What to do with unpredicted failures ?

Slim Bouguerra, Ana Gainaru and Franck Cappello

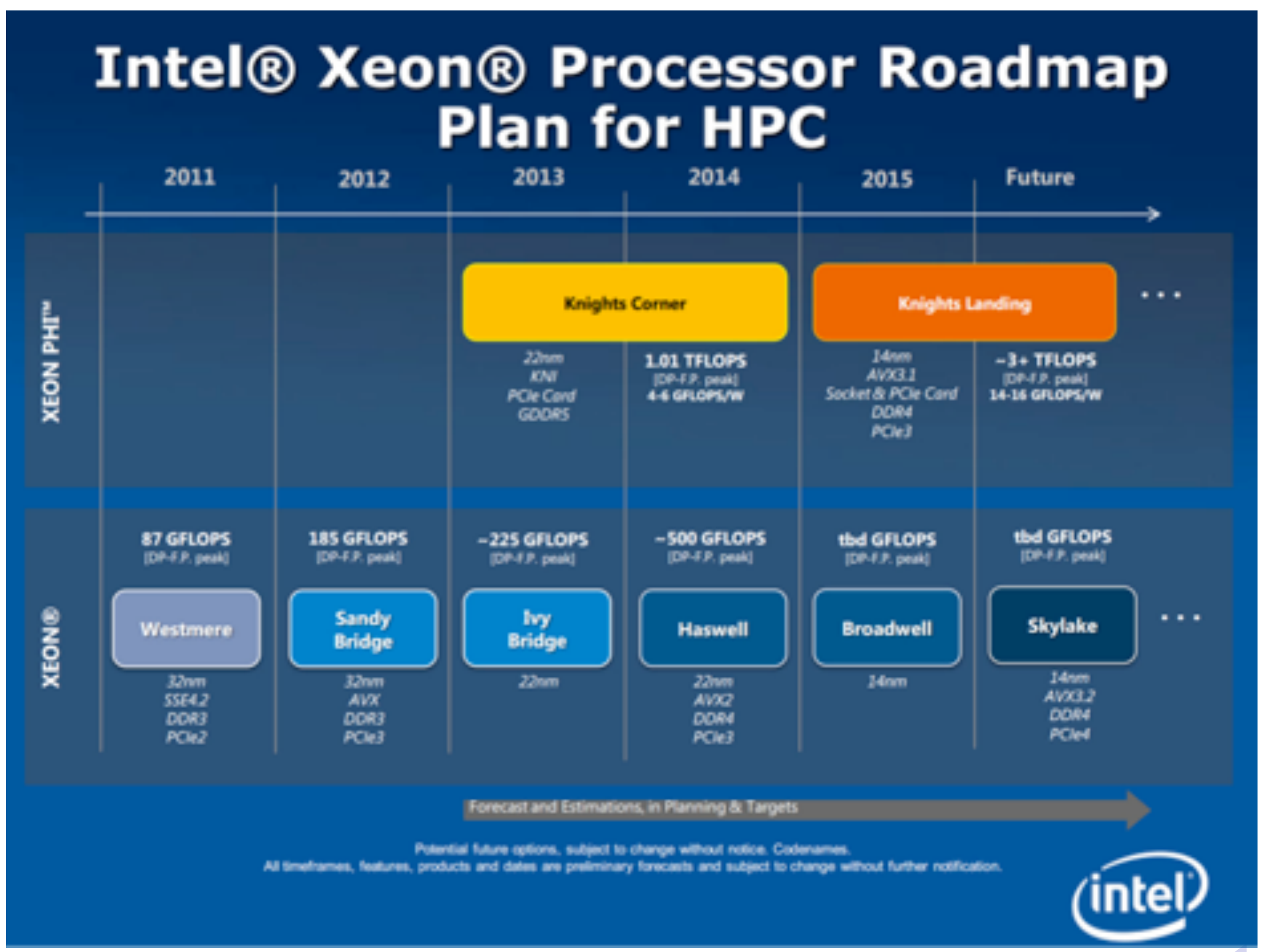
Joint Lab workshop November 2013-UIUC

Source code: scale_up.c

⋮

- Number_of_cores ++ ; // (several Millions)
- Die_shrinking++; // Next generation Xeon Phi on 14 nm.
- Assert(Power $<$ 20 Megawatts); // can not afford the bill

⋮



Source code: scale_up.c

⋮

- Number_of_cores ++ ; // (several Millions)
- Die_shrinking++; // Next generation Xeon Phi on 14 nm.
- Assert(Power < 20 Megawatts); // can not afford the bill

⋮

Failure Isn't An Option, It's a Certainty !!

Motivations

Main Motivation

Effective and efficient combination between proactive and preventive fault tolerance strategies.

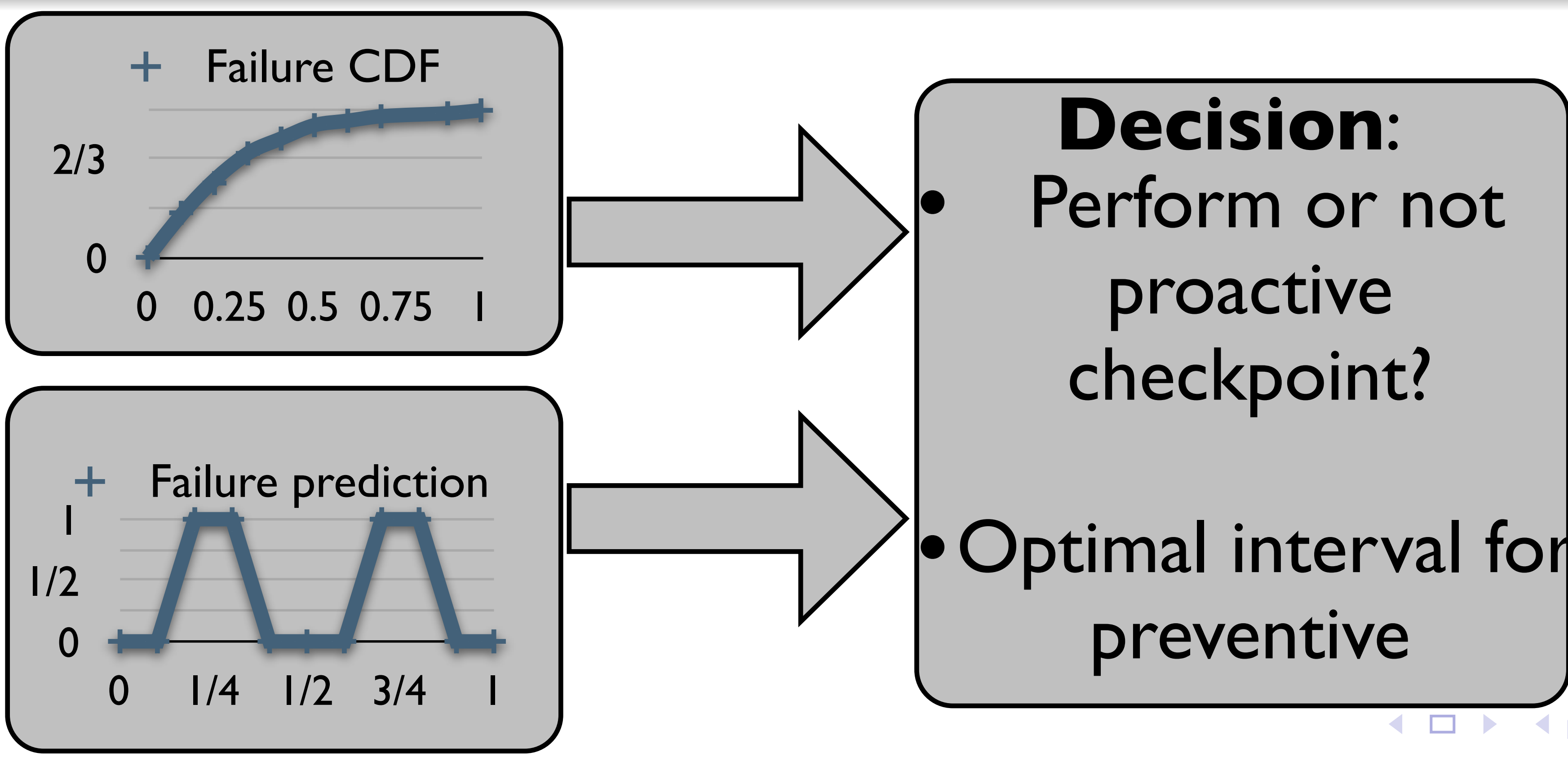
Motivations

Main Motivation

Effective and efficient combination between proactive and preventive fault tolerance strategies.

Target problem

Checkpoint interval selection problem.



Motivations

Main Motivation

Effective and efficient combination between proactive and preventive fault tolerance strategies.

Target problem

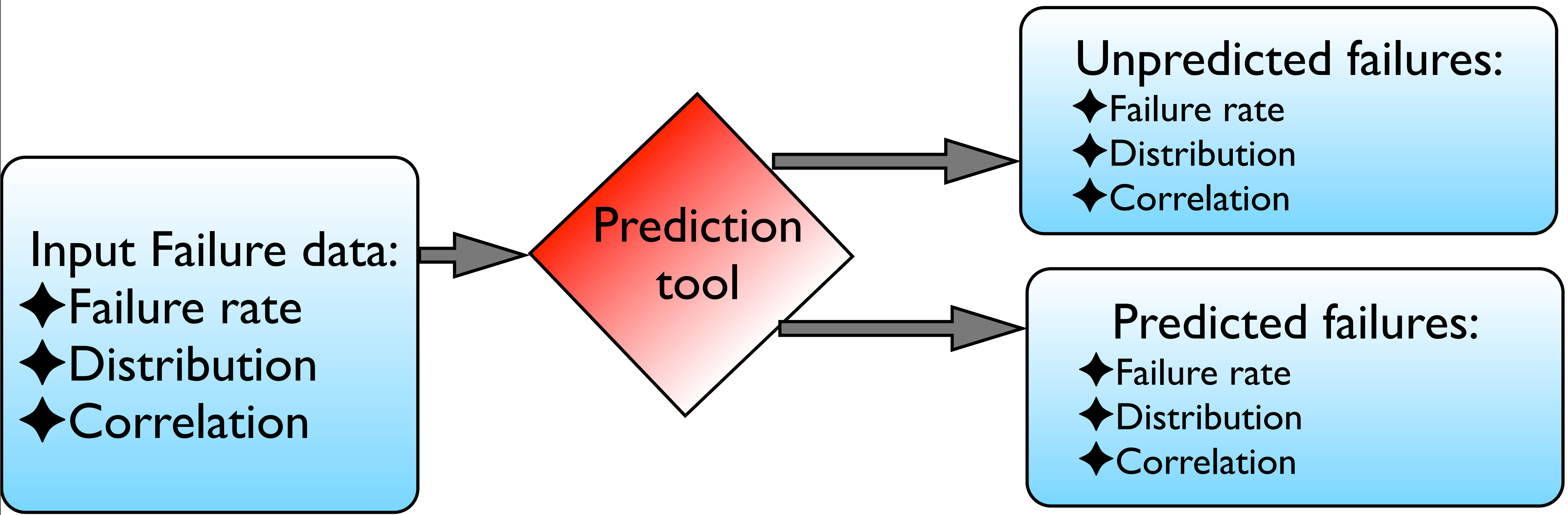
Checkpoint interval selection problem.

Objective

Advanced models to shape the relation between the occurrences of failures and the failure prediction mechanisms in HPC.

