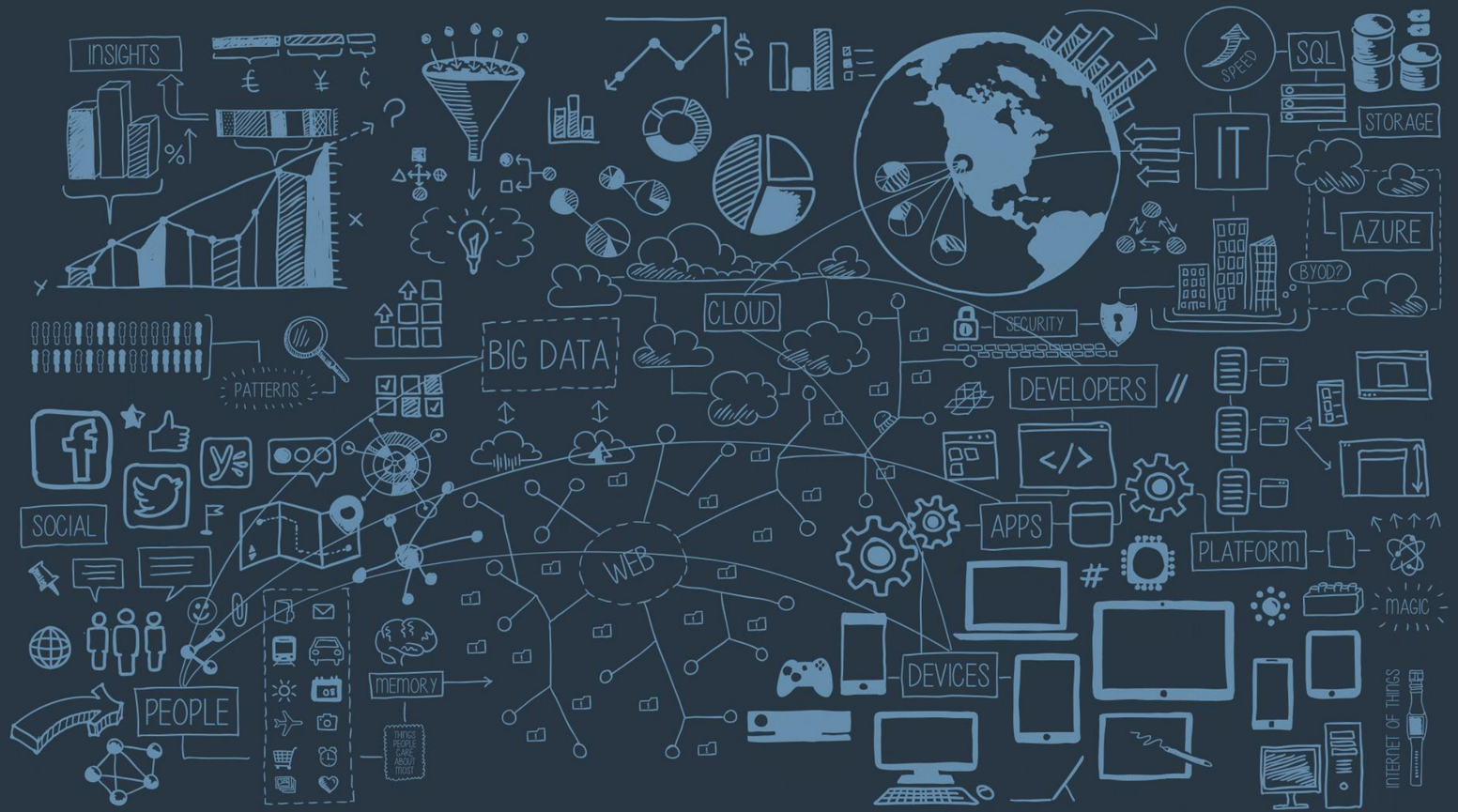
Data Pipeline Selection and Optimization

