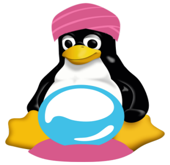
Machine Learning and Configurable Systems: A Gentle Introduction

(tutorial at SPLC'19)

Hugo Martin, Juliana Alves Pereira,
Paul Temple, Mathieu Acher

https://github.com/VaryVary/



Learning Software Configuration Spaces: A Systematic Literature Review

Hugo Martin
hugo.martin@irisa.fr

Juliana Alves Pereira
juliana.alves-pereira@irisa.fr

Paul Temple
paul.temple@unamur.be

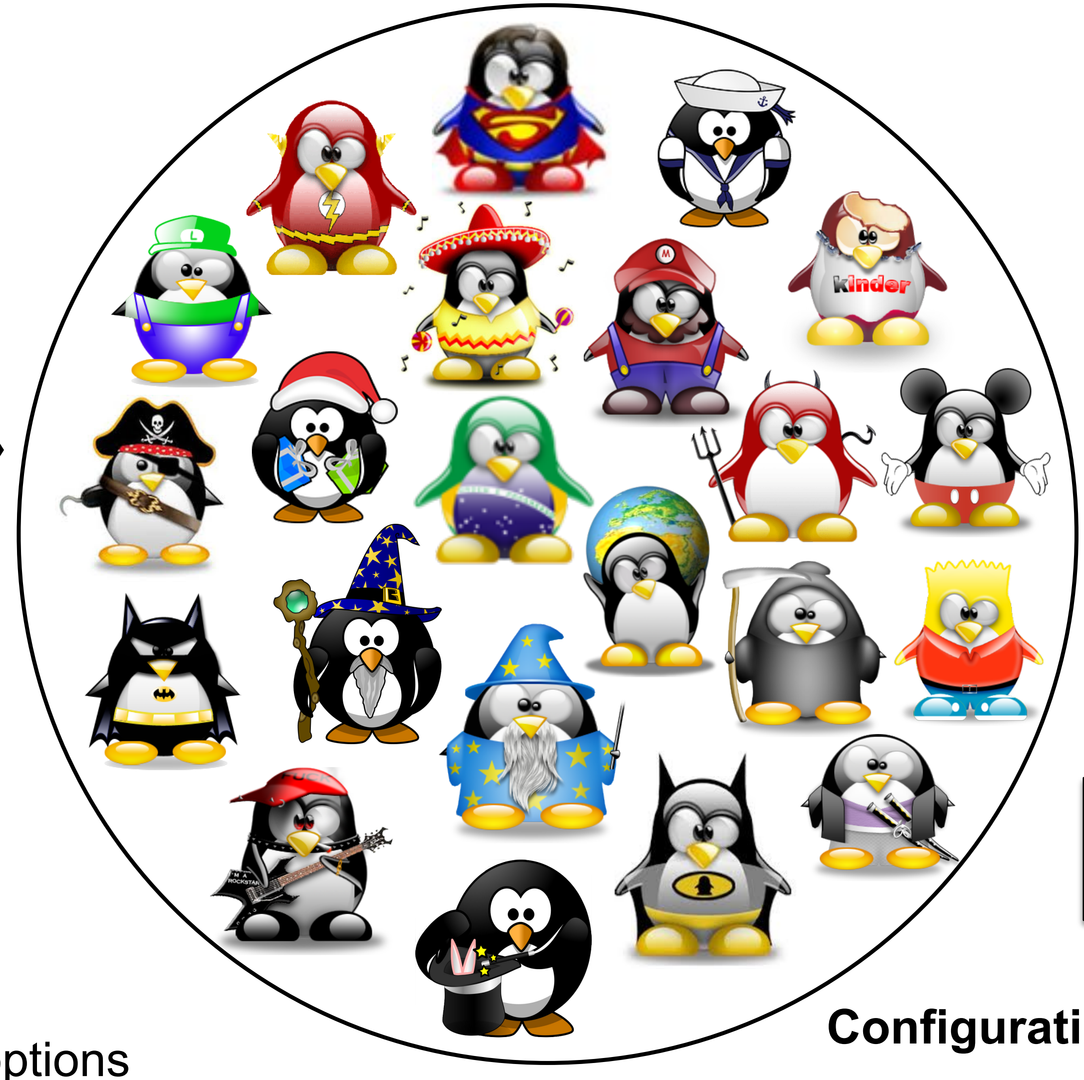
Mathieu Acher
mathieu.acher@irisa.fr

VaryVary: https://varyvary.github.io/
University of Rennes 1, Inria, CNRS, IRISA, France

SPLC 2019, September 9, Paris, France

Software Variability

Features



Configuration space

Sample

Learning

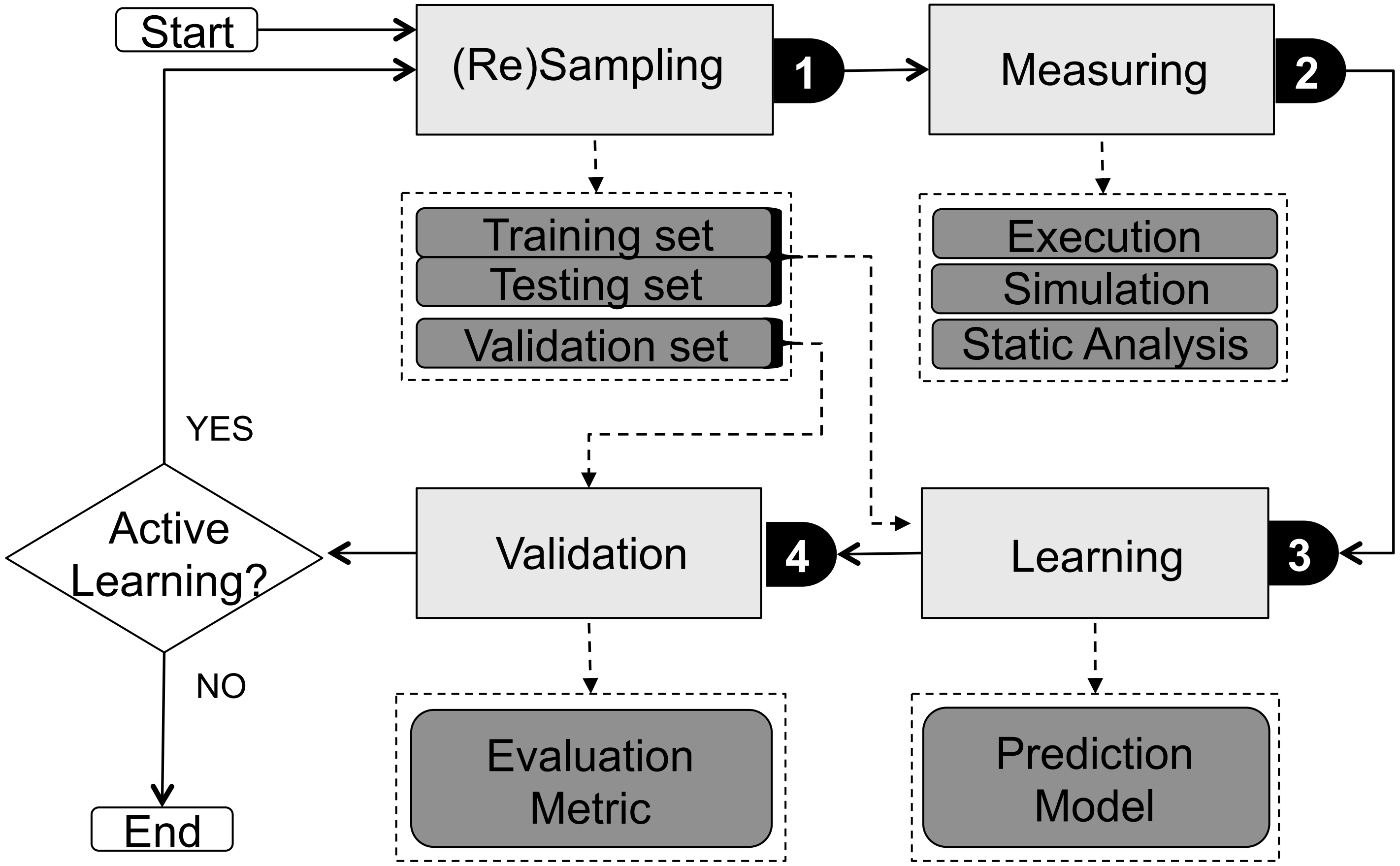
How they can be measured and evaluated?

Linux™

15,000+ features

Tri-state values and numerical options

Learning Stages



- **Research Questions:**

RQ1. What are the **applications** addressed for learning techniques to explore configuration spaces?