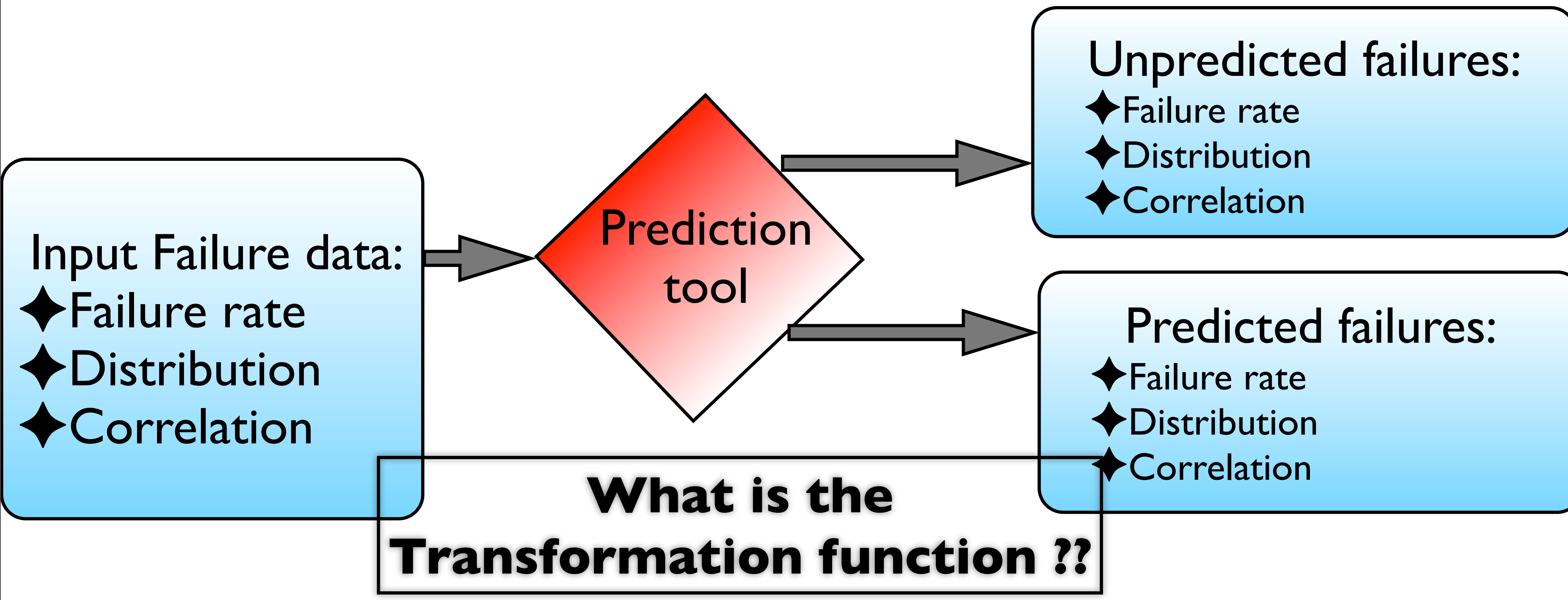
Problem description

1. Investigating the failure prediction transformations.



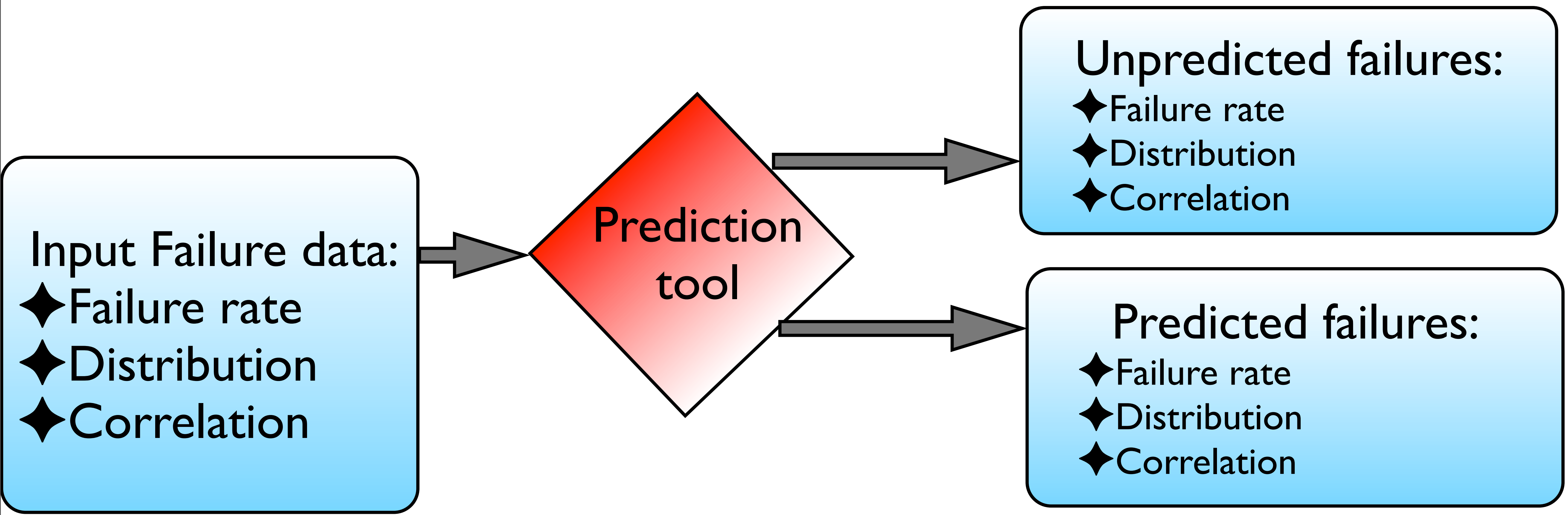
Problem description

1. Investigating the failure prediction transformations.



Problem description

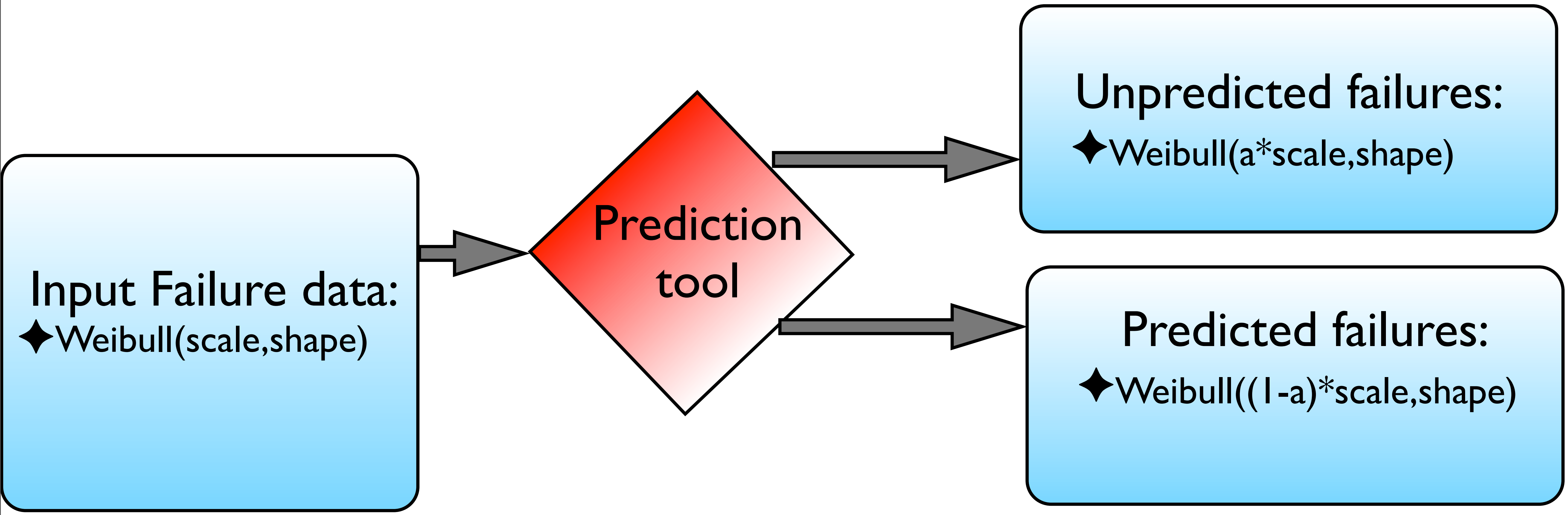
1. Investigating the failure prediction transformations.



2. How to deal with the unpredicted and predicted failures.

The Results

1. The failure prediction mechanism is scaling filter.



The Results

1. The failure prediction mechanism is scaling filter.
2. Correlation between failures isn't bad news and it helps to improve the recall.

The Results

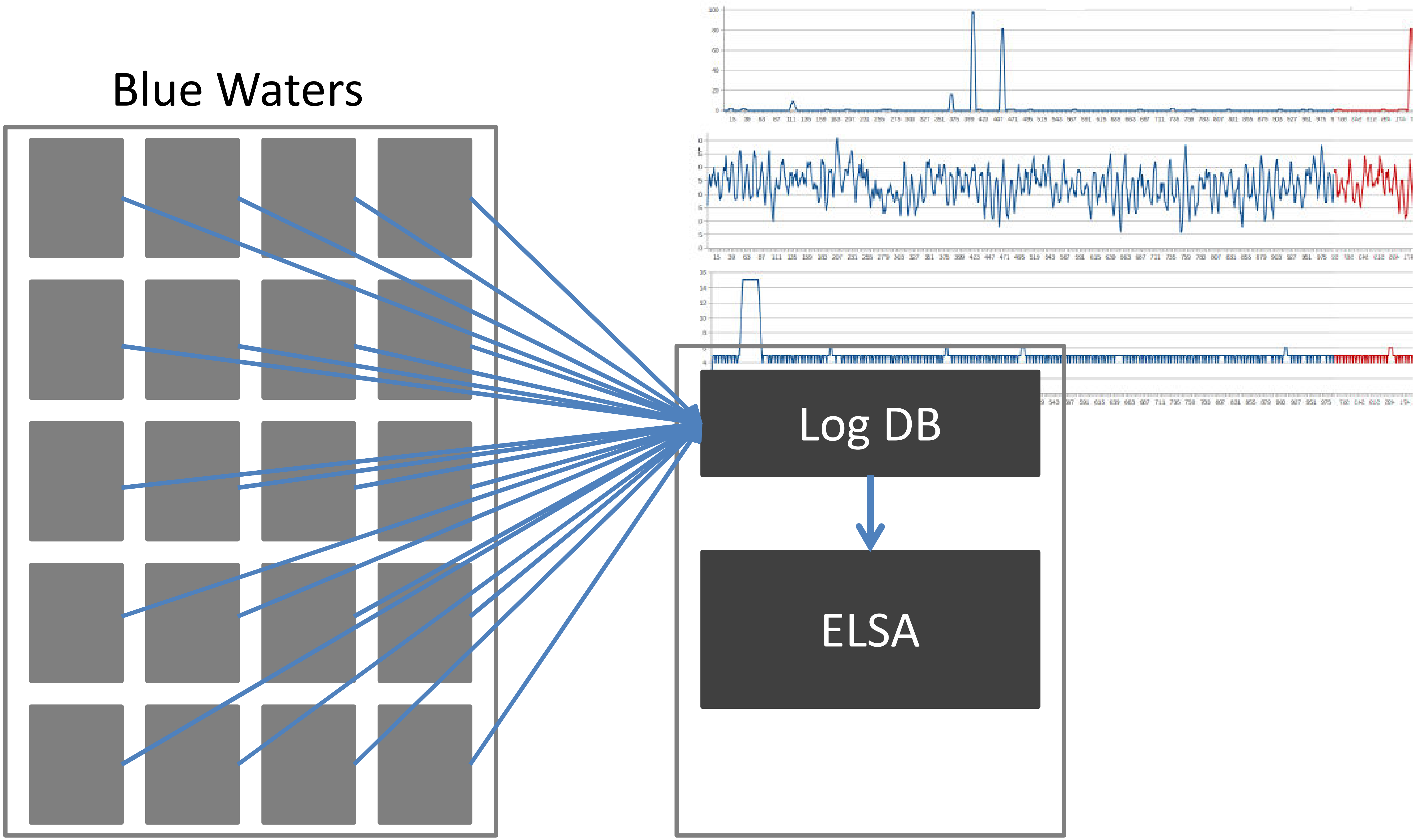
1. The failure prediction mechanism is scaling filter.
2. Correlation between failures isn't bad news and it helps to improve the recall.
3. The failure prediction mechanism catches the the noise (correlations) in data (Easier to infer mathematical models).