DOLAP 2019

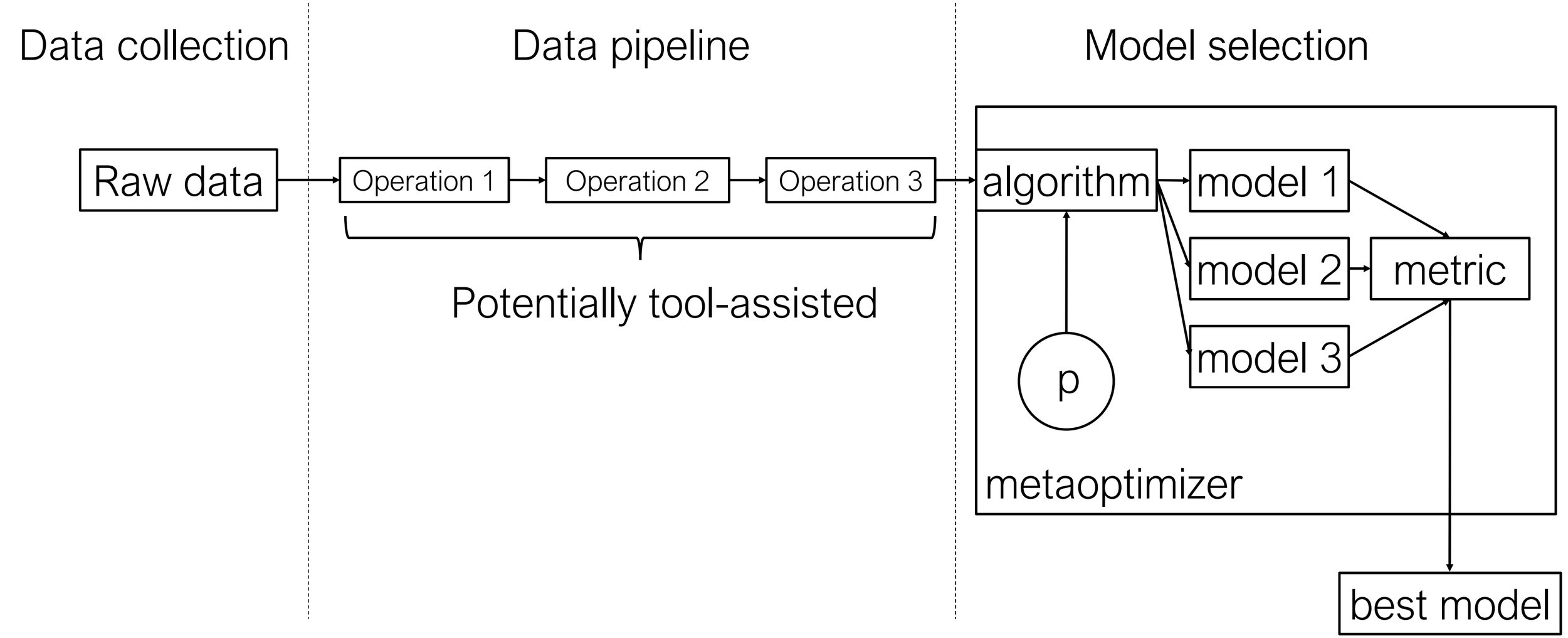


Alexandre Quemy
IBM, Data and AI
Politechnika Poznańska

IBM Poland Software Lab

IBM

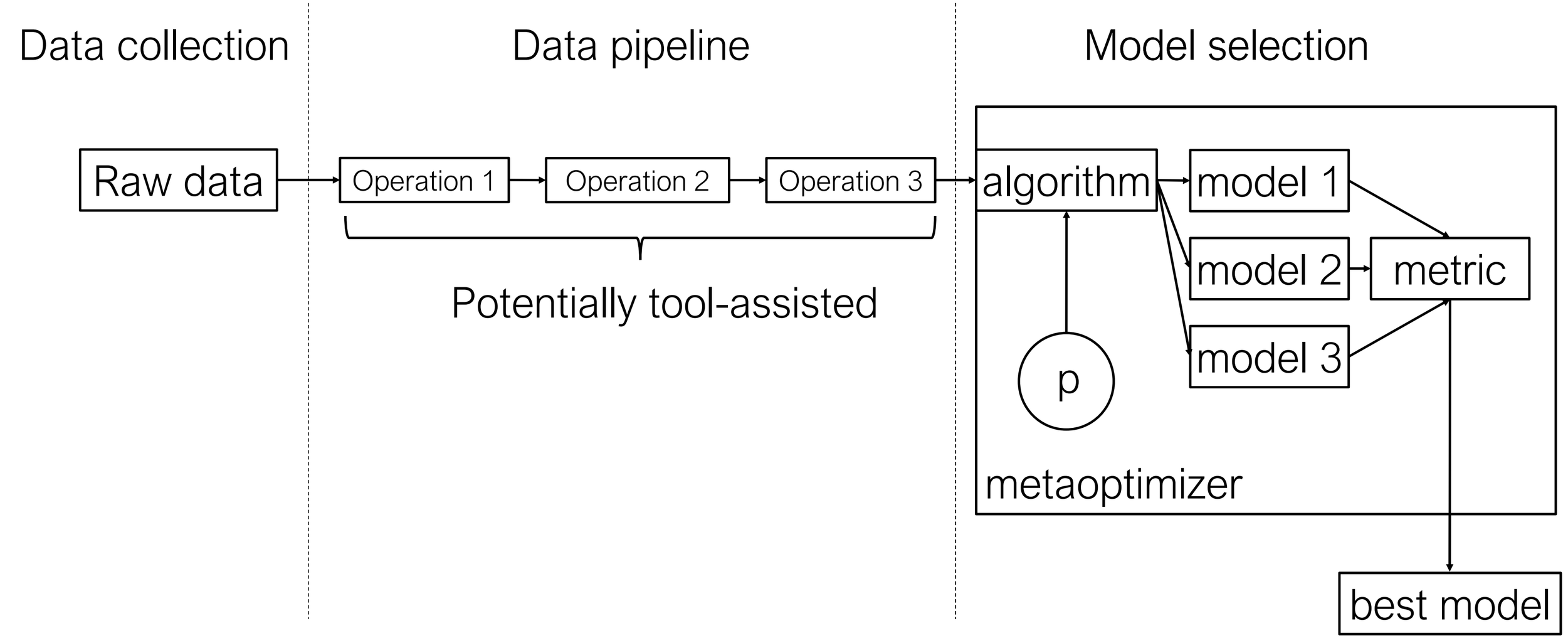
The usual workflow



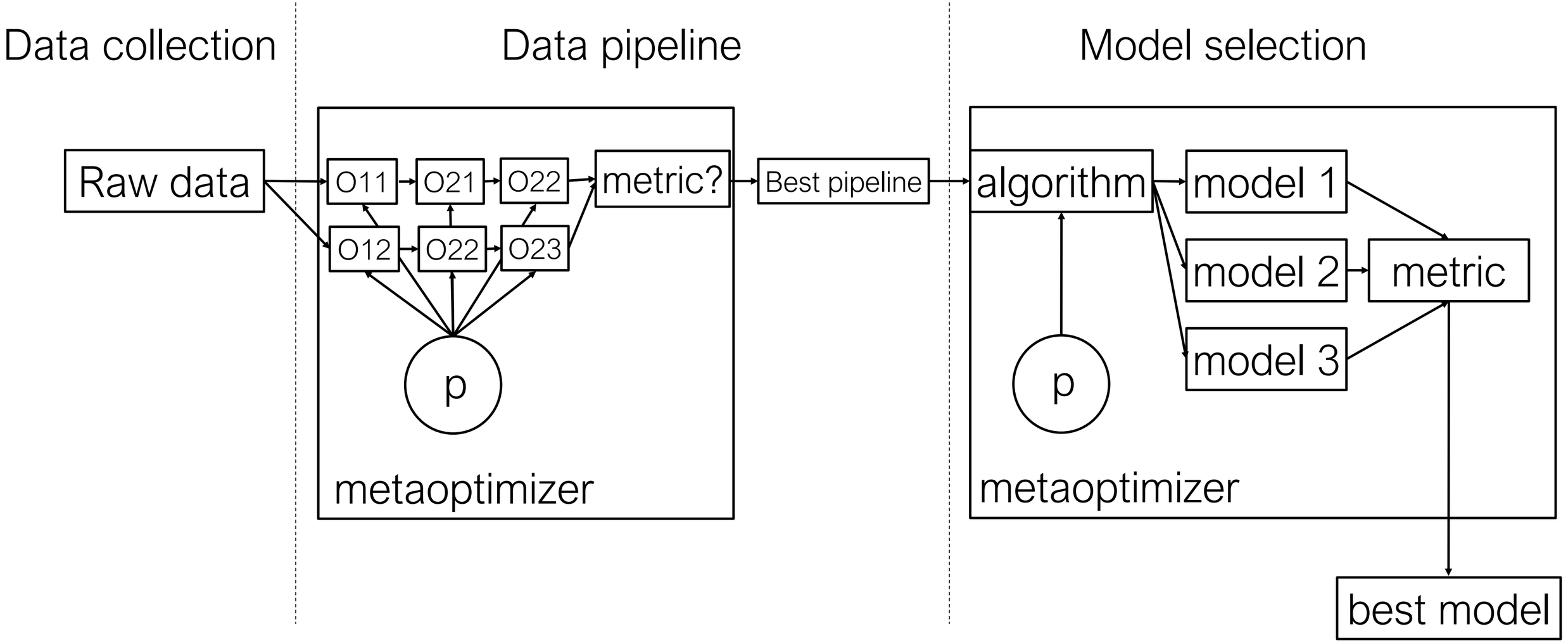
The hard life of data scientists

→ **Dealing with missing value:**
  → Discarding? Row? Column?
  → Imputation? What imputation? Mean? Median? Model-based? What model?
→ **Imbalanced datasets:**
  → Downsampling? Oversampling?
  → Nothing? What bias it implies?
→ **Data too large:**
  → Dimensional reductions: what algorithm? PCA? normalization or not?
  → Subsampling: what technique? what bias?
→ **Outliers detection and curation:**
  → What threshold? What deviation measure?
  → Trimming? Truncating? Censoring? Winsorizing?
→ **Encoding for method domain requirements:**
  → Discretization? Grid? What step? Cluster? What method? What hyperparameter?
  → Categorial encoder? Binary? Hot-One? Helmert? Backward Difference?
→ **NLP:**
  → How many tokens?
  → Size of m-grams?