RQ2. Which **sampling** methods are used by these techniques?

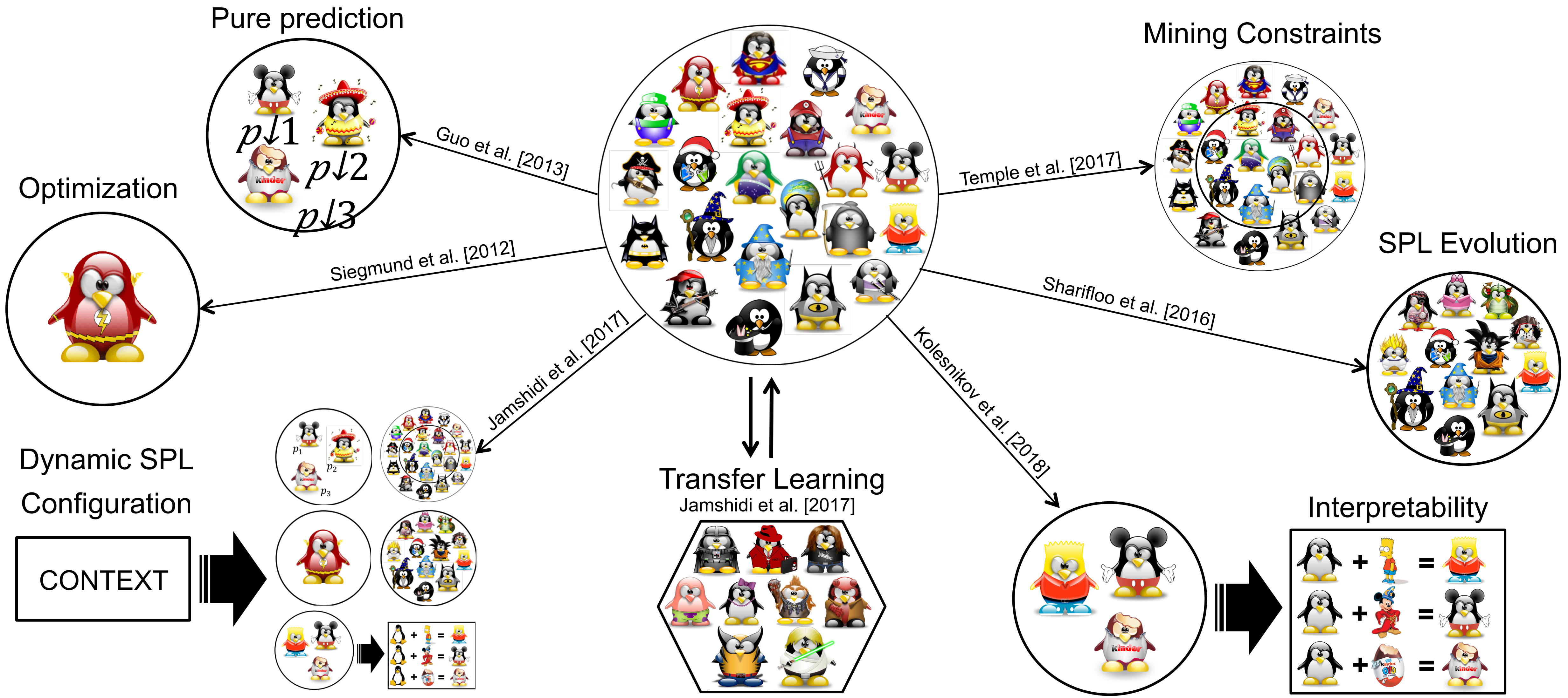
RQ3. How do the proposed techniques deal with **measurements** of non-functional properties?

RQ4. Which **learning** techniques have been proposed in the literature?

RQ5. How are these techniques **validated**?

RQ6. What are the **limitations** faced by the current techniques and open challenges that need attention in the future?

RQ1: Applicability



RQ1: Applicability

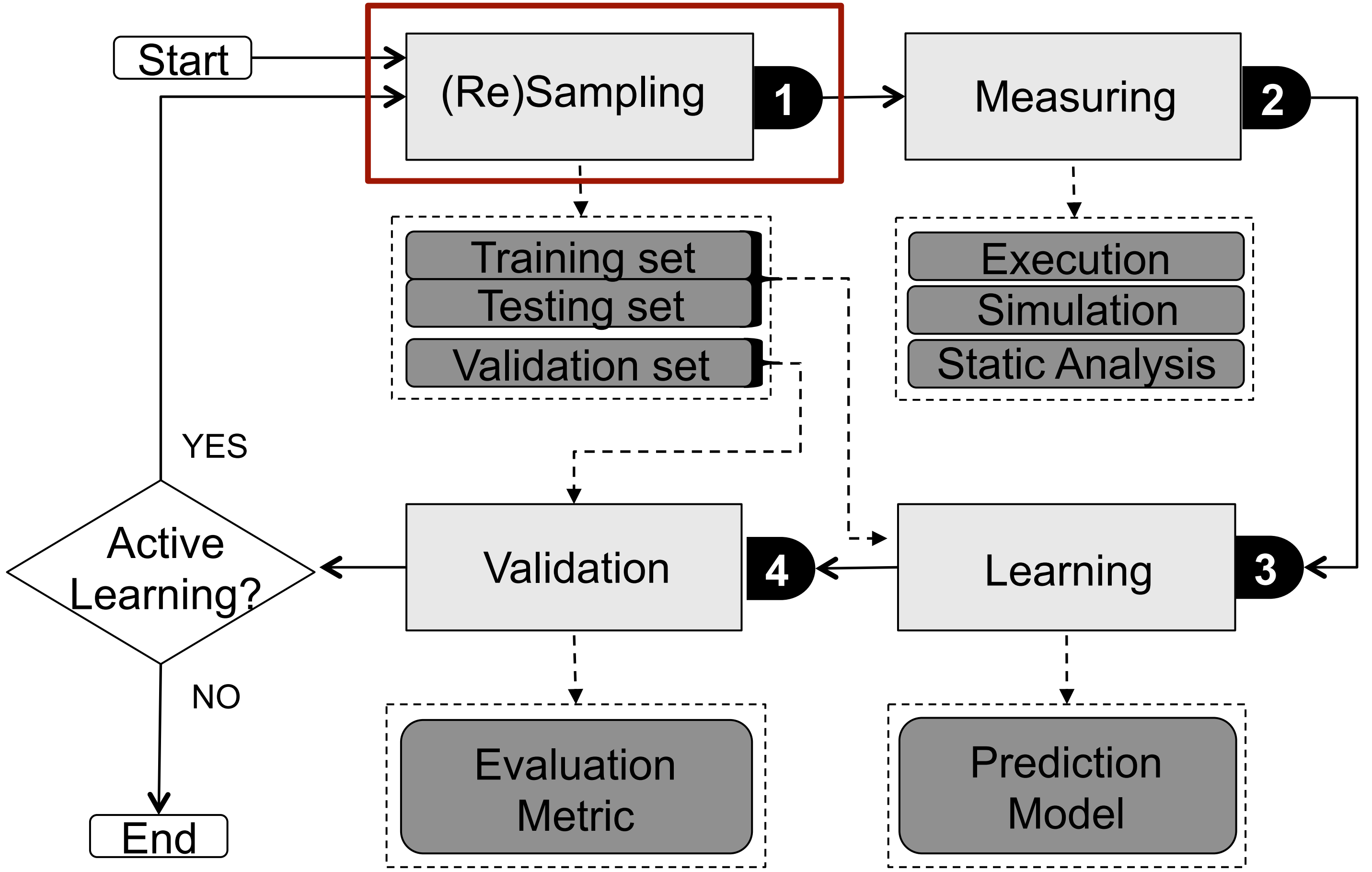
App.	Reference
A1	[65, 46, 29, 66, 36, 37, 38, 39, 67, 68, 69, 70, 71, 72, 44, 40, 73, 74, 75, 76, 77]
A2	[43, 47, 78, 45, 40, 79, 80]
A3	[30, 81, 54, 31, 82, 32, 83, 84, 41, 85, 33, 34, 86, 12, 40, 87, 88, 35, 89, 90, 91, 92, 93, 94, 95, 96, 97]
A4	[42, 56, 57, 58, 59, 9, 55, 98, 40, 79, 99]
A5	[60, 100, 61, 62, 63, 64, 40, 9, 55, 101]
A6	[98, 12, 40]

Table 2: Applicability of selected primary studies. *A1*: Pure Prediction; *A2*: Interpretability of Configurable Systems; *A3*: Optimization; *A4*: Dynamic Configuration; *A5*: Mining Constraints; *A6*: SPL Evolution.

Learning Stages

What sampling technique to use?