The Results

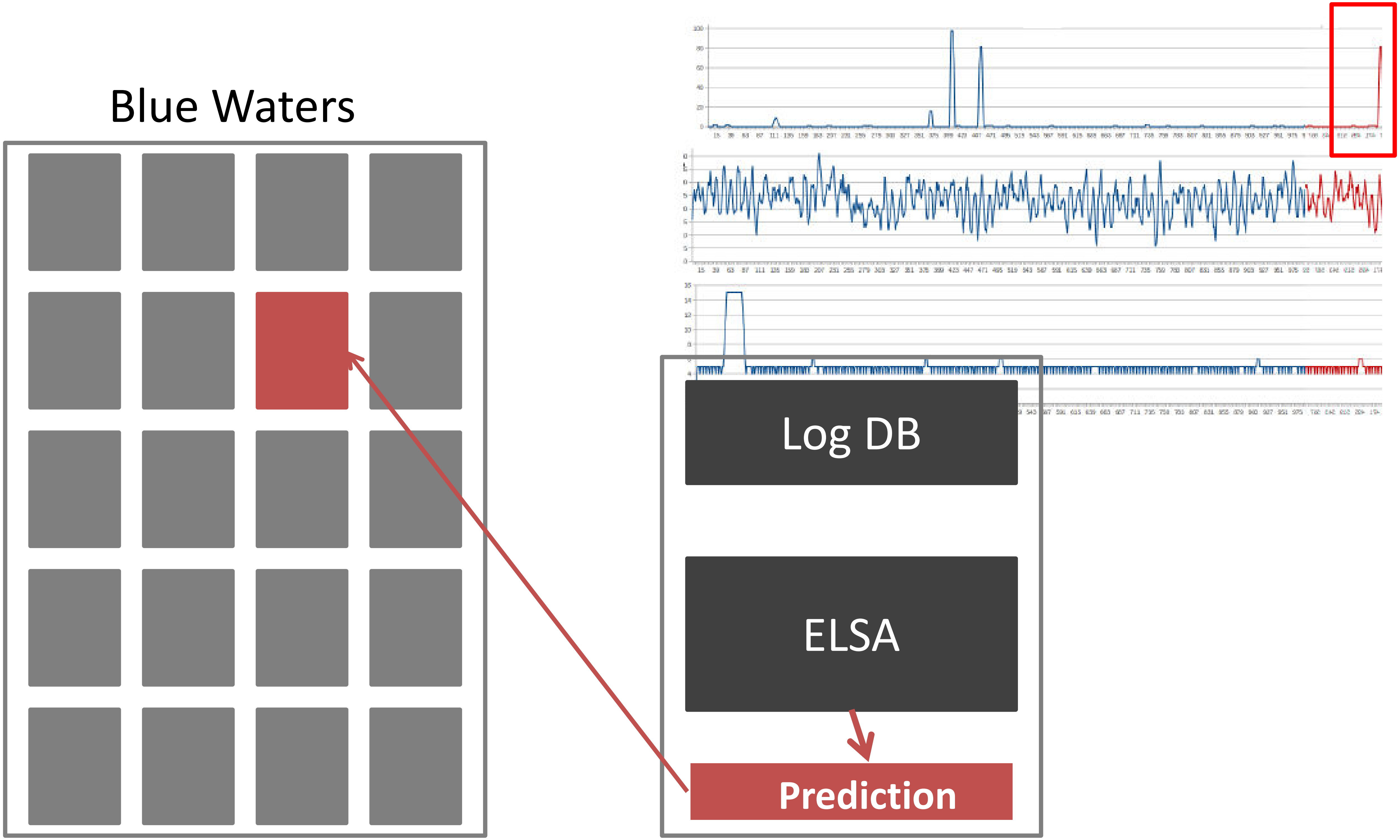
1. The failure prediction mechanism is scaling filter.
2. Correlation between failures isn't bad news and it helps to improve the recall.
3. The failure prediction mechanism catches the the noise (correlations) in data (Easier to infer mathematical models).
4. Combing proactive and preventive checkpointing leads to an improvement of 12 % to 30% of the amount of useful work.

Outline

Let's remember ELSA



Let’s remember ELSA



Online failure prediction terminology

Terminology

- True positive alert (correct prediction)
- False positive alert (misleading prediction)
- False negative alert (unpredicted failure)

Online failure prediction terminology

Terminology

- True positive alert (correct prediction)
- False positive alert (misleading prediction)
- False negative alert (unpredicted failure)

Metric

- Recall:

$$\frac{\#\text{True positive}}{\#\text{True positive} + \#\text{False negative}}$$

- Precision:

$$\frac{\#\text{True positive}}{\#\text{True positive} + \#\text{False positive}}$$

Proactive and preventive fault tolerance

Prediction is feasible

- ELSA: Signal analysis with data mining:
  - 90% precision and 45% recall.
  - At least 10 seconds of lead-time.
  - Failure location is provided.

Proactive and preventive fault tolerance

Prediction is feasible

- ELSA: Signal analysis with data mining:
  - 90% precision and 45% recall.
  - At least 10 seconds of lead-time.
  - Failure location is provided.

Fast checkpointing strategies exist

- FTI (Fault Tolerance Interface):
  - Capable of taking a checkpoint in 5s for 1GB memory.
  - Multi-level checkpoint with 8% overhead.

Outline

Data characteristics

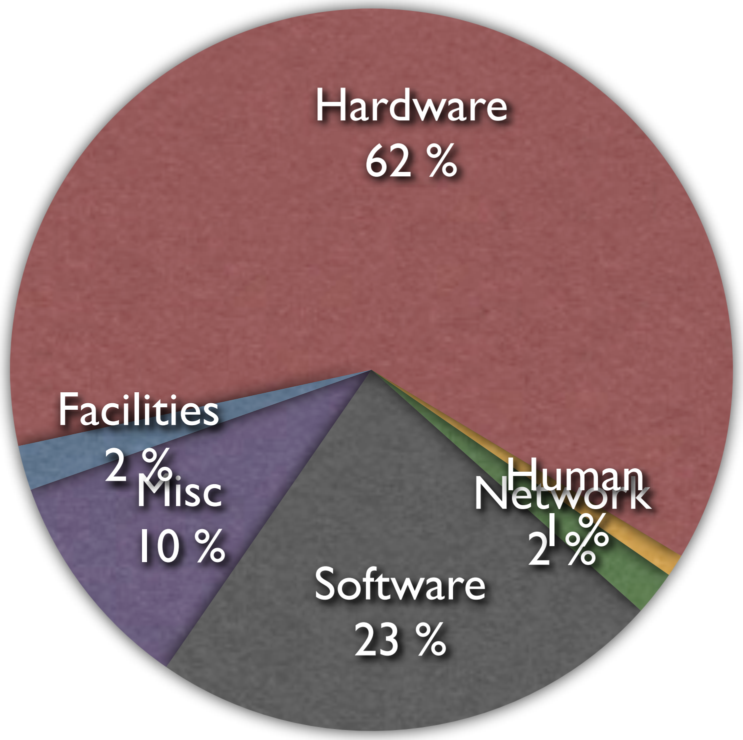
- 22 High performance computing systems from Los Alamos National Lab.
    - December 1996 - November 2005.
    - Different architectures and sizes.
    - 433,490 per system.
    - MTBF, 13 to 215 hours.
    - Failures are manually annotated.

Data characteristics

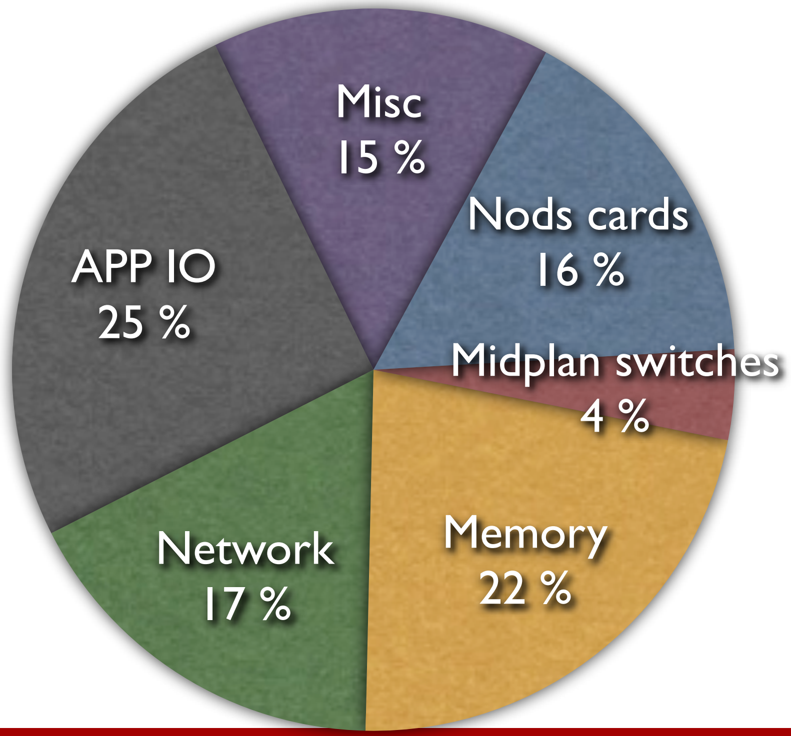
- 22 High performance computing systems from Los Alamos National Lab.
  - December 1996 - November 2005.
  - Different architectures and sizes.
  - 433,490 per system.
  - MTBF, 13 to 215 hours.
  - Failures are manually annotated.
- BlueGene/L at Lawrence Livermore National Lab.
  - June 2005 - january 2006.
  - 128K PowerPc 440 processors.
  - 4,747,963 events.
  - MTBF 24h.
  - Anomaly detection technique.

