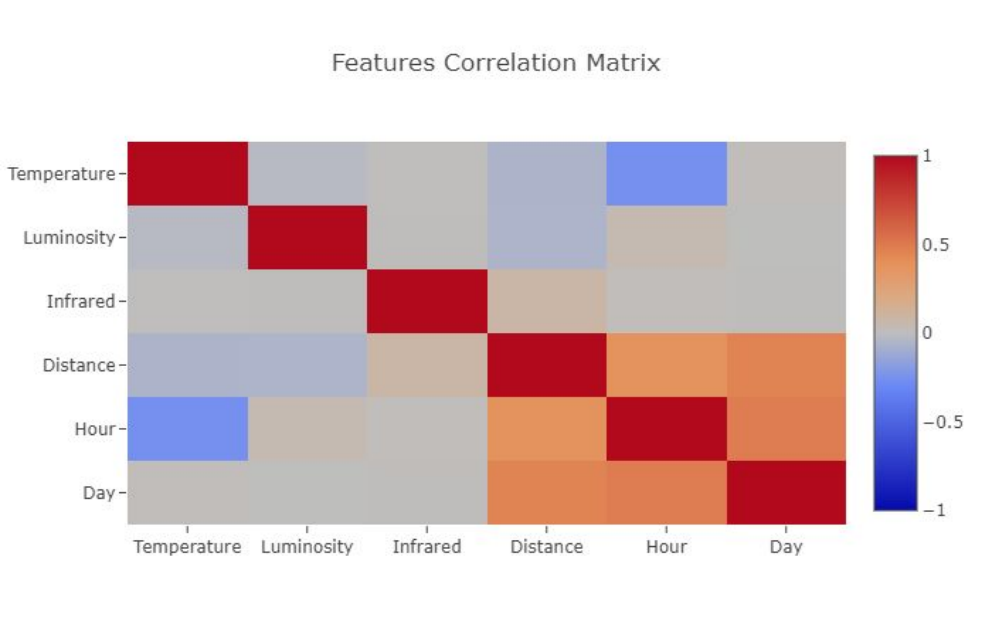
box plot segmentation

- Looking at the diagram we can extract valuable information like:
  - the median bar from Distance is much higher comparing to the other features
  - the min-max values from Temperature and Infrared are not uniformly distributed
  - Temperature has many outliers
- We can use this information later to improve the model accuracy



CORRELATION MATRIX

correlation matrix



- Which features are strongly correlated with the target variable?
- Are there interesting or unexpected strong correlations between other features?
- Correlation factor:
  - near -1 or 1 indicates a strong relationship (proportionality).
  - closer to 0 indicates a weak relationship.
  - 0 indicates no relationship