What is an ideal sample size?



- **Research Questions:**

RQ1. What are the **applications** addressed for learning techniques to explore configuration spaces?

RQ2. Which **sampling** methods are used by these techniques?

RQ3. How do the proposed techniques deal with **measurements** of non-functional properties?

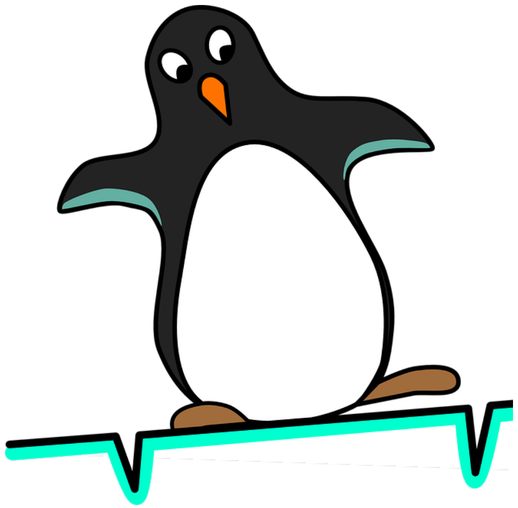
RQ4. Which **learning** techniques have been proposed in the literature?

RQ5. How are these techniques **validated**?

RQ6. What are the **limitations** faced by the current techniques and open challenges that need attention in the future?

RQ2: Samplir

- Multi-objective prob
  - SAT constrained s
  - The "training sam
  - Not infinite resour



Sample Method	Reference
Random	[46, 29, 42, 43, 30, 31, 60, 9, 67, 56, 68, 100, 59, 61, 39, 45, 54, 84, 72, 40, 86, 87, 76, 94, 91, 35, 99, 97]
Knowledge-wise heuristic	[36, 38, 40, 87, 95]
Feature-coverage heuristic	[72, 66, 37, 36, 38, 39, 69, 62, 73, 35]
Feature-frequency heuristic	[66, 37, 39, 36, 38, 82, 73, 35, 80]
Family-based simulation	[74]
Multi-start local search	[96]
Plackett-Burman design	[39, 35]
central composite design	[35]
D-optimal design	[35]
Breakdown	[32]
Sequence type trees	[55]
East-west sampling	[84]
Exemplar sampling	[84]
Constrained-driven sampling	[63]
Diameter uncertainty strategy	[33]
Historical configurations	[41, 89]
Latin hypercube sampling	[34, 85, 90]
Neighborhood sampling	[58, 44]
Input-based clustering	[93]
Distance-based sampling	[72, 88]
Genetic sampling	[83, 85, 101]
Interaction tree discovery	[77]
Arbitrarily chosen	[65, 81, 78, 57, 64, 75, 79, 99]

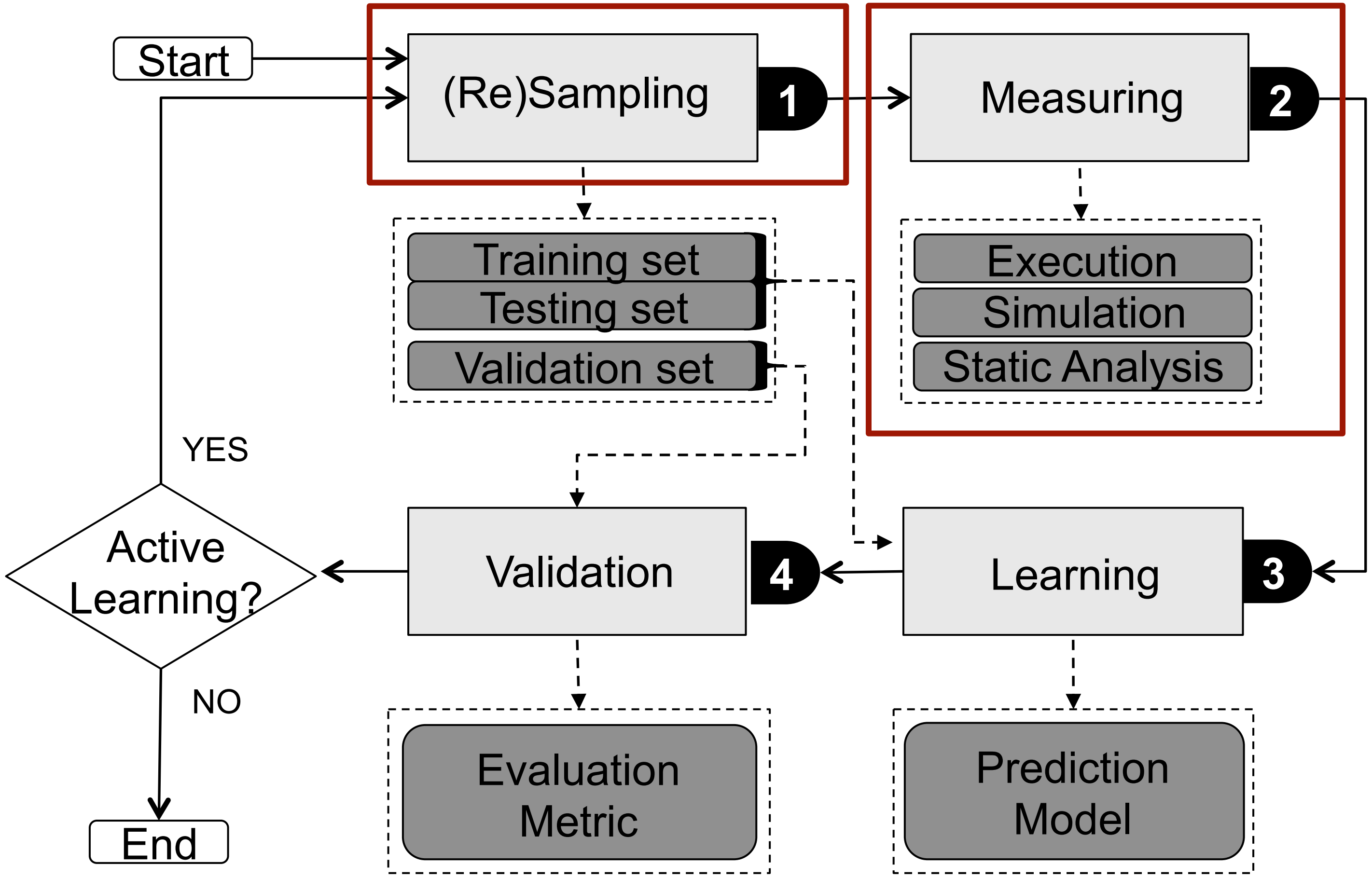
	Sample Method
6, 45, 51, 52]	Random sampling
, 57, 59]	Feature-frequency heuristic
, 61, 63]	Feature-coverage heuristic
	Interaction-wise heuristic
	Higher-order interactions
	Hot-spot features
	Arbitrarily chosen
	Plackett-Burman design
	All valid configurations

Learning Stages

How to measure a sample?

Which properties are measured?

How to measure properties?



- ## Research Questions:

RQ1. What are the **applications** addressed for learning techniques to explore configuration spaces?

RQ2. Which **sampling** methods are used by these techniques?

RQ3. How do the proposed techniques deal with **measurements** of non-functional properties?

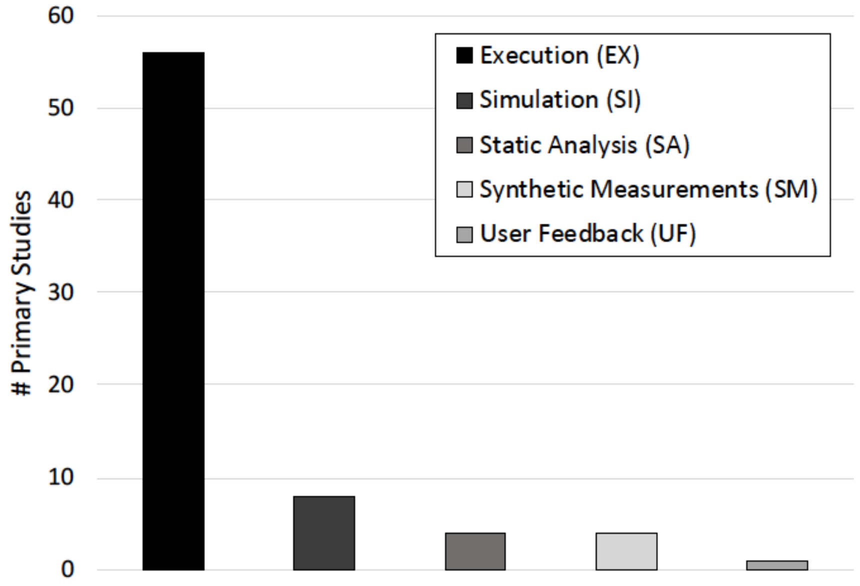
RQ4. Which **learning** techniques have been proposed in the literature?