Failure prediction characteristics

22 HPC systems

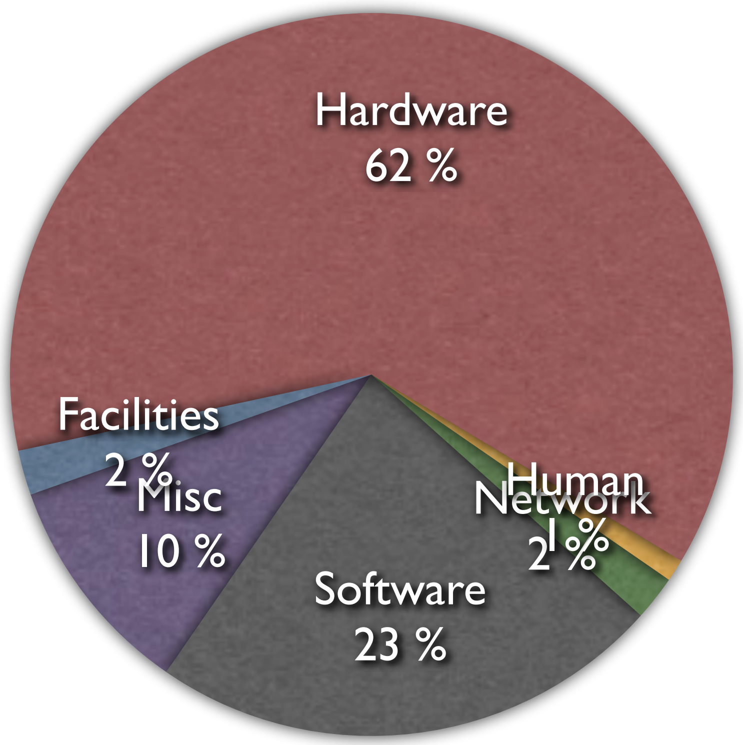


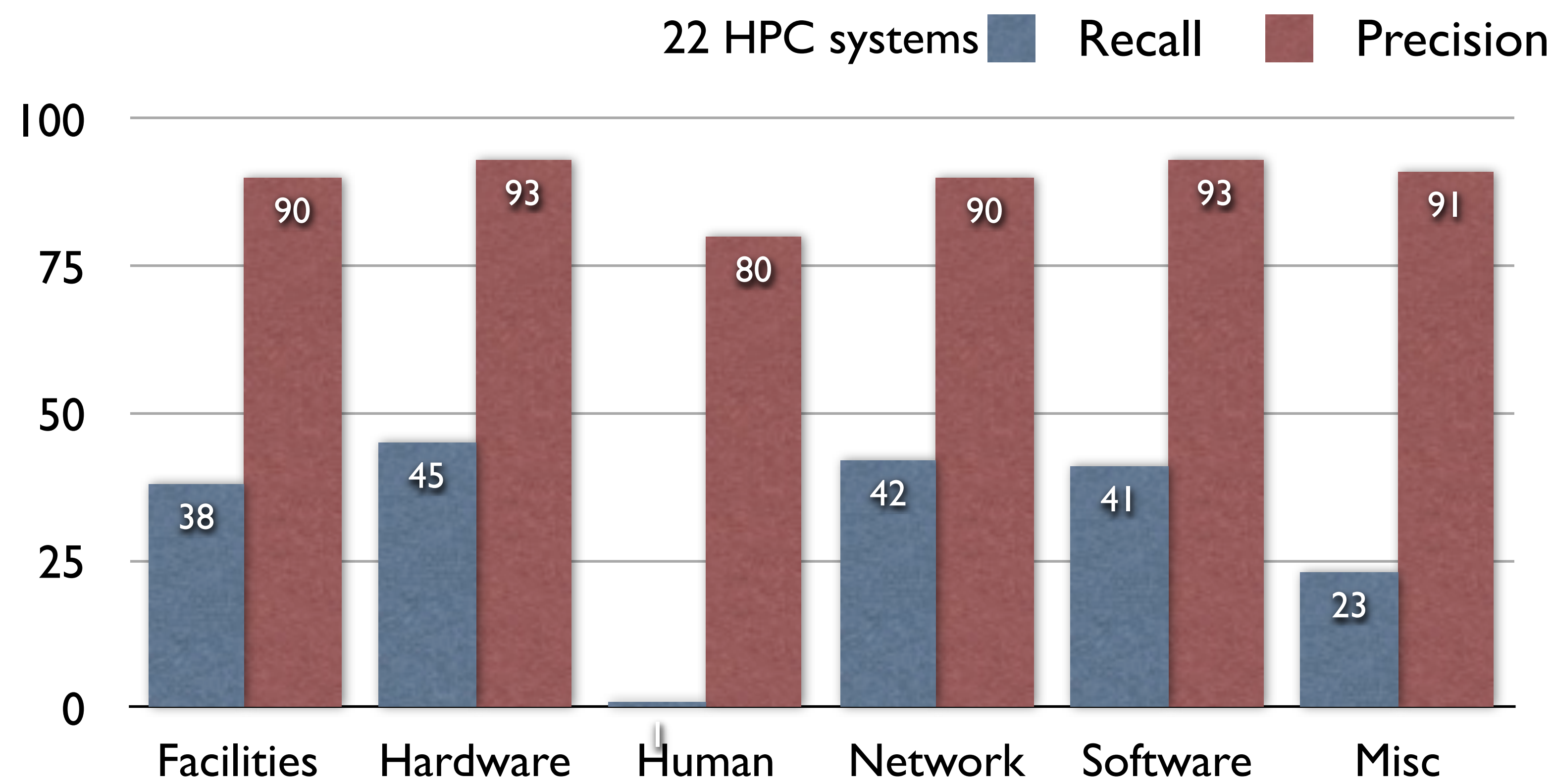
BG/L



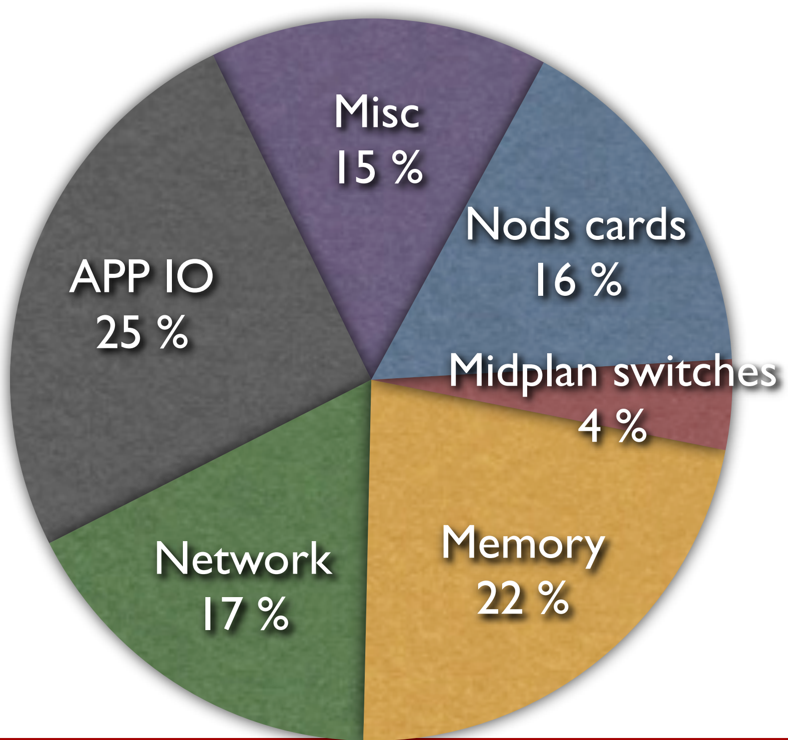
Failure prediction characteristics

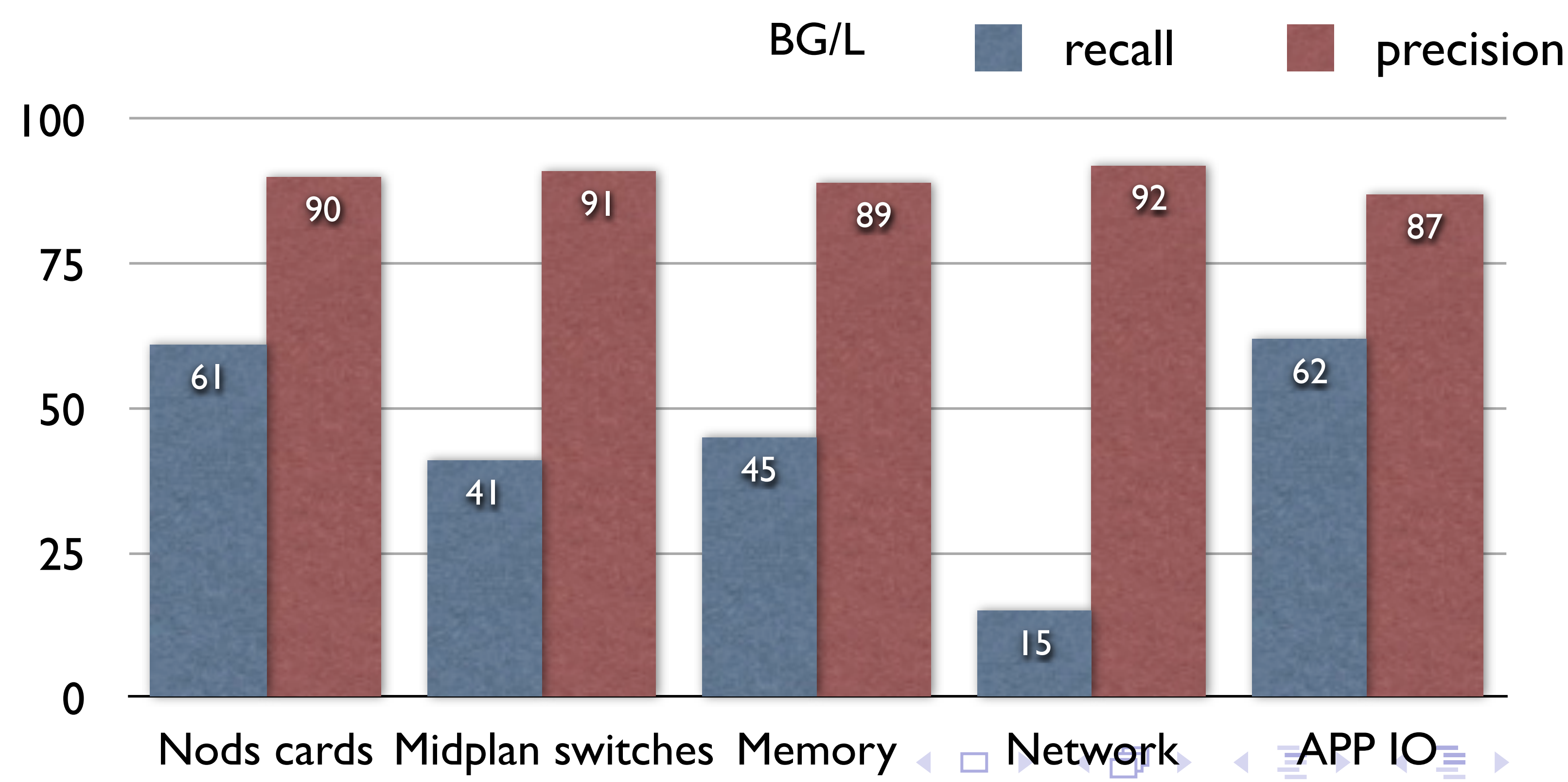
22 HPC systems





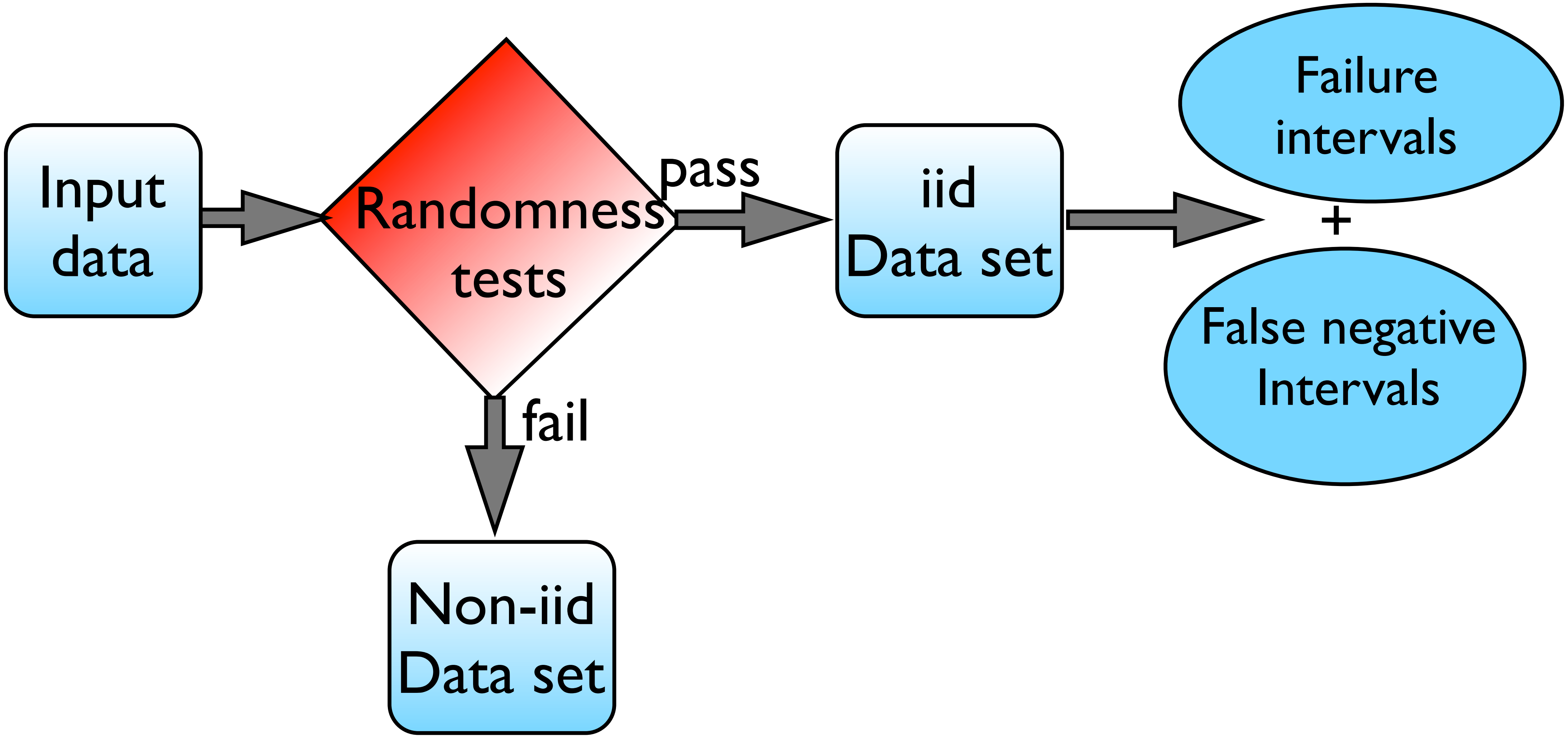
BG/L





Outline

Methodology: Randomness Test



Method:

- Runs test
- Runs up/down test
- Autocorrelation function test (ACF)