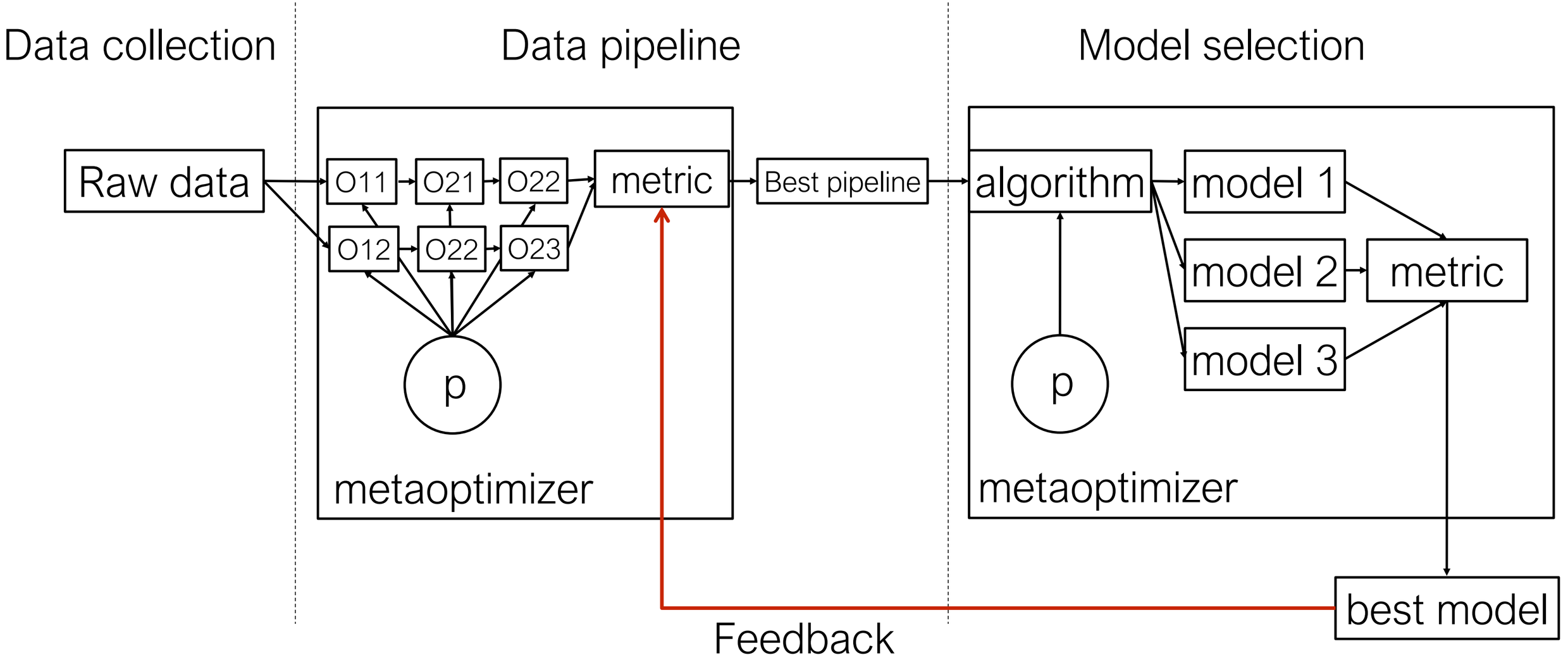
The usual workflow



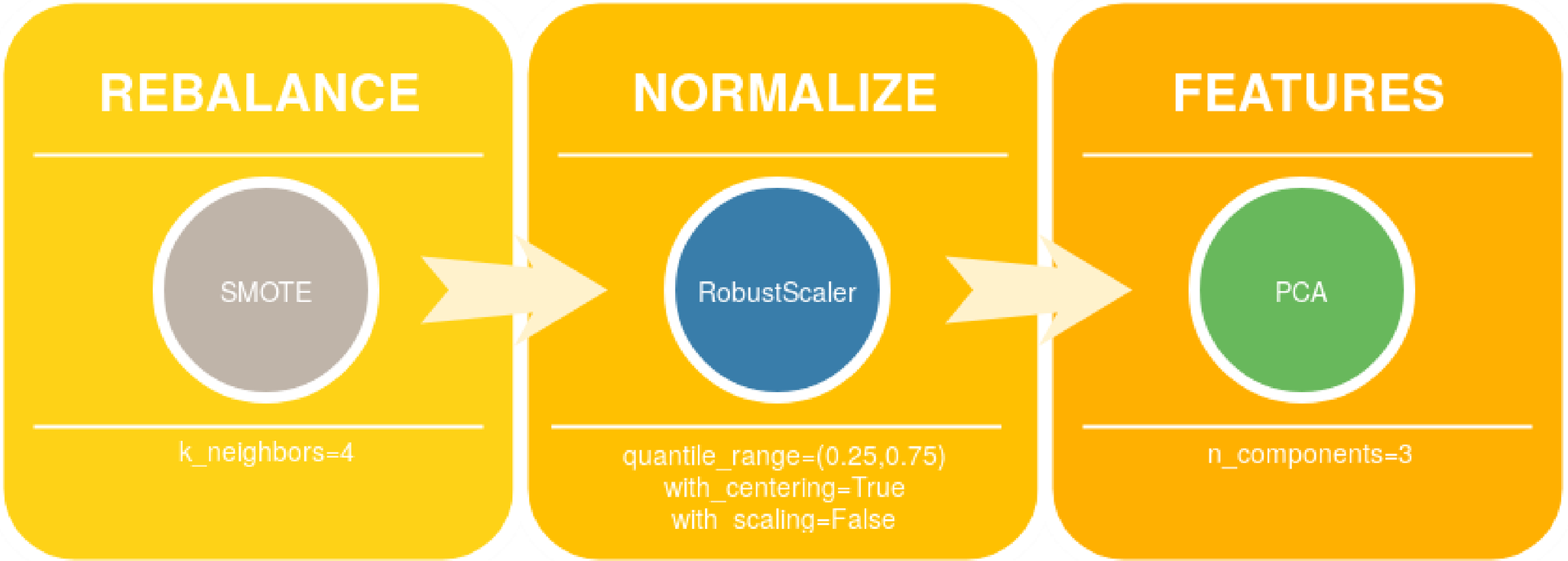
The workflow proposed in the paper



The workflow proposed in the paper



Pipeline prototype



Rebalance: 4 operators
Normalize: 5 operators
Features: 4 operators

Baseline: (Id, Id, Id)

