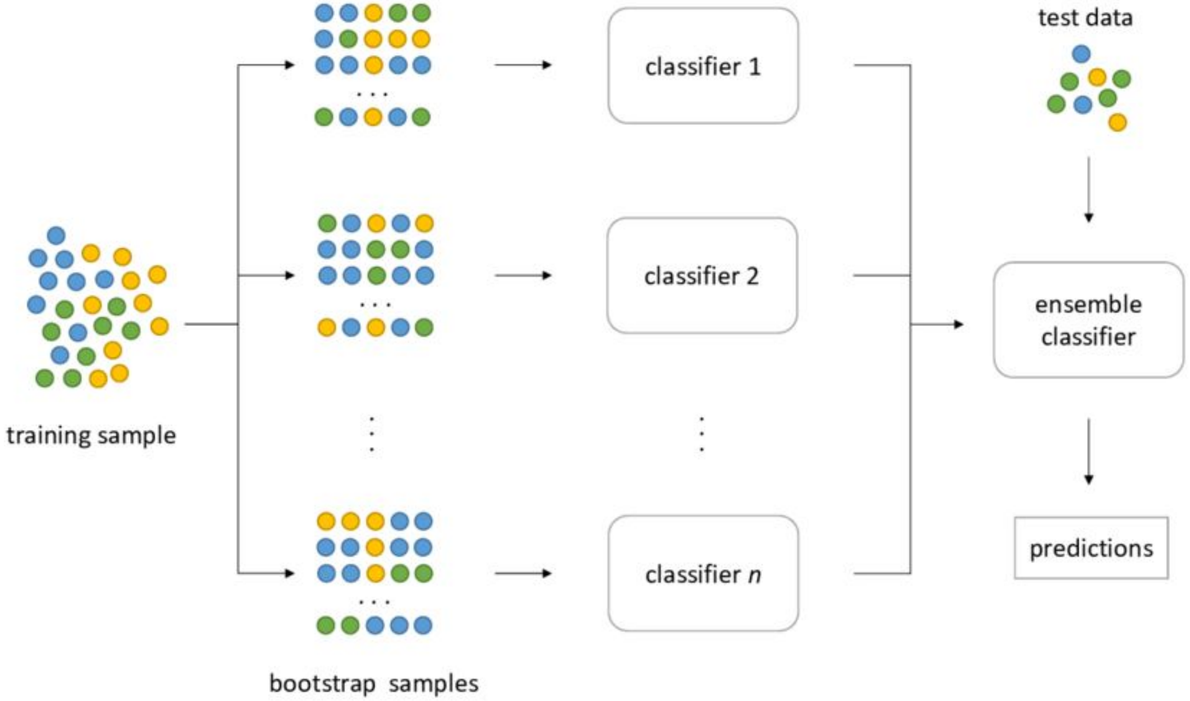
HUDSON AND THAMES

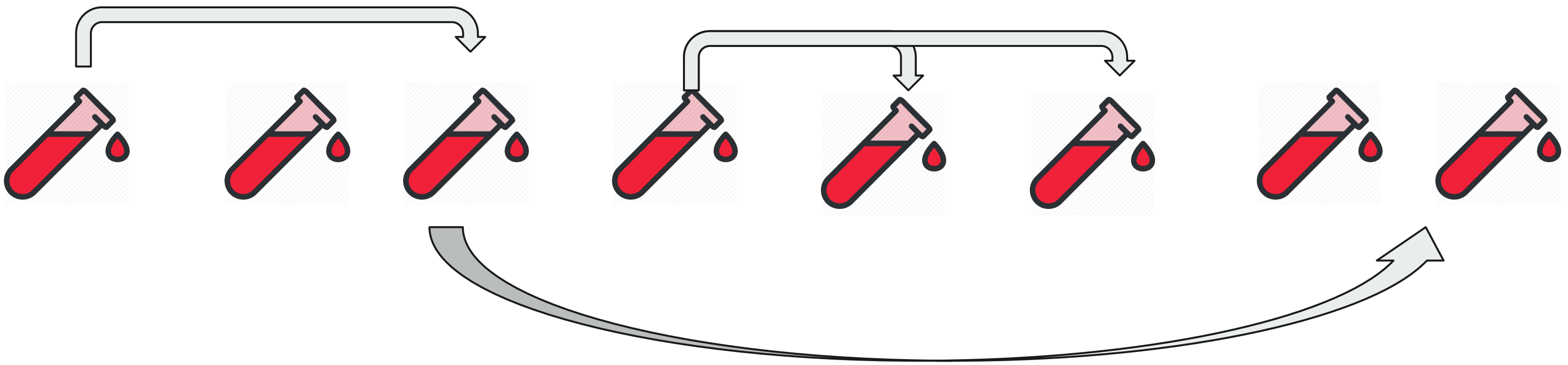
# Bagging in Financial Machine Learning

An open source way of work.