Randomness tests output

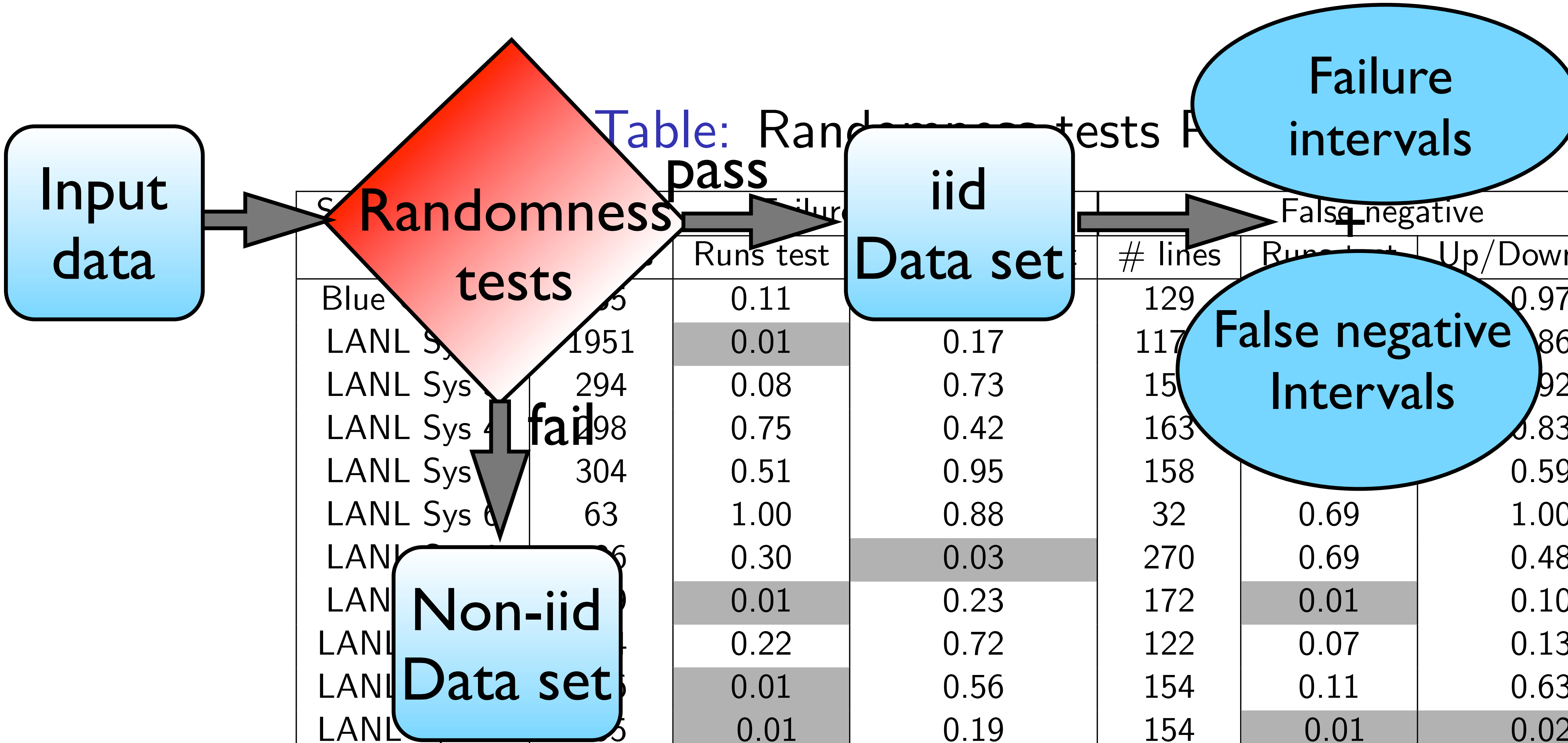


Table: Rand[illegible]tests F[illegible]

[illegible]	[illegible]	Failur[illegible] Runs test	[illegible]	False negative # lines	Runs test	Up/Down test
Blue [illegible]	[illegible]5	0.11	[illegible]	129	[illegible]	0.97
LANL Sy[illegible]	1951	0.01	0.17	117	[illegible]	[illegible]86
LANL Sys [illegible]	294	0.08	0.73	15[illegible]	[illegible]	[illegible]92
LANL Sys 4	[illegible]98	0.75	0.42	163	[illegible]	[illegible].83
LANL Sys [illegible]	304	0.51	0.95	158	[illegible]	0.59
LANL Sys 6	63	1.00	0.88	32	0.69	1.00
LAN[illegible]	[illegible]6	0.30	0.03	270	0.69	0.48
LAN[illegible]	[illegible]	0.01	0.23	172	0.01	0.10
LANL [illegible]	[illegible]	0.22	0.72	122	0.07	0.13
LANL [illegible]	[illegible]	0.01	0.56	154	0.11	0.63
LANL [illegible]	[illegible]5	0.01	0.19	154	0.01	0.02
LANL Sys 13	194	0.04	0.74	123	0.80	0.53
LANL Sys 14	120	0.06	0.36	75	0.49	0.17
LANL Sys 15	53	0.01	0.87	32	0.50	0.51
LANL Sys 16	245	0.04	0.98	159	0.62	0.97
LANL Sys 18	3917	0.01	0.01	2195	0.66	0.74
LANL Sys 19	3235	0.03	0.54	1785	0.08	0.86
LANL Sys 20	2400	0.01	0.14	1310	0.01	0.85
LANL Sys 21	105	0.02	0.01	76	0.39	0.96
LANL Sys 22	17	not	enough	lines		
LANL Sys 23	226	0.32	0.41	129	0.15	0.55
LANL Sys 24	23	not	enough	lines		

Randomness tests output

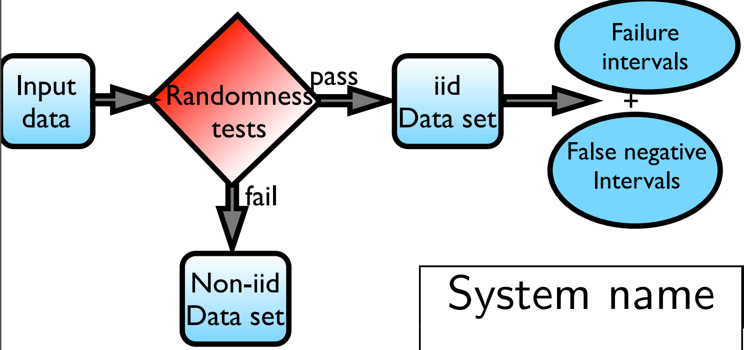
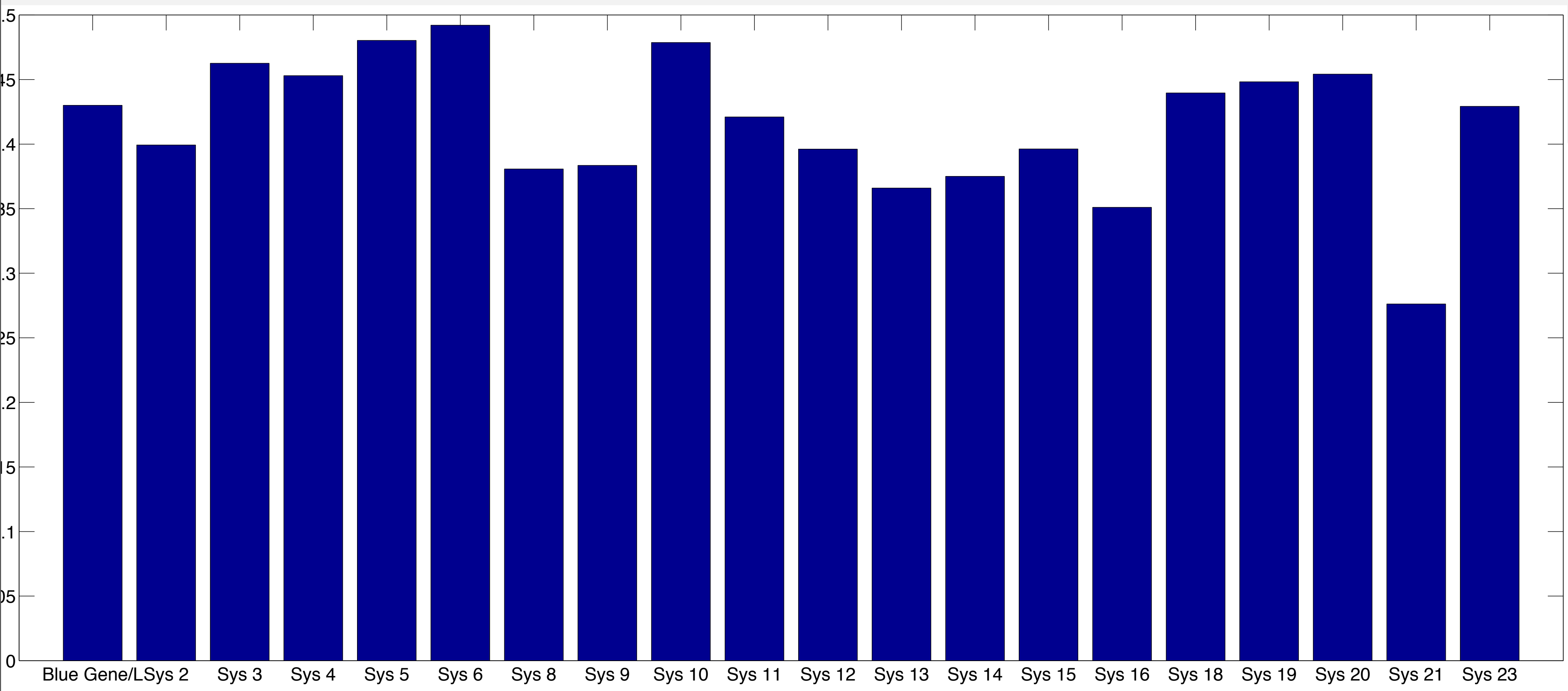
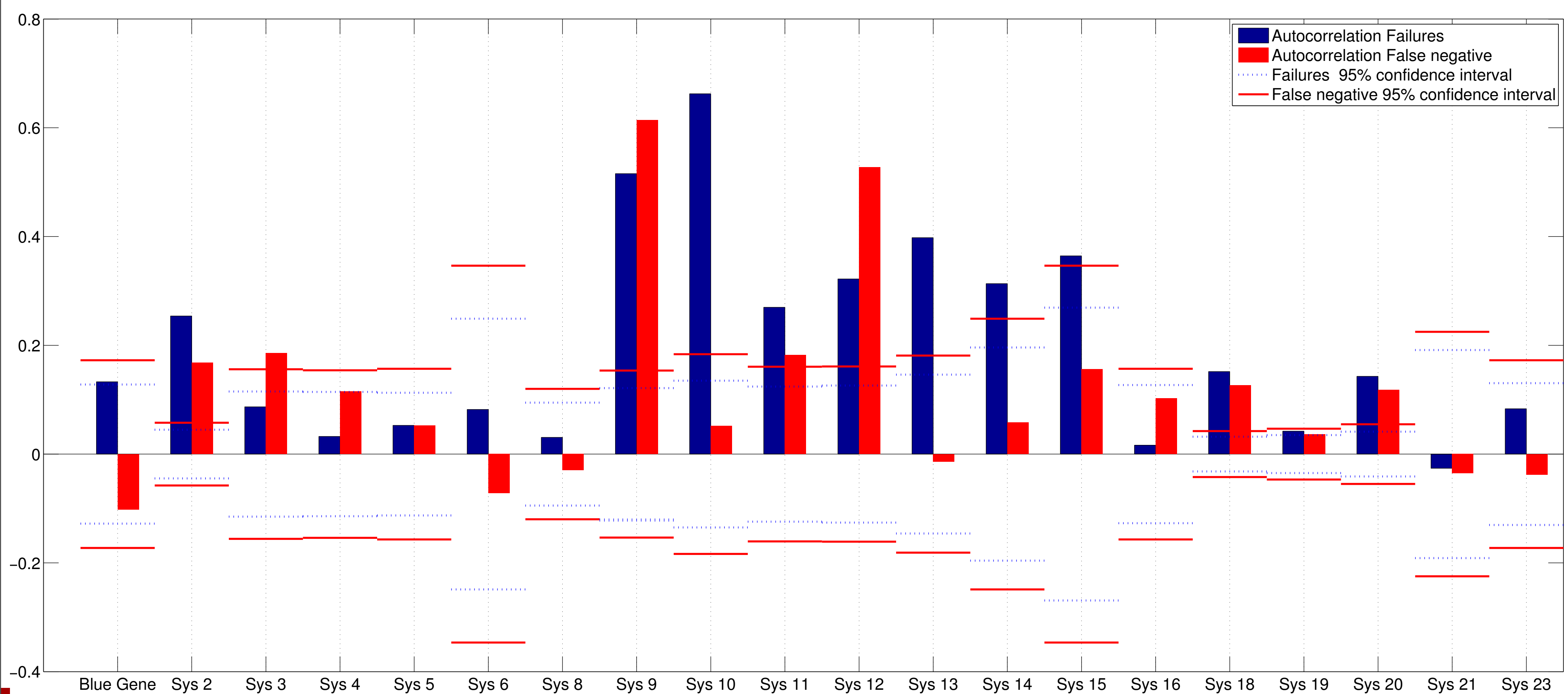
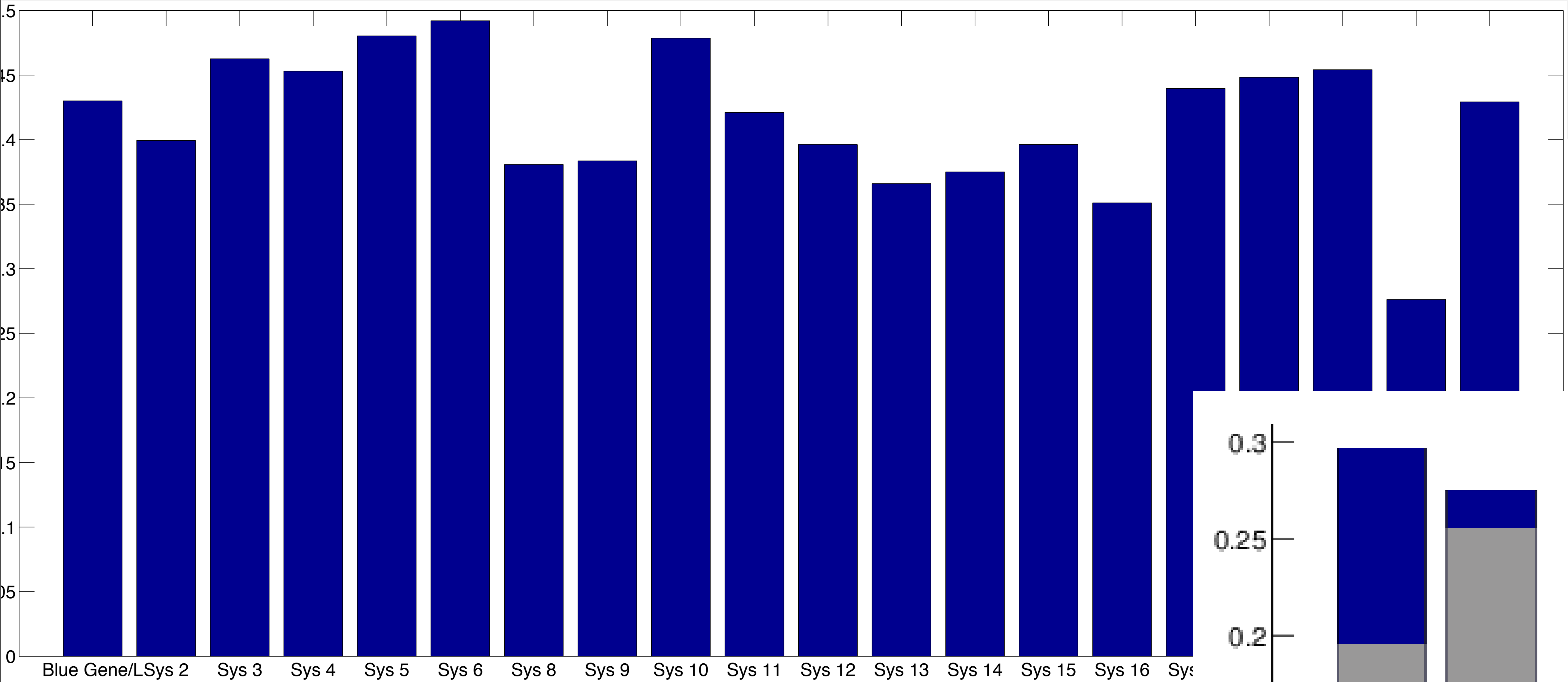