Configuration space: 4750 configurations

Protocol

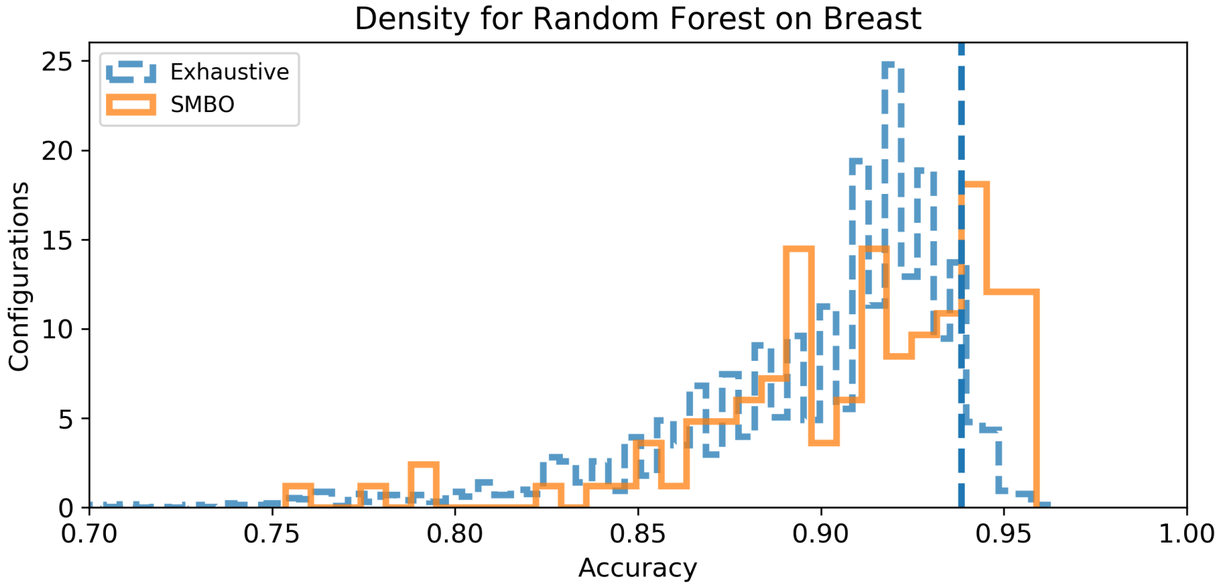
- **Datasets:** Breast, Iris, Wine.
- **Methods:** SVM, Random Forest, Neural Network, Decision Tree.
- **Dataset split:** 60% for training set, 40% for test set.
- **Pipeline configuration space size:** 4750 configurations.
- **Performance metric:** Cross-validation accuracy
- **Metaoptimizer:** Tree Parzen Estimator (hyperopt)
- **Budget:** 100 configurations (~2% of the space)

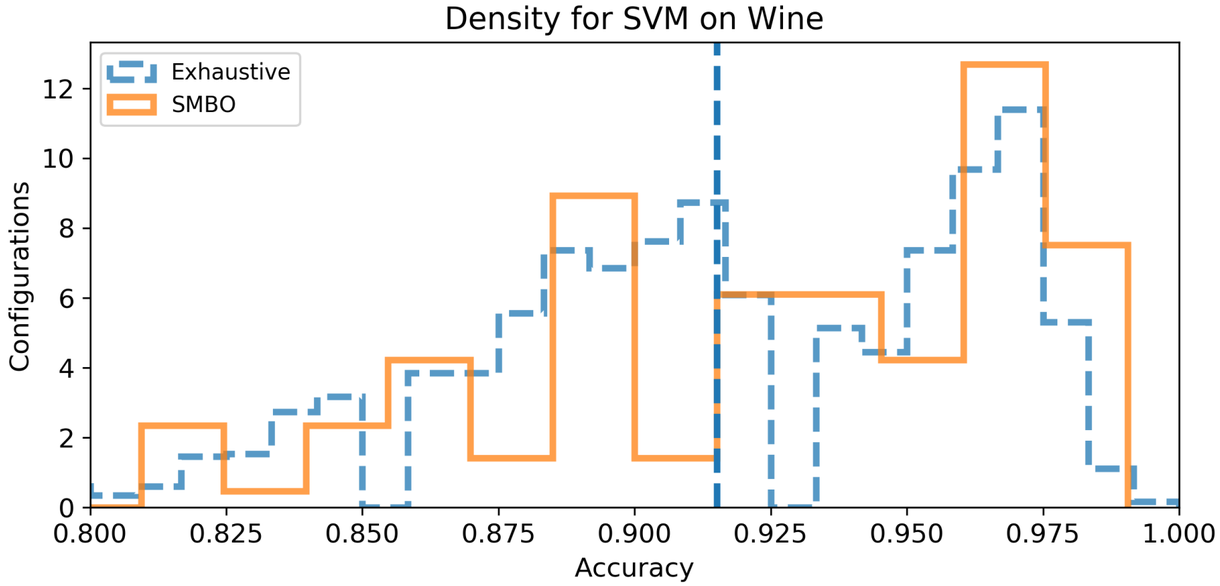
No algorithm hyperparameter tuning!