DATA VISUALIZATION DEMO

DATA PREPARATION

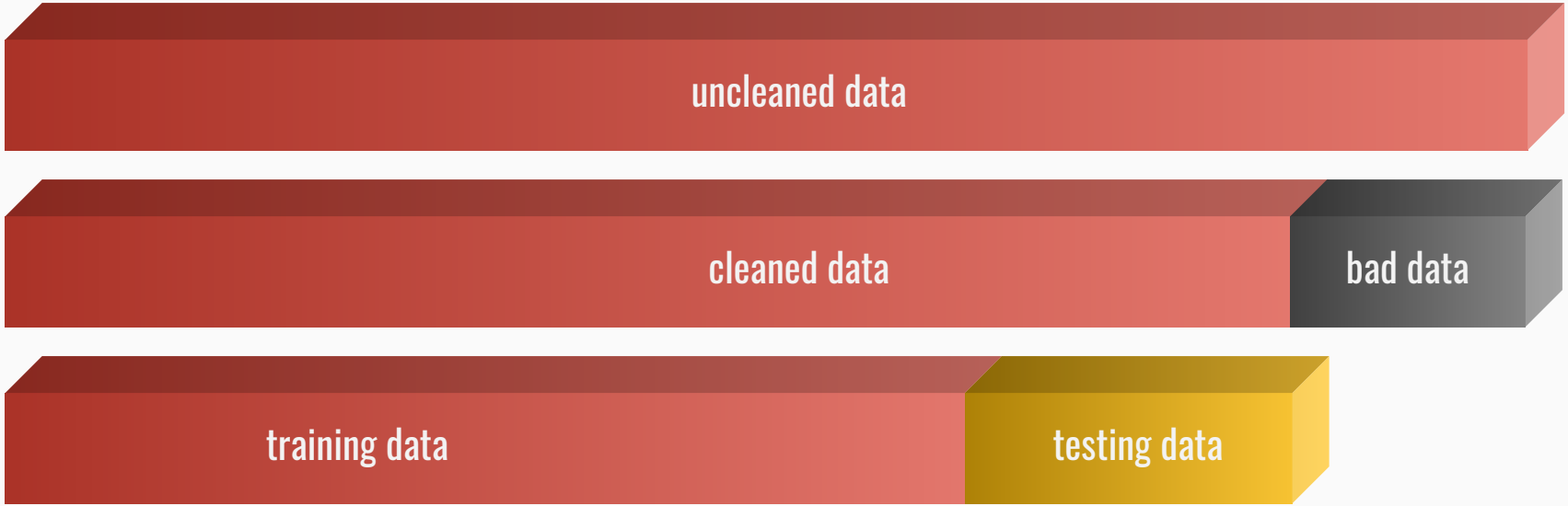
LESS PREPARATION - MORE INSIGHT

Credits: Talend.com

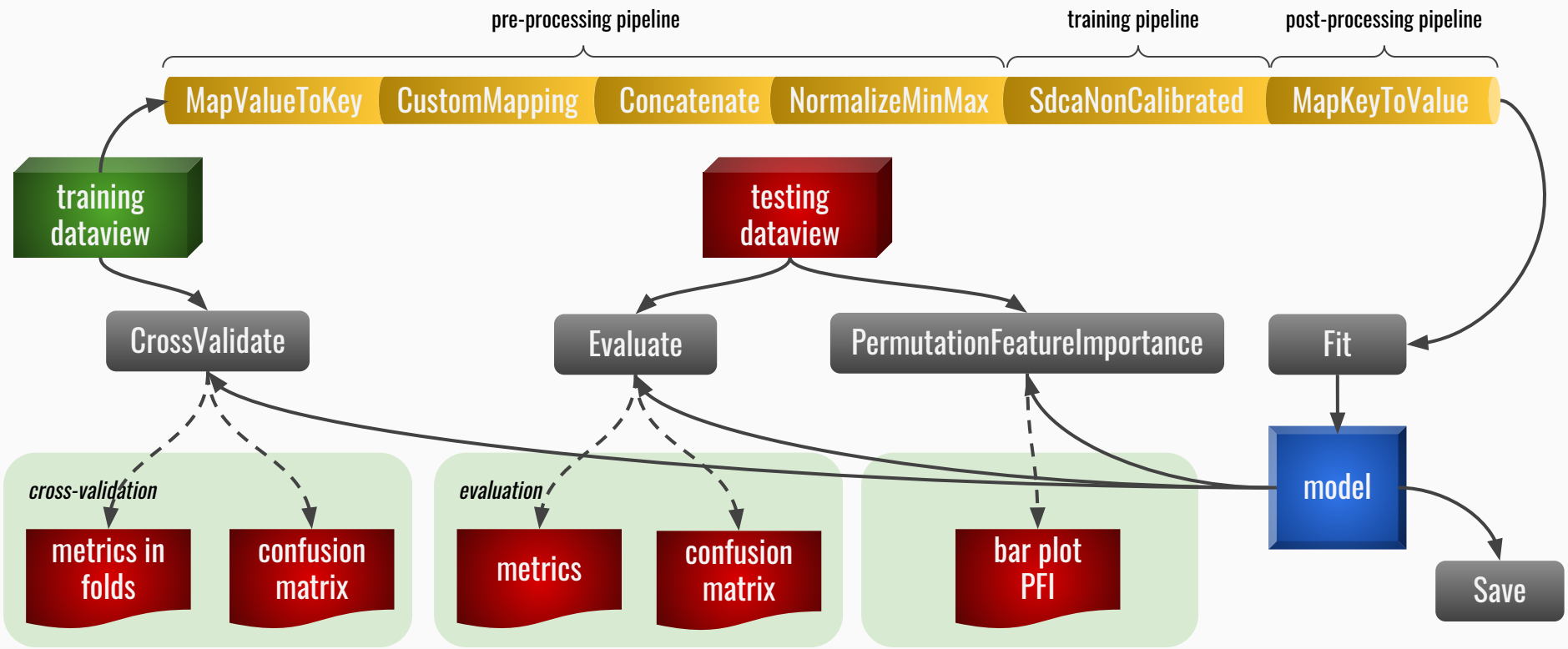
DATA FILTERING (CLEANING)

- Filter unwanted data
  - Duplicate - Dataframe
  - Irrelevant - Dataframe
- Fix structural errors
  - Typos, capitalization - Categorical distribution chart
- Filter unwanted outliers
  - Experimental error - Numerical distribution chart, Box plot segmentation chart