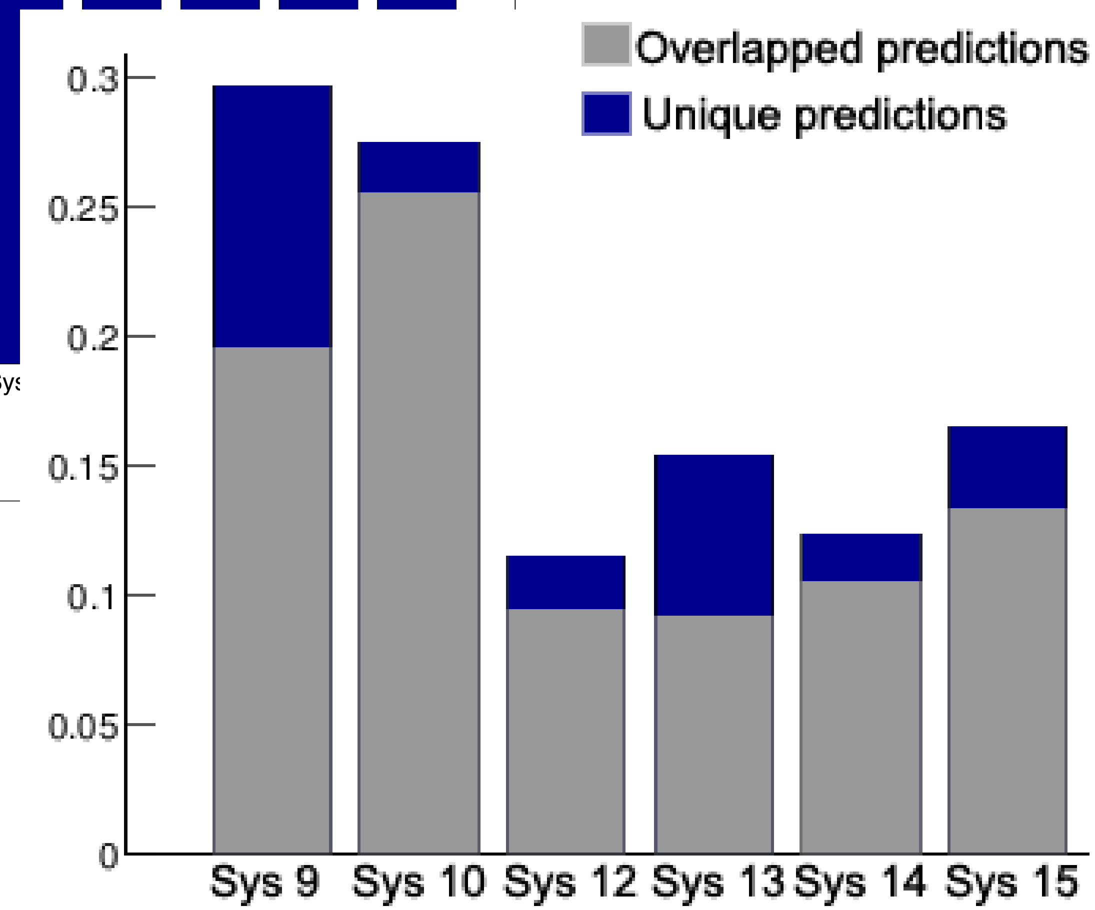
Table: Randomness tests P-values

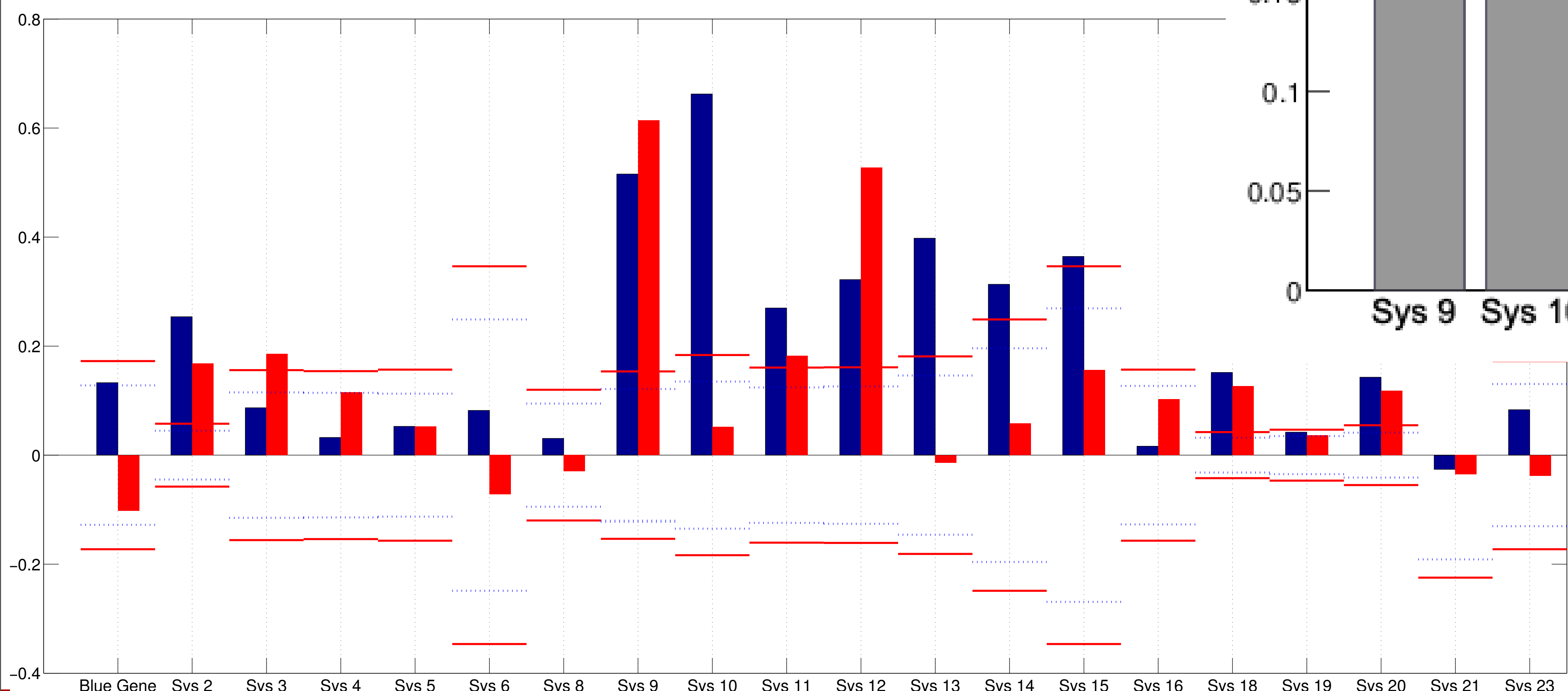
System name	Failures			False negative		
	# lines	Runs test	Up/Down test	# lines	Runs test	Up/Down test
Blue Gene/L	235	0.11	0.17	129	0.70	0.97
LANL Sys 2	1951	0.01	0.17	1172	0.01	0.86
LANL Sys 3	294	0.08	0.73	158	0.36	0.92
LANL Sys 4	298	0.75	0.42	163	0.15	0.83
LANL Sys 5	304	0.51	0.95	158	0.83	0.59
LANL Sys 6	63	1.00	0.88	32	0.69	1.00
LANL Sys 8	436	0.30	0.03	270	0.69	0.48
LANL Sys 9	279	0.01	0.23	172	0.01	0.10
LANL Sys 10	234	0.22	0.72	122	0.07	0.13
LANL Sys 11	266	0.01	0.56	154	0.11	0.63
LANL Sys 12	255	0.01	0.19	154	0.01	0.02
LANL Sys 13	194	0.04	0.74	123	0.80	0.53
LANL Sys 14	120	0.06	0.36	75	0.49	0.17
LANL Sys 15	53	0.01	0.87	32	0.50	0.51
LANL Sys 16	245	0.04	0.98	159	0.62	0.97
LANL Sys 18	3917	0.01	0.01	2195	0.66	0.74
LANL Sys 19	3235	0.03	0.54	1785	0.08	0.86
LANL Sys 20	2400	0.01	0.14	1310	0.01	0.85
LANL Sys 21	105	0.02	0.01	76	0.39	0.96
LANL Sys 22	17	not	enough	lines		
LANL Sys 23	226	0.32	0.41	129	0.15	0.55
LANL Sys 24	23	not	enough	lines		


0
Blue Gene/LSys 2
Sys 3
Sys 4
Sys 5
Sys 6
Sys 8
Sys 9
Sys 10
Sys 11
Sys 12
Sys 13
Sys 14
Sys 15
Sys 16
Sys 18
Sys 19
Sys 20
Sys 21
Sys 23