# Ensemble models and Bagging



# Blood sample example.



# Sequential Bootstrapping. Indicator Matrix

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 |
| 3 | 0 | 1 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 |
| 5 | 0 | 1 | 0 | 1 |
| 6 | 0 | 1 | 0 | 1 |
| 7 | 0 | 0 | 0 | 1 |
| 8 | 0 | 0 | 0 | 1 |

# Sequential Bootstrapping. Step 1

The probability on the first step is uniformly distributed. Let's assume that sample # 1 was drawn

| | 1 | 1 | Σ |
|---|---|---|---|
| 1 | 1 | 1 | 2 |
| 2 | 1 | 1 | 2 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |

The first sample average uniqueness:
$(½ + ½)/2 = ½$

# Sequential Bootstrapping. Step 1

| | 1 | 2 | Σ |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 2 | 1 | 1 | 2 |
| 3 | 0 | 1 | 1 |
| 4 | 0 | 1 | 1 |
| 5 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |

The second sample average uniqueness:
$(\frac{1}{2} + 1 + 1 + 1 + 1)/5 = 9/10$

# Sequential Bootstrapping. Step 1

| | 1 | 3 | Σ |
|---|---|---|---|
| 1 | 1 | 0 | 1 |
| 2 | 1 | 0 | 1 |
| 3 | 0 | 1 | 1 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |

The third sample average uniqueness:
(1)/1 = 1

# Sequential Bootstrapping. Step 1

The probability of a sample being drawn is based on sample uniqueness:

P = [0.147, 0.264, 0.294, 0.294]

As you can see SB penalizes repeating samples

| | 1 | 4 | Σ |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 1 | 1 |
| 6 | 0 | 1 | 1 |
| 7 | 0 | 1 | 1 |
| 8 | 0 | 1 | 1 |

The fourth sample average uniqueness:
(1+1+1+1)/4 = 1

Let’s say the third sample was drawn

HUDSON AND THAMES

# Sequential Bootstrapping. Step 2

| | 1 | 3 | 1 | Σ |
|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 2 |
| 2 | 1 | 0 | 1 | 2 |
| 3 | 0 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 |

The first sample average uniqueness:
(½ + ½ )/2 = 1/2

# Sequential Bootstrapping. Step 2

| | 1 | 3 | 2 | Σ |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 |
| 2 | 1 | 0 | 1 | 2 |
| 3 | 0 | 1 | 1 | 2 |
| 4 | 0 | 0 | 1 | 1 |
| 5 | 0 | 0 | 1 | 1 |
| 6 | 0 | 0 | 1 | 1 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 |

The second sample average uniqueness:
(½ + ½ + 1 + 1 + 1)/5= 4/5

# Sequential Bootstrapping. Step 2

| | 1 | 3 | 3 | Σ |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 1 |
| 3 | 0 | 1 | 1 | 2 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 |
| 6 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 |

The third sample average uniqueness:
(1)/2 = ½

# Sequential Bootstrapping. Step 2

Probability of being drawn on Step 2:

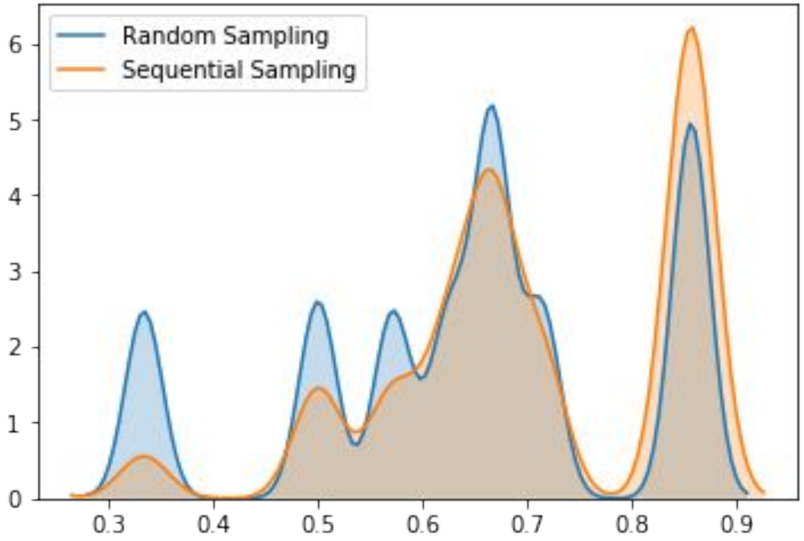
P = [0.126, 0.304, 0.19, 0.38]

Despite the fact that, sample #2 is the most overlapping, SB penalises already drawn samples to increase the uniqueness of bootstrapped data set

| | 1 | 3 | 4 | Σ |
|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 |
| 2 | 1 | 0 | 0 | 1 |
| 3 | 0 | 1 | 0 | 1 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 1 | 1 |
| 6 | 0 | 0 | 1 | 1 |
| 7 | 0 | 0 | 1 | 1 |
| 8 | 0 | 0 | 1 | 1 |

The fourth sample average uniqueness: (1+1+1+1)/4 = 1

# Sequential Bootstrapping. Monte-Carlo simulations




HUDSON
AND THAMES
Thank You