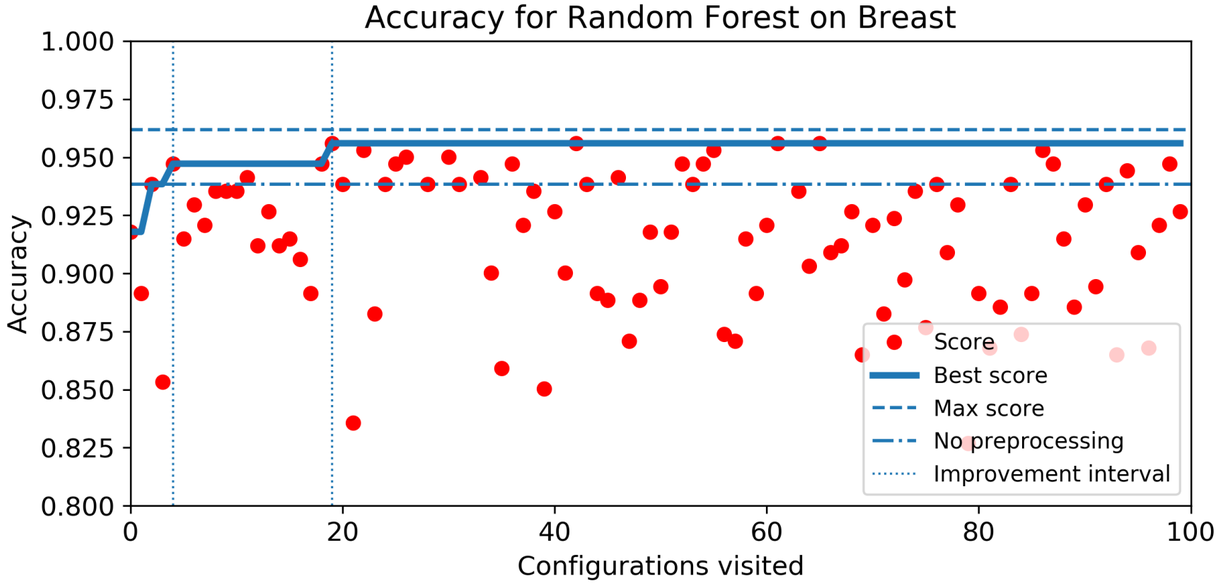
$\Rightarrow$ We want to quantify the influence of data pipeline

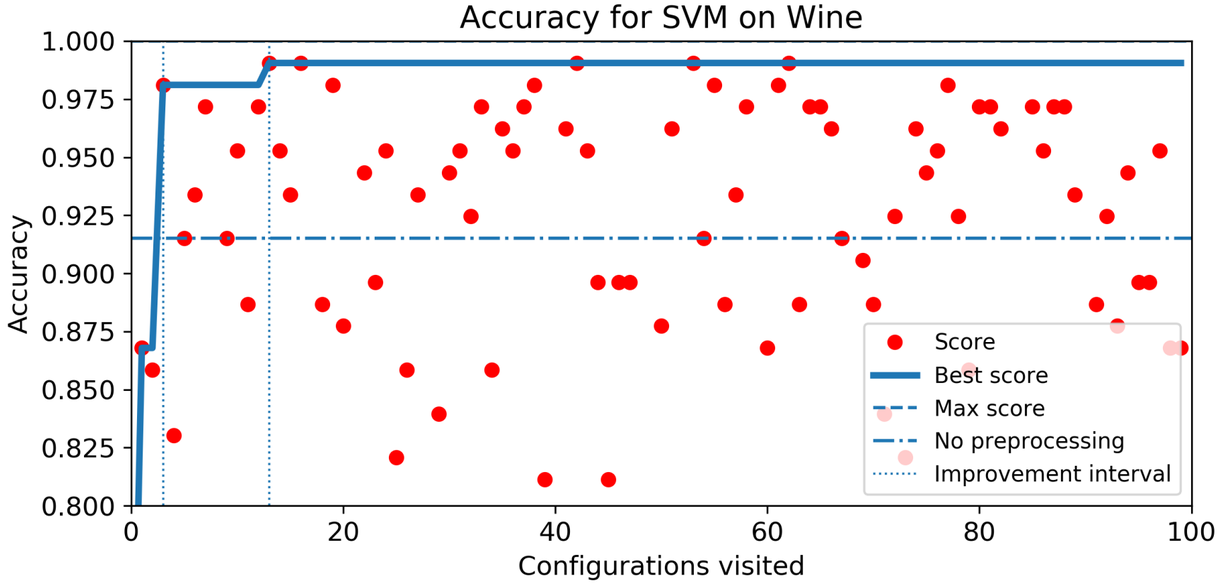
Exhaustive search to compare between baseline and max score.