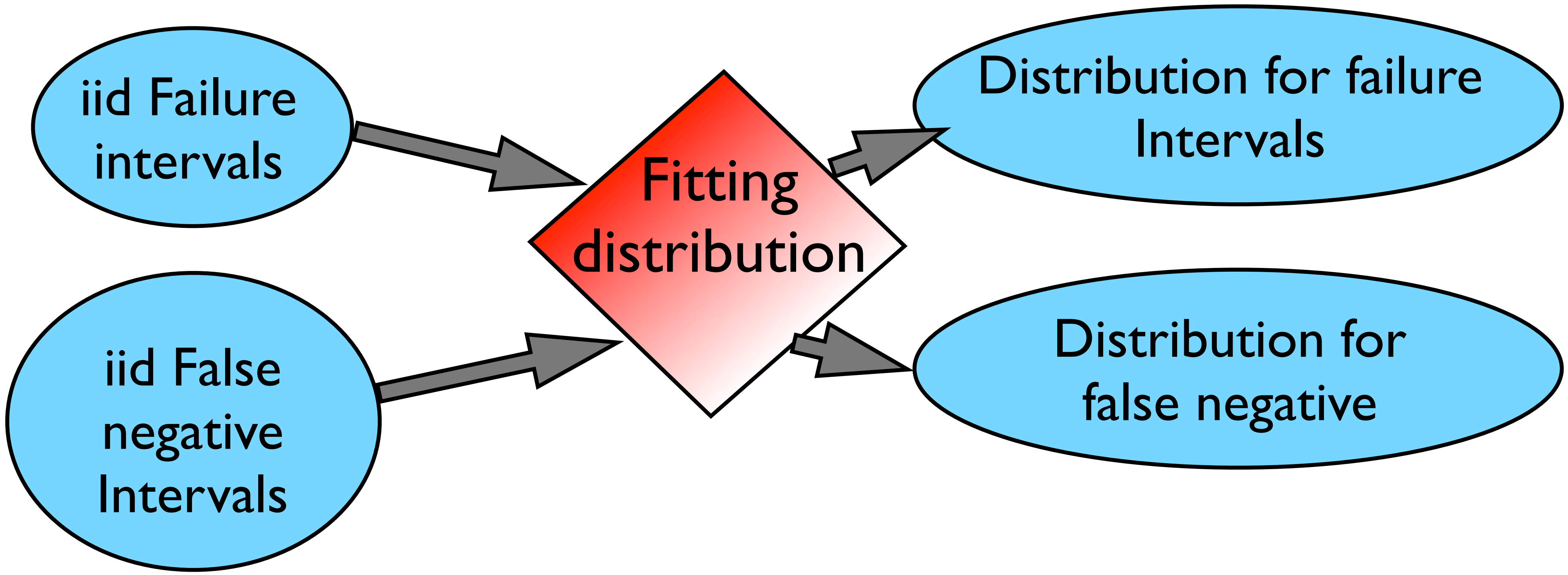

Autocorrelation Failures
Autocorrelation False negative
Failures 95% confidence interval
False negative 95% confidence interval
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Blue Gene
Sys 2
Sys 3
Sys 4
Sys 5
Sys 6
Sys 8
Sys 9
Sys 10
Sys 11
Sys 12
Sys 13
Sys 14
Sys 15
Sys 16
Sys 18
Sys 19
Sys 20
Sys 21
Sys 23


0
Blue Gene/L
Sys 2
Sys 3
Sys 4
Sys 5
Sys 6
Sys 8
Sys 9
Sys 10
Sys 11
Sys 12
Sys 13
Sys 14
Sys 15
Sys 16


Overlapped predictions
Unique predictions
0.3
0.25
0.2
0.15
0.1
0.05
0
Sys 9
Sys 10
Sys 12
Sys 13
Sys 14
Sys 15


0.8
0.6
0.4
0.2
0
-0.2
-0.4
Blue Gene
Sys 2
Sys 3
Sys 4
Sys 5
Sys 6
Sys 8
Sys 9
Sys 10
Sys 11
Sys 12
Sys 13
Sys 14
Sys 15
Sys 16
Sys 18
Sys 19
Sys 20
Sys 21
Sys 23

Methodology: Fitting



Method:

- Maximum Likelihood Estimation (MLE)

Target Distributions: Exponential, Weibull, Log-normal and Gamma.

Fitting output

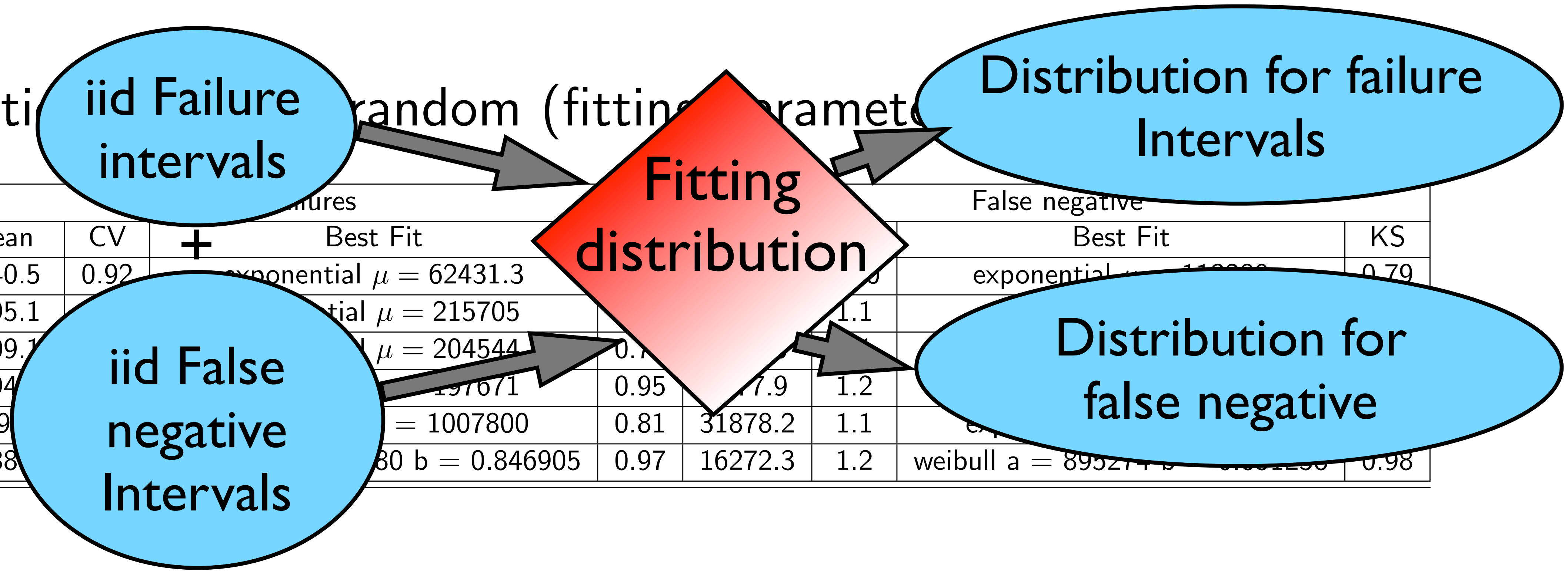


Table: Statisti[illegible] random (fitting parameters[illegible]

System name	Failures				False negative			
	Mean	CV	Best Fit	KS	Mean	CV	Best Fit	KS
Blue Gene/L	1040.5	0.92	exponential $\mu = 62431.3$	[illegible]	[illegible]	[illegible]	exponential [illegible]	0.79
LANL Sys 3	3595.1	[illegible]	[illegible]tial $\mu = 215705$	[illegible]	[illegible]	1.1	[illegible]	[illegible]
LANL Sys 4	3409.[illegible]	[illegible]	[illegible] $\mu = 204544$	0.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
LANL Sys 5	3294[illegible]	[illegible]	[illegible]197671	0.95	[illegible]7.9	1.2	[illegible]	[illegible]
LANL Sys 6	1679[illegible]	[illegible]	[illegible] = 1007800	0.81	31878.2	1.1	[illegible]	[illegible]
LANL Sys 23	9288[illegible]	[illegible]	[illegible]80 b = 0.846905	0.97	16272.3	1.2	weibull a = 8952[illegible]	0.98

Table: Statistical Fitting false negative random

System name	False negative			
	Mean	CV	Best Fit	KS
LANL Sys 8	7859.6	1.4	weibull a = 401499 b = 0.767798	0.74
LANL Sys 10	8247.0	3.6	weibull a = 318087 b = 0.647838	0.29
LANL Sys 11	6353.5	3.0	weibull a = 232647 b = 0.609348	0.61
LANL Sys 13	8164.3	3.9	lognormal $\mu = 11.5257$ $\sigma = 1.87004$	0.14
LANL Sys 14	11351.0	2.5	weibull a = 391931 b = 0.559039	0.77
LANL Sys 15	12136.7	1.2	exponential $\mu = 728203$	0.17
LANL Sys 16	3430.6	1.3	weibull a = 182624 b = 0.810939	0.69
LANL Sys 18	818.6	1.5	lognormal $\mu = 10.1123$ $\sigma = 1.28677$	0.37
LANL Sys 19	863.6	1.4	exponential $\mu = 29000.5$	0.18
LANL Sys 21	1986.9	2.3	lognormal $\mu = 10.6382$ $\sigma = 1.46402$	0.85

Fitting output

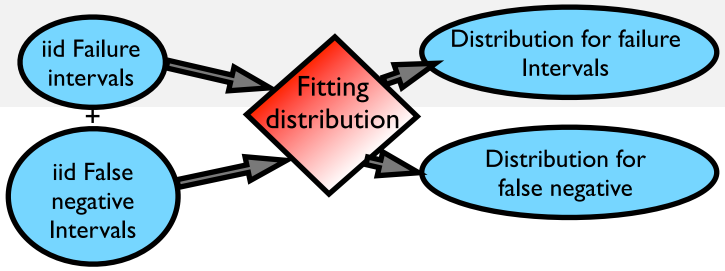


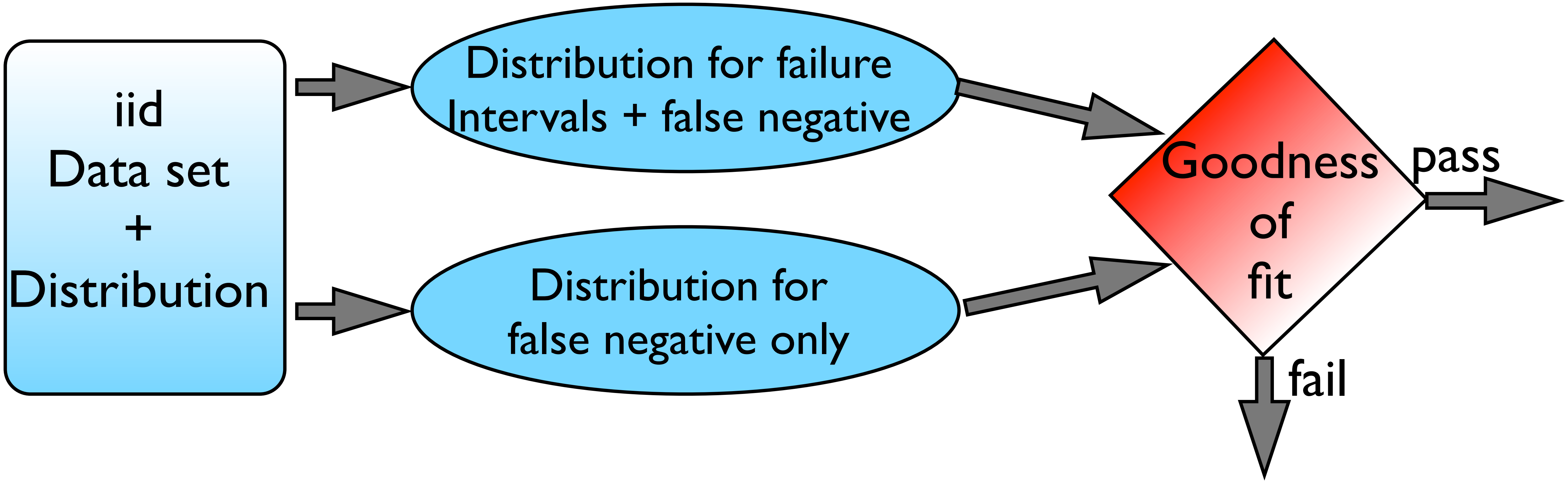
Table: Statistical fitting all random (fitting parameters scale are in seconds)