DATA CLEANING AND SPLITTING



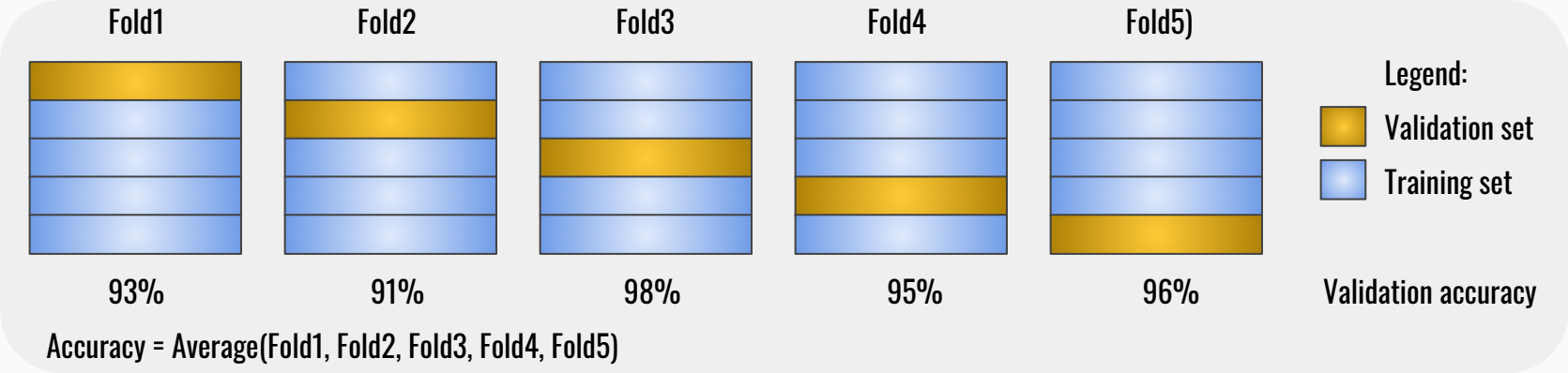
MACHINE LEARNING PIPELINES



MODEL VALIDATION

MODEL VALIDATION

- Cross-validation is a useful technique for ML applications. It helps estimate the variance of the model quality from one run to another and also eliminates the need to extract a separate test set for evaluation.
- Analyze metrics
  - Standard deviation, confidence interval for MicroAccuracy, MacroAccuracy, LogLoss - see Metrics



VALIDATION METRICS

metrics in folds

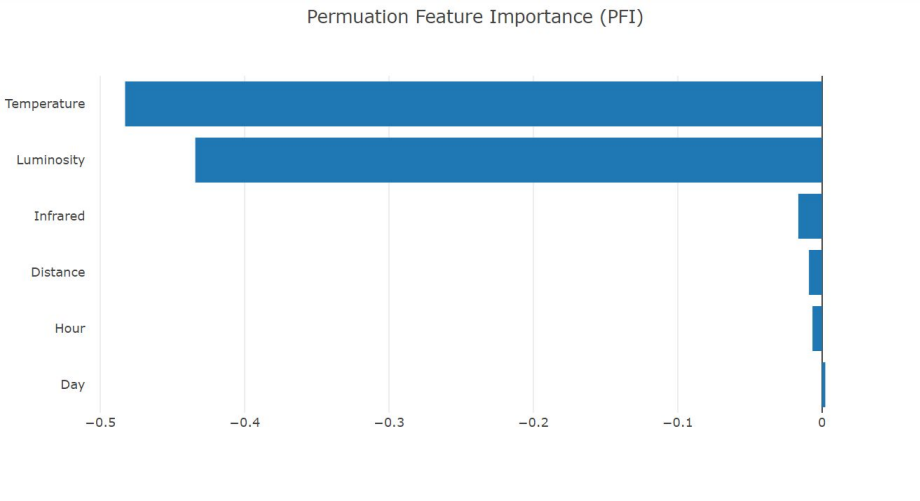
- The metrics are calculated for each fold and they are aggregates showing the average, standard deviation and confidence interval

CROSS-VALIDATION: multi-class classification	Average	Standard deviation	Confidence interval (95%)
MacroAccuracy	0.952	0.020	0.020
MicroAccuracy	0.949	0.019	0.018
LogLoss	8.483	11.457	11.228
LogLossReduction	-5.434	8.738	8.563

PERFORMANCE FEATURE IMPORTANCE (PFI)

bar plot
PFI

- We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature
- A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction



MODEL EVALUATION

MODEL EVALUATION

- EVALUATE MODEL
  - Analyze metrics
    - MicroAccuracy, MacroAccuracy, LogLoss - see Metrics
  - Confusion matrix