Results


Density for Random Forest on Breast
Exhaustive
SMBO
Configurations
Accuracy
0
5
10
15
20
25
0.70
0.75
0.80
0.85
0.90
0.95
1.00


Density for SVM on Wine
Exhaustive
SMBO
Configurations
Accuracy
0
2
4
6
8
10
12
0.800
0.825
0.850
0.875
0.900
0.925
0.950
0.975
1.000


Accuracy for Random Forest on Breast
Score
Best score
Max score
No preprocessing
Improvement interval
Accuracy
Configurations visited
0.800
0.825
0.850
0.875
0.900
0.925
0.950
0.975
1.000
0
20
40
60
80
100

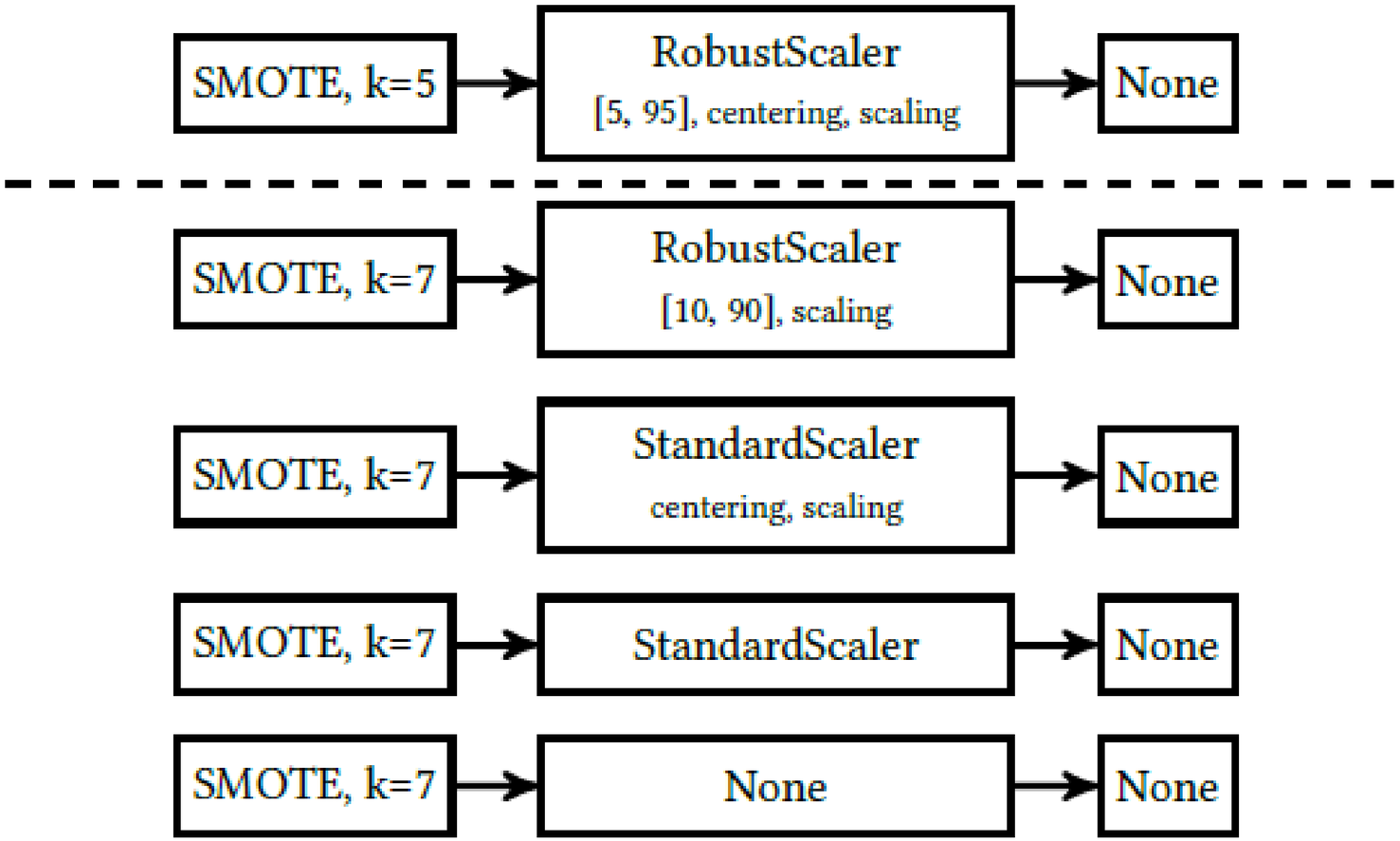

Accuracy for SVM on Wine
Score
Best score
Max score
No preprocessing
Improvement interval
Accuracy
Configurations visited
0.800
0.825
0.850
0.875
0.900
0.925
0.950
0.975
1.000
0
20
40
60
80
100

	Baseline	Exhaustive	SMBO	SMBO (norm.)	Imp. Inter.
			Iris		
SVM	0.9667	0.9889	0.9778	0.9831	[11, 11]
Random Forest	0.9222	0.9778	0.9667	0.9828	[8, 27]
Neural Net	0.9667	0.9889	0.9778	0.9831	[17, 17]
Decision Tree	0.9222	0.9889	0.9889	1.0000	[1, 83]
			Breast		
SVM	0.9501	0.9765	0.9765	1.0000	[12, 20]
Random Forest	0.9384	0.9619	0.9560	0.9780	[4, 19]
Neural Net	0.9326	0.9765	0.9707	0.9903	[1, 7]
Decision Tree	0.9296	0.9619	0.9589	0.9900	[0, 67]
			Wine		
SVM	0.9151	1.0000	0.9906	0.9811	[3, 13]
Random Forest	0.9623	0.9906	0.9811	0.9818	[5, 20]
Neural Net	0.9057	0.9906	0.9906	1.0000	[1, 25]
Decision Tree	0.9057	0.9811	0.9811	1.0000	[5, 35]

In average, with **20 iterations** (**0.42%** of the search space):

1. decrease of error by **58%** compared to the baseline
2. **98.92%** in the normalized score space)

How close are we from the optimal pipeline?