RQ5. How are these techniques **validated**?

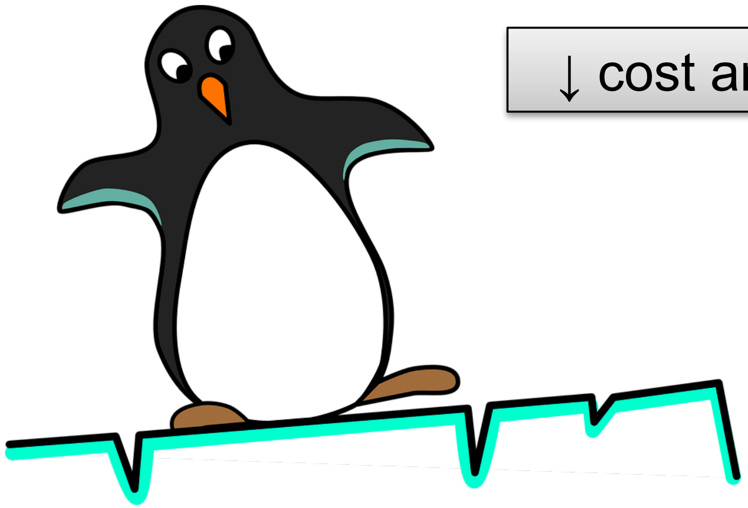
RQ6. What are the **limitations** faced by the current techniques and open challenges that need attention in the future?

RQ3: Measuring

- Measuring properties of configurations
  - Quantitative and qualitative measurements
  - Execution vs. simulation vs. static analysis vs. user feedback vs. synthetic measurements
  - Consider hardware and external influences
    - Some properties are less costly



Tool support: Sincero et al. [2010] and Siegmund et al. [2012]



↓ cost and ↓ measurement error

video quality

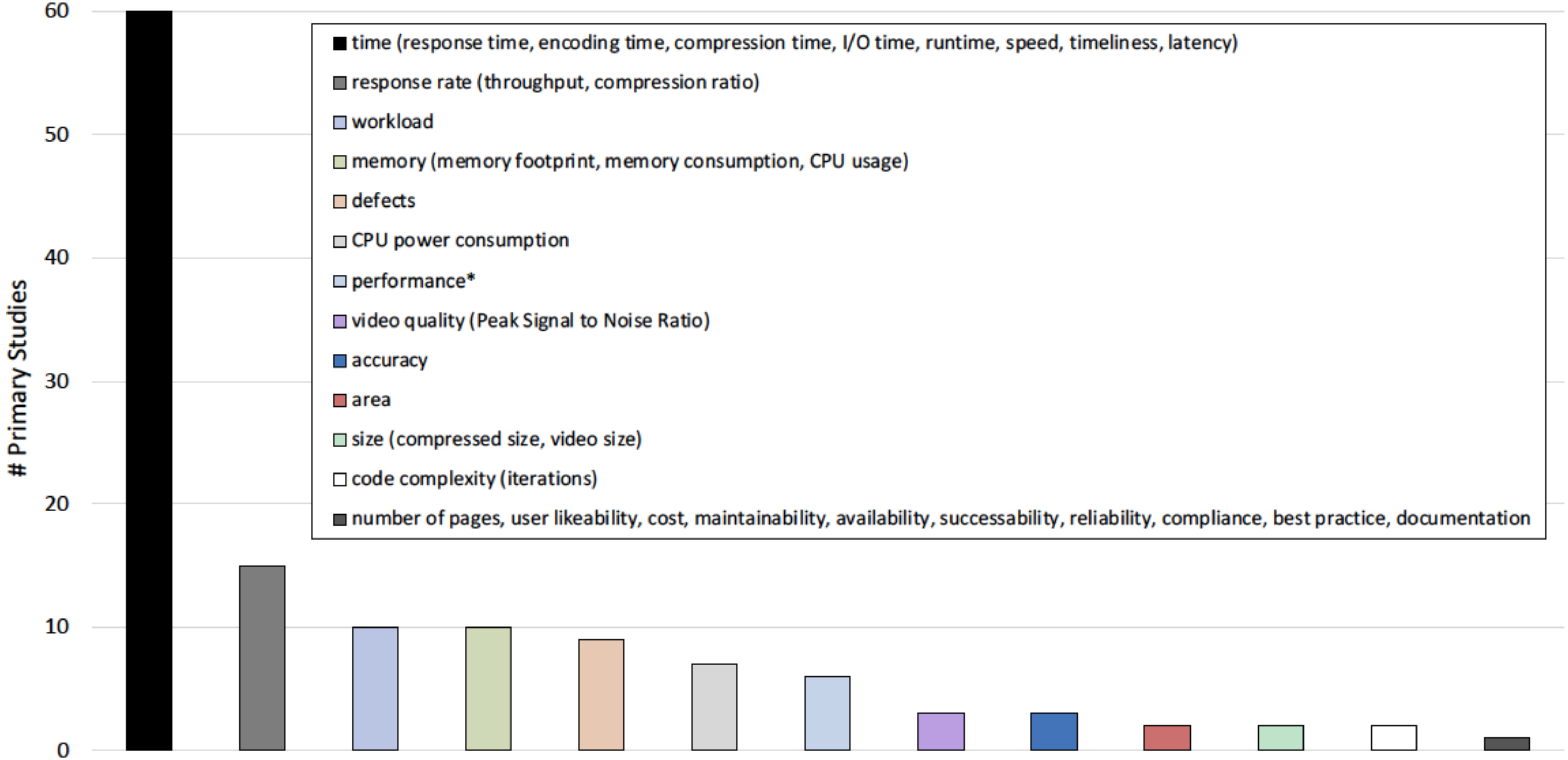
response time

code complexity

CPU power consumption

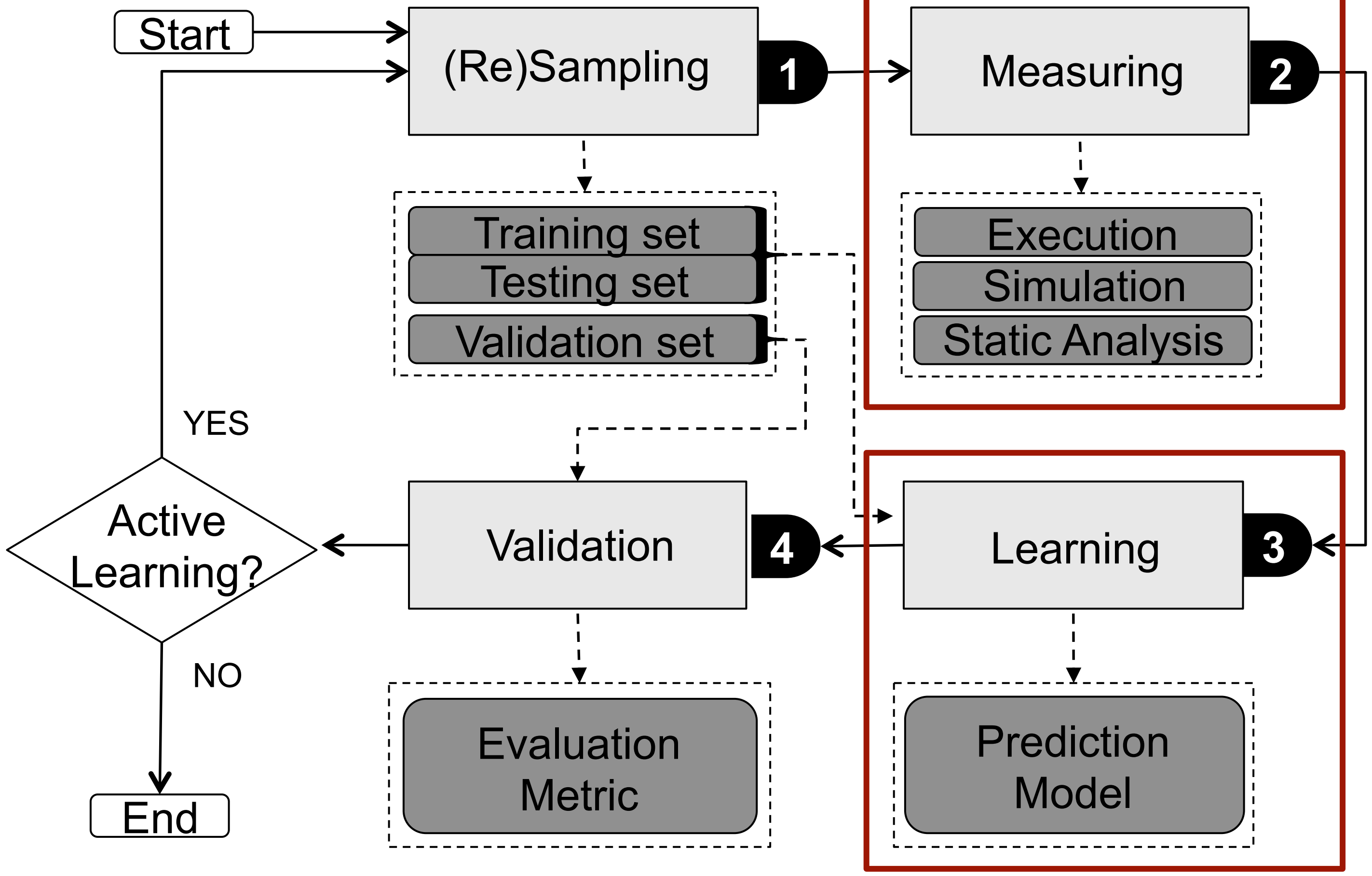
Non-Functional Property	Type	Non-Functional Property	Type
area	QT	Peak Signal to Noise Ratio	QT
compressed size	QT	response rate	QT
CPU power consumption	QT	response time	QT
CPU usage	QT	runtime	QT
encoding time	QT	throughput	QT
iterations	QT	video quality	QL
latency	QT	workload	QT
footprint	QT	code complexity	QT
performance[12]	QT	memory consumption	QT