EVALUATION: multi-class classification	Class	Value	Note
MacroAccuracy		0.986	the closer to 1, the better
MicroAccuracy		0.988	the closer to 1, the better
LogLoss		24.447	the closer to 0, the better
LogLoss per Class	FlashLight	30.563	the closer to 0, the better
	Infrared	31.036	the closer to 0, the better
	Day	2.382	the closer to 0, the better
	Lighter	20.740	the closer to 0, the better

EVALUATION METRICS

metrics

- Micro-Accuracy aggregates the contributions of all classes to compute the average metric
  - The closer to 1.00, the better
  - In a multi-class classification task, micro-accuracy is preferable over macro-accuracy if you suspect there might be class imbalance
- Macro-Accuracy is the average accuracy at the class level. The accuracy for each class is computed and the macro-accuracy is the average of these accuracies
  - The closer to 1.00, the better
- Log-loss measures the performance of a classification model where the prediction input is a probability value between 0.00 and 1.00
  - The closer to 0.00, the better.
  - The goal of our machine learning models is to minimize this value.
- Log-Loss can be interpreted as the advantage of the classifier over a random prediction
  - Ranges from -inf and 1.00, where 1.00 is perfect predictions and 0.00 indicates mean predictions.
  - For example, if the value equals 0.20, it can be interpreted as "the probability of a correct prediction is 20% better than random guessing"

CONFUSION MATRIX

confusion matrix

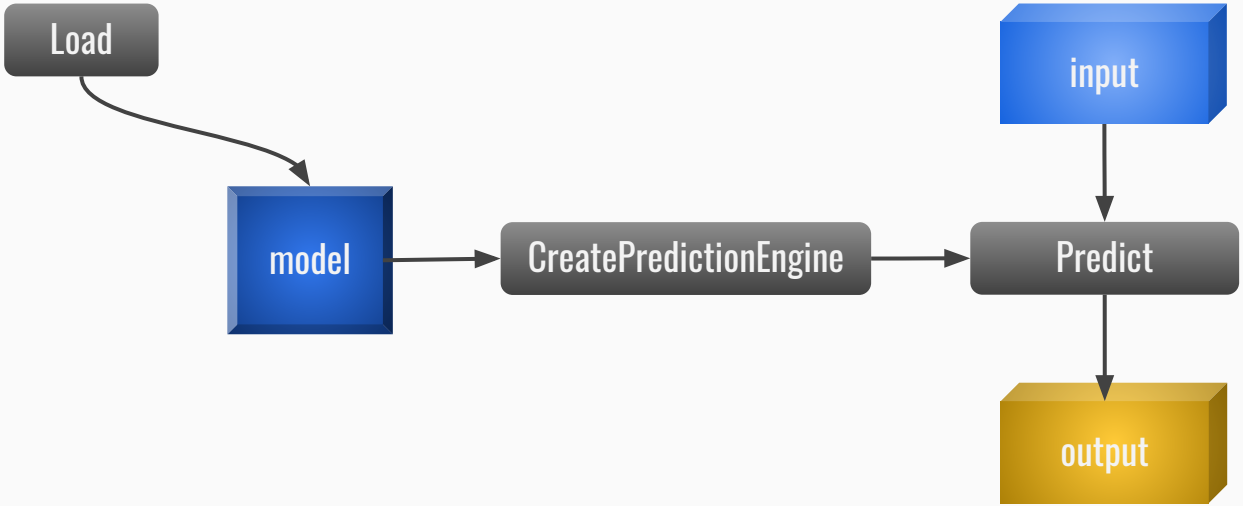
Using the testing dataset we can make predictions and compare the predicted results to the actual results.

Confusion Matrix		Predicted				
		FlashLight	**Infrared**	**Day**	**Lighter**	Recall
Truth	**FlashLight**	66	1	0	0	0.9851
	Infrared	0	49	0	0	1
	Day	1	0	28	0	0.9655
	Lighter	1	0	0	24	0.96
	Precision	0.9706	0.98	1	1	total = 170

DATA PREPARATION DEMO

DATA PREDICTION

DATA PREDICTION



From "Introduction to Microsoft Azure" by David Chappell

DATA PREDICTION DEMO

Q & A

Daniel Costea
developer, trainer & speaker

LinkedIn	https://linkedin.com/in/danielcostea	
Twitter	https://twitter.com/dfcostea	
TwitchTV	https://www.twitch.tv/daniel_apexcode	Subscribe!
Email	daniel_costea@ymail.com	
GitHub	https://github.com/dcostea/	
Presentation repo	https://github.com/dcostea/SmartFireAlarm	