A solution for Euclidian space

Definition 3.2 (Mean Absolute Deviation).

$$\mathrm{MAD}(\mathbf{p}^*, r) = \frac{1}{N}\sum_{i=1}^{N}|p_i^* - r|$$

Definition 3.3 (Normalized Mean Absolute Deviation).

$$\mathrm{NMAD}(\mathbf{p}^*, r) = \frac{1}{K}||\mathrm{MAD}(\mathbf{p}^*, r)||_1$$

N: number of algorithms
K: dimension of the configuration space
r: a reference point
p*: sample of optimal configurations

For each optimal configuration r:

1. Build the sample w.r.t. to the algorithms:

→ For each algorithm, select the optimal point that is the closest from the reference point.

2.Express the sample in normalized conf. space
3.Calculate the NMAD on the sample

Results on two datasets for text classification

Method	(n, k)	accuracy
ECHR		
Decision Tree	(5, 50000)	0.900
Neural Network	(5, 50000)	0.960
Random Forest	(3, 10000), (4, 10000), (5, 50000)	0.910
Linear SVM	(3, 50000), (4, 50000), (5, 50000)	0.921
Newsgroup		
Decision Tree	(4, 5000), (4, 100000)	0.889
Neural Network	(5, 50000)	0.953
Random Forest	(3, 10000)	0.931
Linear SVM	(2, 100000)	0.946

ECHR	
Point	NMAD
(5, 50000)	0
(3, 10000)	0.275
(4, 10000)	0.213
(3, 50000)	0.175
(4, 50000)	0.094

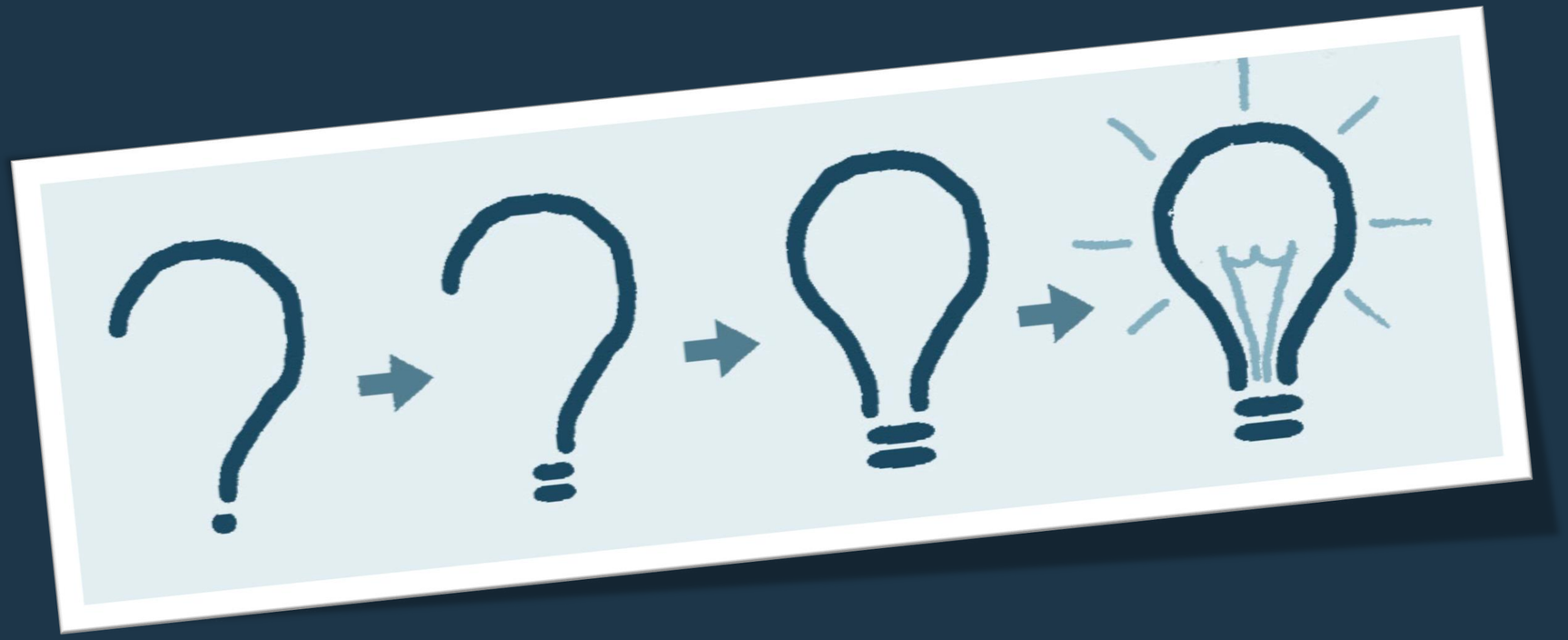
Newsgroup	
Point	NMAD
(4, 5000)	0.306
(4, 100000)	0.300
(5, 50000)	0.356
(3, 10000)	0.294
(2, 100000)	0.362

Future work

Work in progress:

→ Tests on larger configuration spaces.
→ Online architecture.
→ Metric between pipelines

Thank you



Don’t forget the poster session!

IBM