RQ3: Measuring



Learning Stages

What learning technique to use?



- **Research Questions:**

RQ1. What are the **applications** addressed for learning techniques to explore configuration spaces?

RQ2. Which **sampling** methods are used by these techniques?

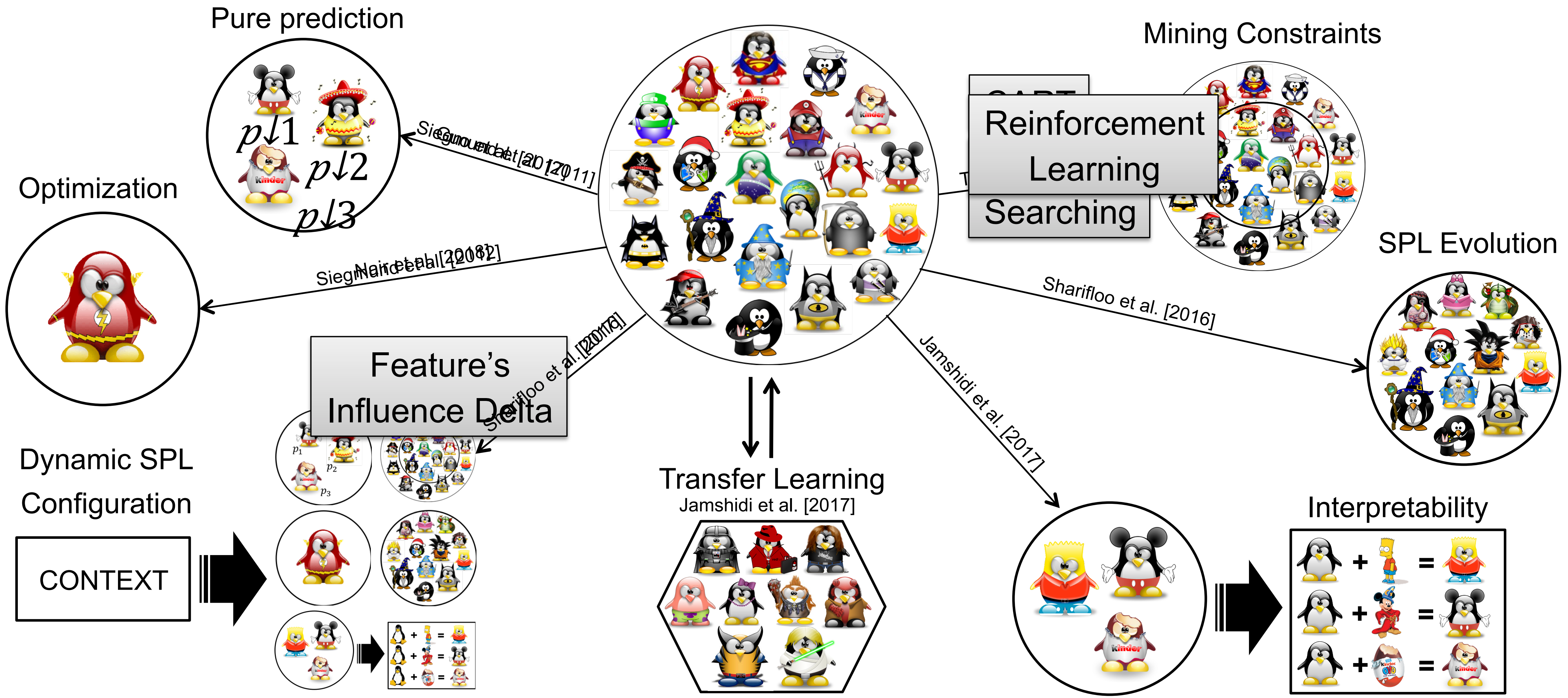
RQ3. How do the proposed techniques deal with **measurements** of non-functional properties?

RQ4. Which **learning** techniques have been proposed in the literature?

RQ5. How are these techniques **validated**?

RQ6. What are the **limitations** faced by the current techniques and open challenges that need attention in the future?

RQ4: Learning



RQ4: Learning

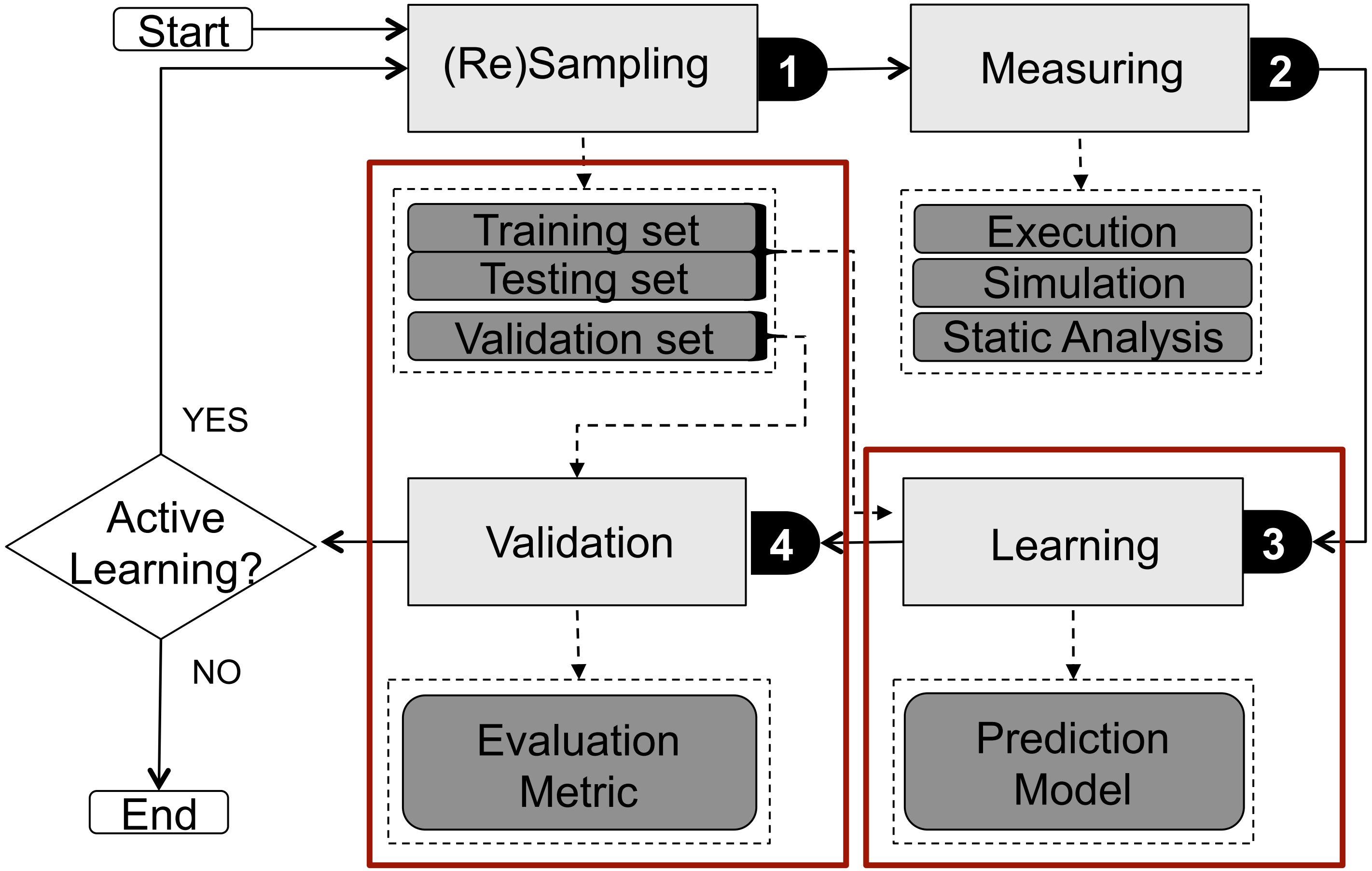
Reference	Learning Technique	Applicability
[69]	Adaptive ELAs, Multi-Class FDA-CIT, Static Error Locating Arrays (ELAs), Ternary-Class FDA-CIT, Test Case-Aware CIT, Traditional CIT	*A1*
[67]	Bagging	*A1*
[68]	Fourier Learning of Boolean Functions	*A1*
[70]	Frequent Item Set Mining	*A1*
[74]	Graph Family-Based Variant Simulator	*A1*
[71]	Implicit Path Enumeration Technique (IPET)	*A1*
[75]	Naive Bayes	*A1*
[43, 47, 39, 56, 70, 72, 59, 87]	Step-Wise Linear Regression	*A1, A2, A3, A4*
[46, 29, 43, 81, 54, 31, 66, 60, 9, 67, 45, 32, 100, 84, 62, 55, 58, 12, 76, 77]	Classification and Regression Trees (CART)	*A1, A2, A3, A4, A5*
[40]	Kernel Density Estimation and NSGA-II	*A1, A2, A3, A4, A5, A6*
[36, 37, 82, 38]	Feature's Influence Delta	*A1, A3*
[86, 32, 42, 41, 44, 85, 33]	Gaussian Process Model	*A1, A3, A4*
[81, 57, 65, 73]	Linear Regression	*A1, A3, A4*

Reference	Learning Technique	Applicability
[67, 64, 75, 90, 94]	Random Forest	*A1, A3, A5*
[64, 61, 67, 92]	Support Vector Machine	*A1, A3, A5*
[64, 75, 77]	C4.5 (J48)	*A1, A5*
[78, 80]	Covariance Analysis	*A2*
[43]	Multinomial Logistic Regression	*A2*
[79]	K-Plane Algorithm	*A2, A4*
[89]	AdaRank	*A3*
[81]	Bagging Ensembles of CART, Bagging Ensembles of MLPs	*A3*
[83]	Data Mining Interpolation Technique	*A3*
[41]	Factor Analysis, k-means, Ordinary Least Squares	*A3*
[32, 91]	Genetic Programming (GP)	*A3*
[32]	Kriging	*A3*
[93]	Max-Apriori Classifier, Exhaustive Feature Subsets Classifiers, All Features Classifier, Incremental Feature Examination classifier	*A3*
[97]	Quick Optimization via Guessing	*A3*

Learning Stages

Does it work in practice? How to evaluate?

What are the challenging configurable systems?



- **Research Questions:**

RQ1. What are the **applications** addressed for learning techniques to explore configuration spaces?

RQ2. Which **sampling** methods are used by these techniques?

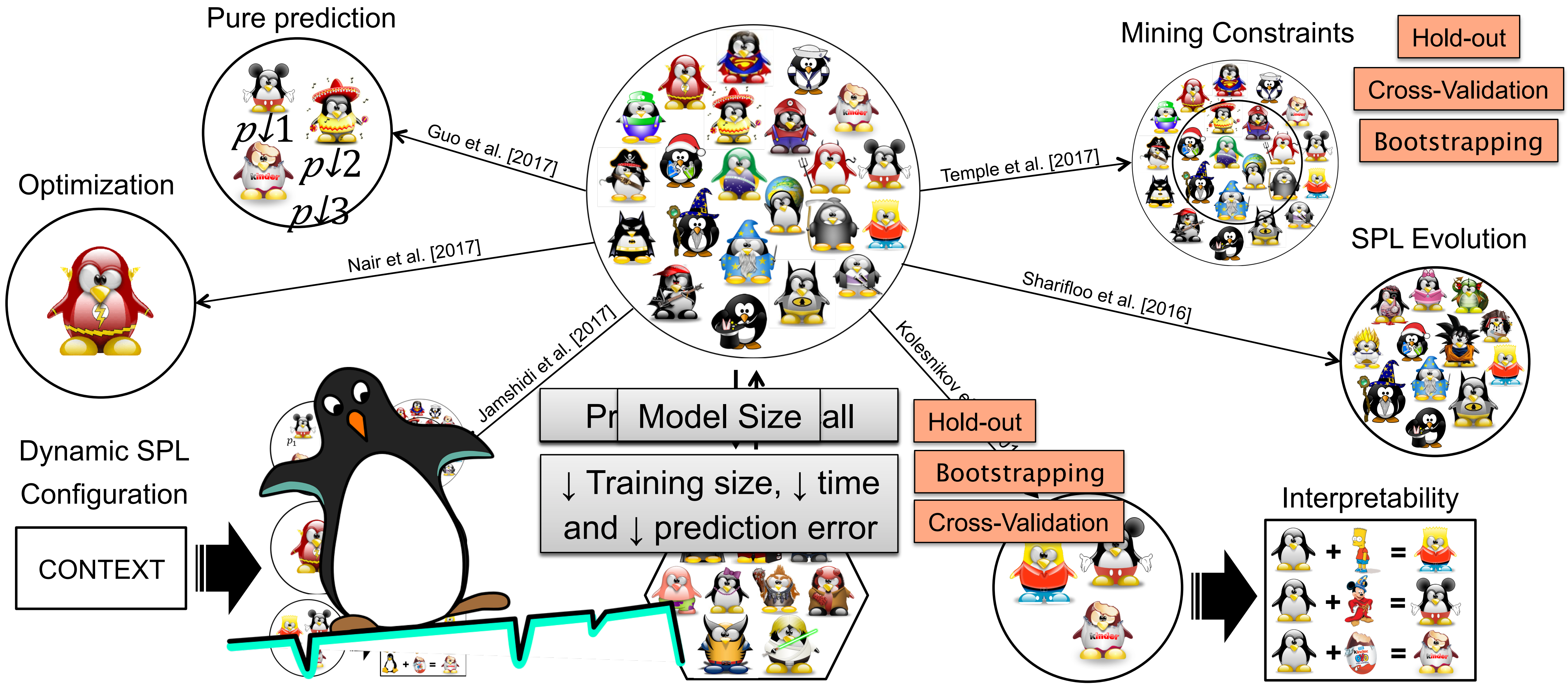
RQ3. How do the proposed techniques deal with **measurements** of non-functional properties?

RQ4. Which **learning** techniques have been proposed in the literature?

RQ5. How are these techniques **validated**?

RQ6. What are the **limitations** faced by the current techniques and open challenges that need attention in the future?

RQ5: Validation



RQ5: Validation

Reference	Evaluation Metric	Applicability
[67]	Closeness Range, Winner Probability	*A1*
[76, 77]	Coverage	*A1*
[70]	Jaccard Similarity	*A1*
[76]	Performance-Relevant Feature Interactions Detection Accuracy	*A1*
[69]	t-masked metric	*A1*
[75]	True Positive (TP) Rate, False Positive (FP) Rate, Receiver Operating Characteristic (ROC)	*A1*
[65, 46, 29, 47, 36, 37, 38, 39, 67, 45, 32, 68, 84, 71, 72, 59, 44, 35, 94, 79]	Mean Relative Error (MRE)	*A1, A2, A3, A4*
[86, 66, 56, 31, 84, 35, 79, 77]	Sampling Cost	*A1, A2, A3, A4*
[32, 68]	Highest Error (HE)	*A1, A3*
[42, 30, 81, 83, 44, 85, 74, 88, 73, 92, 90]	Mean Absolute Error (MAE)	*A1, A3, A4*
[72, 101, 96]	Mann-Whitney U-test	*A1, A3, A5*
[57, 72]	F-test	*A1, A4*
[60, 9, 100, 70, 64]	Precision, Recall	*A1, A4, A5*
[62, 69, 64, 101, 75]	Precision, Recall, F-measure	*A1, A5*
[80]	GAP	*A2*
[43]	Kullback-Leibler (KL), Pearson Linear Correlation, Spearman Correlation Coefficient	*A2*
[45, 47]	Structure of Prediction Models	*A2*

Reference	Evaluation Metric	Applicability
[43, 59]	Rank Correlation	*A2, A4*
[41, 35]	Domain Experts Feedback	*A3*
[97]	Error Probability	*A3*
[83]	Global Confidence, Neighbors Density Confidence, Neighbors Similarity Confidence	*A3*
[32]	LT15, LT30	*A3*
[54, 31]	Mean Rank Difference	*A3*
[81]	Median Magnitude of the Relative Error (MdMRE)	*A3*
[33]	Pareto Prediction Error	*A3*
[90, 91, 89]	Rank Accuracy (RA)	*A3*
[41, 93, 95, 97]	Tuning Time	*A3*
[81, 92, 57]	Mean Square Error (MSE)	*A3, A4*
[57]	p-value, R2, Residual Standard Error (RSE)	*A4*
[99]	Reward	*A4*
[58]	Statistical Significance	*A4*
[61, 58]	Qualitative Analysis	*A4, A5*
[55]	Ranking Constraints	*A4, A5*
[101]	Delaney's Statistics	*A5*
[101]	Distance Function, Hyper-volume (HV)	*A5*
[63]	Equivalence among Combinatorial Models	*A5*
[63]	Failure Index Delta (FID), Totally Repaired Models (TRM)	*A5*

Pure prediction: *Siegmund et al. Predicting Performance via Automated Feature-Interaction* Detection. ICSE, 2012.

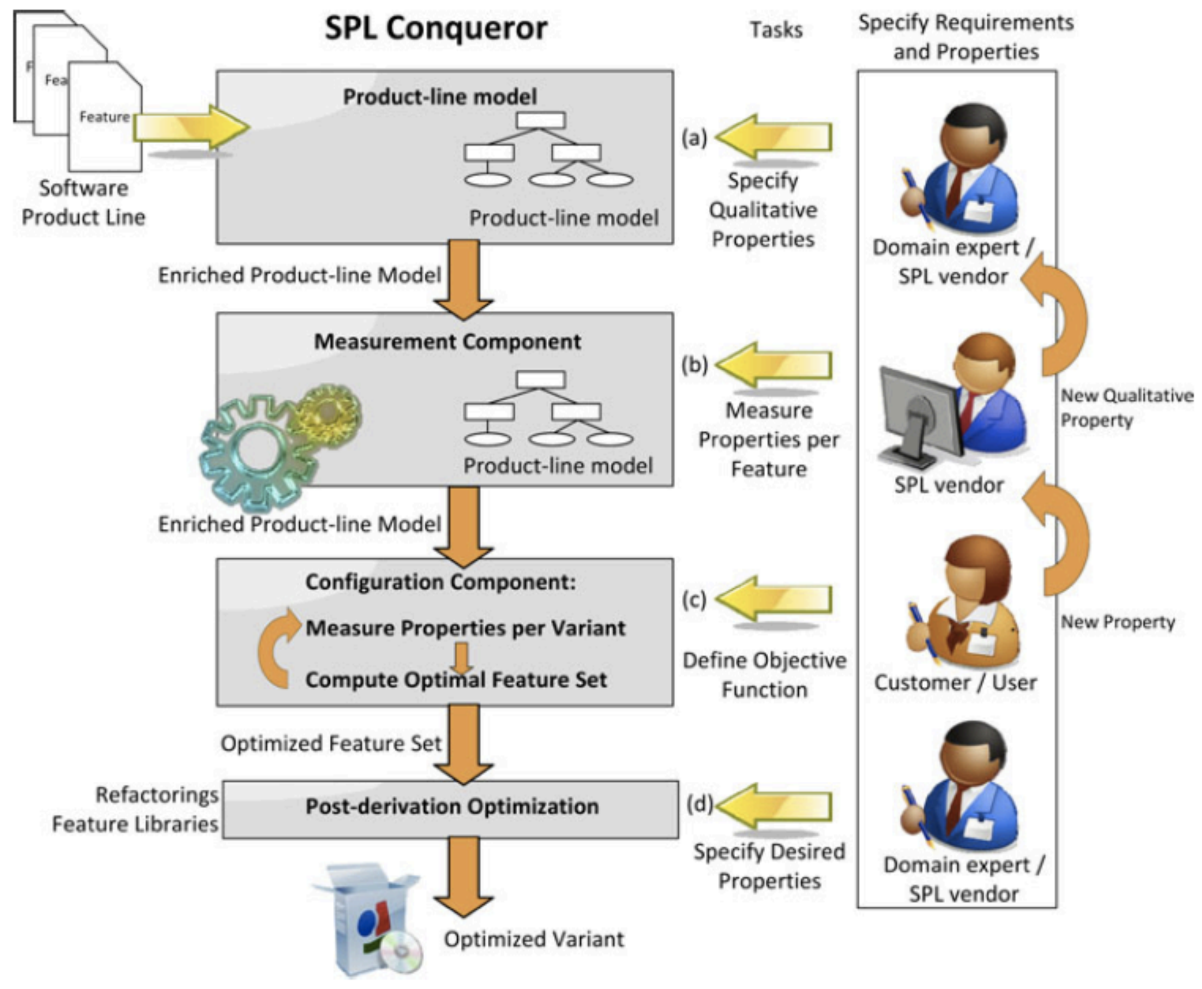


Fig. 5 Process of SPL Conqueror including the different tasks of measurement, configuration and optimization

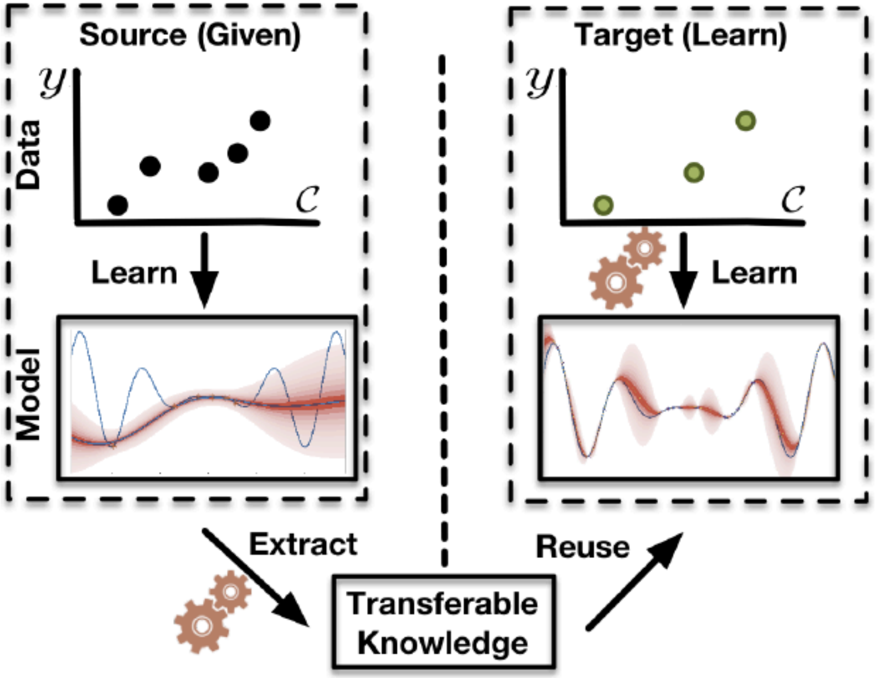
- Sampling: FW, PW, HO, HS, and BF
- Measuring: request rate and time
- Learning: Feature's Influence Delta
- Validation: Merge + Mean Relative Error

EVALUATION RESULTS FOR SIX CASE STUDIES; APPROACHES (APPR.): FEATURE-WISE (FW), PAIR-WISE (PW), HIGHER-ORDER (HO), HOT-SPOT (HS), BRUTE FORCE (BF). MEAN: MEAN FAULT RATE OF PREDICTIONS, STD: STANDARD DEVIATION OF PREDICTIONS.

		Effort			Fault Rate (in %)	
Program	Appr.	Measurements	Time (in h)	Interactions	Distribution	Mean±Std
Berkeley CE	FW	15 (0.6 %)	3	0		44.1±42.3
	PW	139 (5.4 %)	23	14		3.9±5.3
	HO	160 (6.3 %)	27	22		2.8±3.7
	HS	164 (6.4 %)	27	22		2.8±3.7
	BF	2 560 (100 %)	426	-	-40 -20 0 20 40	—
Berkeley JE	FW	10 (3 %)	8.4	0		17.7±19.6
	PW	48 (12 %)	40	24		8.5±9.6
	HO	116 (29 %)	97	51		3.8±5.7
	HS	162 (40.5 %)	137	69		1.7±3.5
	BF	400 (100 %)	335	-	-40 -20 0 20 40	—
Apache	FW	9 (4.7 %)	10	0		14.9±24.8
	PW	29 (15.1 %)	32	18		7.7±11.2
	HO	80 (41.7 %)	89	44		11.6±22.7
	HS	143 (74.5 %)	159	73		5.3±10.8
	BF	192 (100 %)	213	-	-40 -20 0 20 40	—

Interpretability: *Jamshidi et al. Transfer Learning for Performance Modeling of Configurable Systems: An Exploratory Analysis*. ASE, 2017.

- Sampling: Random
- Measuring: time
- Learning: Linear Regression , CART, and Multinomial Logistic Regression
- Validation: Kullback-Leibler, Pearson Linear Correlation, Spearman Correlation Coefficient, Rank correlation,



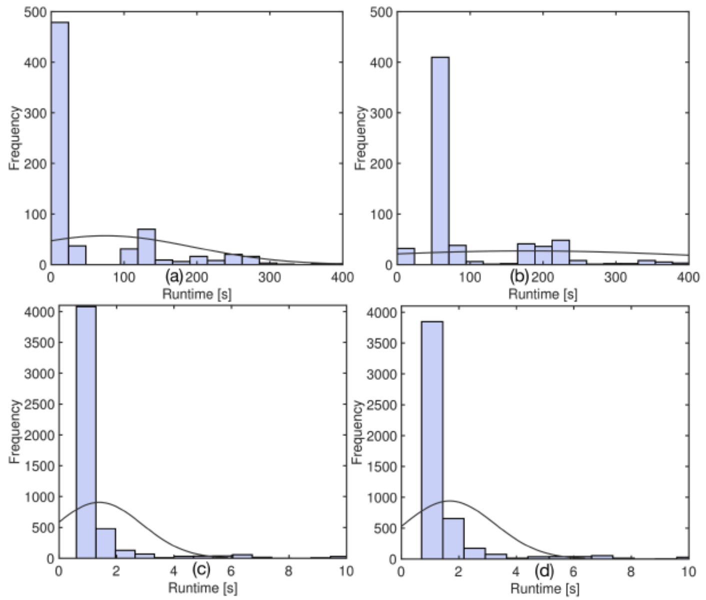


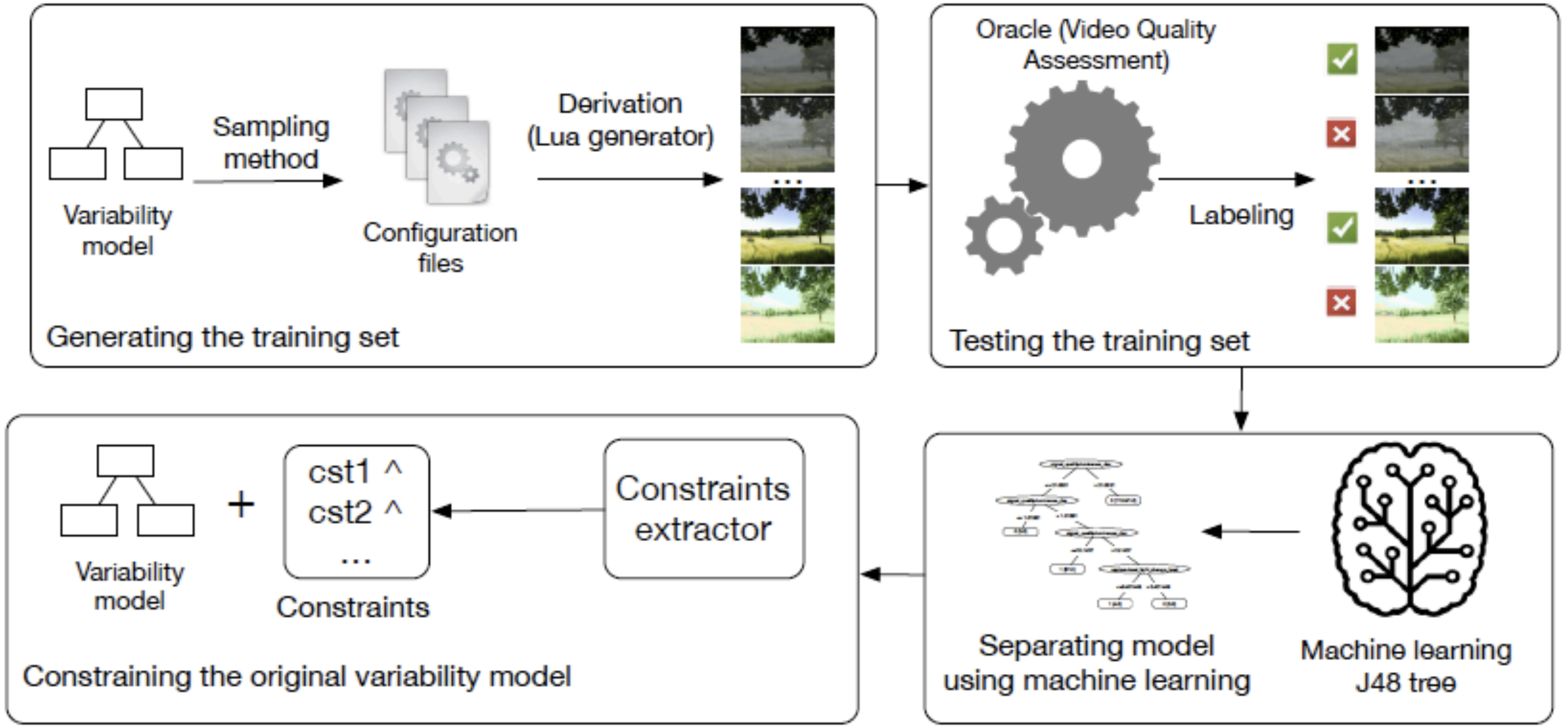
Fig. 2: Performance distributions of environments, depending on the severity of change, may be dissimilar, $D_{KL}^{ec_1} = 25.02$ (a,b), or very similar, $D_{KL}^{ec_{13}} = 0.32$ (c,d).

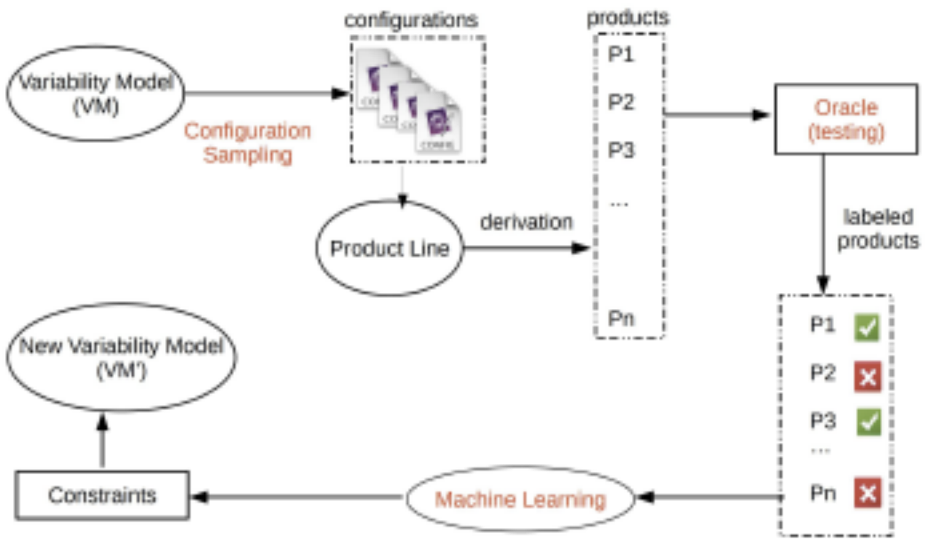
System	Domain	d	$\lvert\mathcal{C}\rvert$	$\lvert H\rvert$	$\lvert W\rvert$	$\lvert V\rvert$
SPEAR	SAT solver	14	16 384	3	4	2
x264	Video encoder	16	4 000	2	3	3
SQLite	Database	14	1 000	2	14	2
SaC	Compiler	50	71 267	1	10	1

d: configuration options; $\mathcal{C}$: configurations; H: hardware environments; W: analyzed workload; V: analyzed versions.

Mining Constraints: *Temple et al. Using Machine Learning to Infer Constraints for Product Lines*. SPLC, 2016.

- Sampling: Random
- Measuring: video quality
- Learning: CART
- Validation: HO + Precision and Recall





		Oracle	
		Faulty	Non-faulty
variability model (VM')	Faulty (invalid)	234 ($\pm$57.899)	69.5 ($\pm$26.973)
	Non-faulty (valid)	141.1 ($\pm$60.440)	3566.2 ($\pm$25.804)

Systematic Literature Review (SLR)