System name	Failures				False negative			
	Mean	CV	Best Fit	KS	Mean	CV	Best Fit	KS
Blue Gene/L	1040.5	0.92	exponential $\mu = 62431.3$	0.10	1888.1	1.10	exponential $\mu = 113289$	0.79
LANL Sys 3	3595.1	1.1	exponential $\mu = 215705$	0.98	6559.0	1.1	exponential $\mu = 393538$	0.70
LANL Sys 4	3409.1	1.1	exponential $\mu = 204544$	0.77	6187.0	1.1	exponential $\mu = 371218$	0.99
LANL Sys 5	3294.5	1.1	exponential $\mu = 197671$	0.95	6377.9	1.2	exponential $\mu = 382671$	0.35
LANL Sys 6	16796.7	0.9	exponential $\mu = 1007800$	0.81	31878.2	1.1	exponential $\mu = 1912690$	0.99
LANL Sys 23	9288.2	1.3	weibull a $= 509380$ b $= 0.846905$	0.97	16272.3	1.2	weibull a $= 895274$ b $= 0.851258$	0.98

Table: Statistical Fitting false negative random

System name	False negative			
	Mean	CV	Best Fit	KS
LANL Sys 8	7859.6	1.4	weibull a $= 401499$ b $= 0.767798$	0.74
LANL Sys 10	8247.0	3.6	weibull a $= 318087$ b $= 0.647838$	0.29
LANL Sys 11	6353.5	3.0	weibull a $= 232647$ b $= 0.609348$	0.61
LANL Sys 13	8164.3	3.9	lognormal $\mu = 11.5257$ $\sigma = 1.87004$	0.14
LANL Sys 14	11351.0	2.5	weibull a $= 391931$ b $= 0.559039$	0.77
LANL Sys 15	12136.7	1.2	exponential $\mu = 728203$	0.17
LANL Sys 16	3430.6	1.3	weibull a $= 182624$ b $= 0.810939$	0.69
LANL Sys 18	818.6	1.5	lognormal $\mu = 10.1123$ $\sigma = 1.28677$	0.37
LANL Sys 19	863.6	1.4	exponential $\mu = 29000.5$	0.18
LANL Sys 21	1986.9	2.3	lognormal $\mu = 10.6382$ $\sigma = 1.46402$	0.85

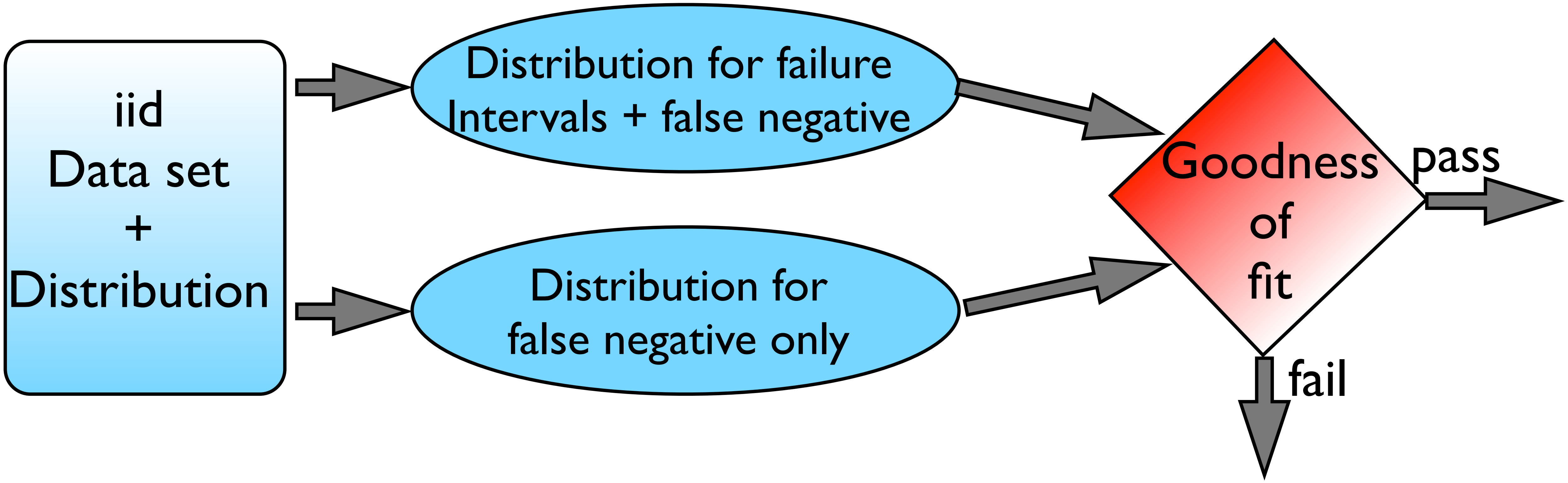
Methodology: Goodness of fit



Method:

- Kolmogorov-Smirnov test
- Probability-Probability plot (PP-plot).

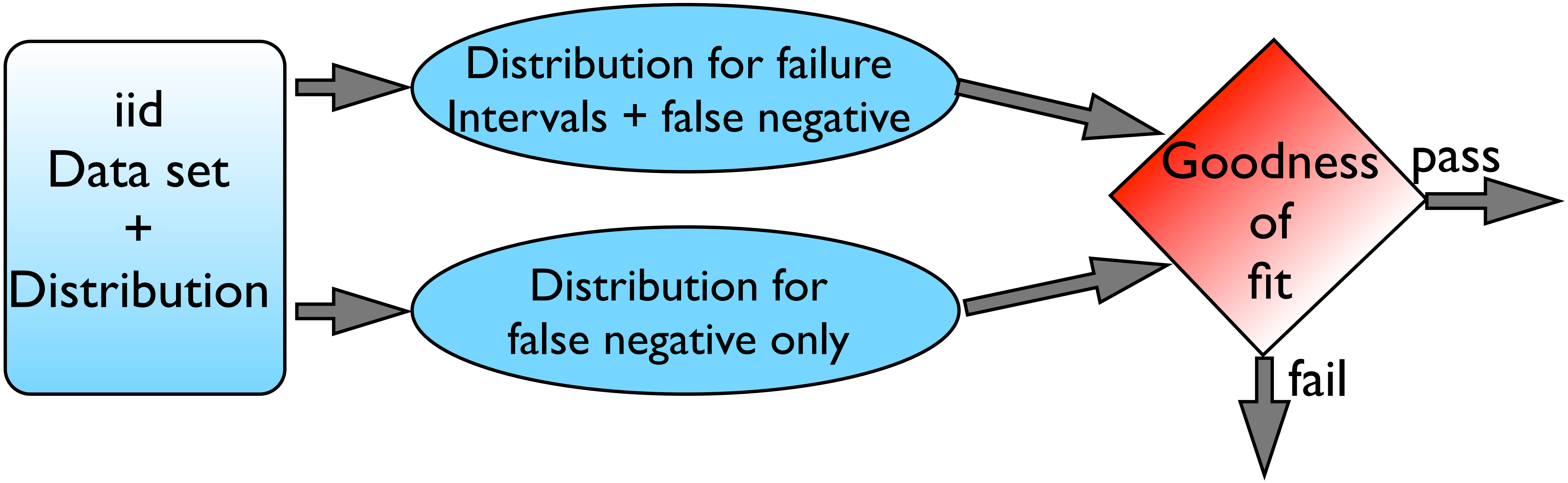
Goodness of fit outputs



System name	Failures				False negative			
	Mean	CV	Best Fit	KS	Mean	CV	Best Fit	KS
Blue Gene/L	1040.5	0.92	exponential $\mu = 62431.3$	0.10	1888.1	1.10	exponential $\mu = 113289$	0.79
LANL Sys 3	3595.1	1.1	exponential $\mu = 215705$	0.98	6559.0	1.1	exponential $\mu = 393538$	0.70
LANL Sys 4	3409.1	1.1	exponential $\mu = 204544$	0.77	6187.0	1.1	exponential $\mu = 371218$	0.99
LANL Sys 5	3294.5	1.1	exponential $\mu = 197671$	0.95	6377.9	1.2	exponential $\mu = 382671$	0.35
LANL Sys 6	16796.7	0.9	exponential $\mu = 1007800$	0.81	31878.2	1.1	exponential $\mu = 1912690$	0.99
LANL Sys 23	9288.2	1.3	weibull a = 509380 b = 0.846905	0.97	16272.3	1.2	weibull a = 895274 b = 0.851258	0.98

System name	False negative			
	Mean	CV	Best Fit	KS
LANL Sys 8	7859.6	1.4	weibull a = 401499 b = 0.767798	0.74
LANL Sys 10	8247.0	3.6	weibull a = 318087 b = 0.647838	0.29
LANL Sys 11	6353.5	3.0	weibull a = 232647 b = 0.609348	0.61
LANL Sys 13	8164.3	3.9	lognormal $\mu = 11.5257$ $\sigma = 1.87004$	0.14
LANL Sys 14	11351.0	2.5	weibull a = 391931 b = 0.559039	0.77
LANL Sys 15	12136.7	1.2	exponential $\mu = 728203$	0.17
LANL Sys 16	3430.6	1.3	weibull a = 182624 b = 0.810939	0.69
LANL Sys 18	818.6	1.5	lognormal $\mu = 10.1123$ $\sigma = 1.28677$	0.37
LANL Sys 19	863.6	1.4	exponential $\mu = 29000.5$	0.18
LANL Sys 21	1986.9	2.3	lognormal $\mu = 10.6382$ $\sigma = 1.46402$	0.85

Goodness of fit outputs

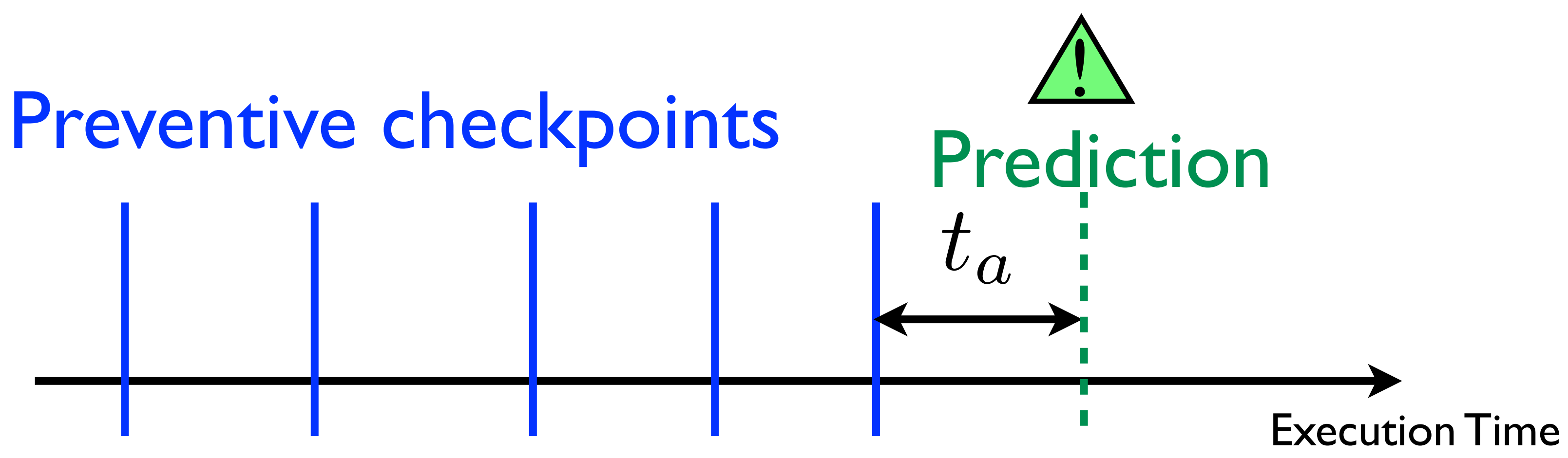


System name	Failures				False negative			
	Mean	CV	Best Fit	KS	Mean	CV	Best Fit	KS
Blue Gene/L	1040.5	0.92	exponential $\mu = 62431.3$	0.10	1888.1	1.10	exponential $\mu = 113289$	0.79
LANL Sys 3	3595.1	1.1	exponential $\mu = 215705$	0.98	6559.0	1.1	exponential $\mu = 393538$	0.70
LANL Sys 4	3409.1	1.1	exponential $\mu = 204544$	0.77	6187.0	1.1	exponential $\mu = 371218$	0.99
LANL Sys 5	3294.5	1.1	exponential $\mu = 197671$	0.95	6377.9	1.2	exponential $\mu = 382671$	0.35
LANL Sys 6	16796.7	0.9	exponential $\mu = 1007800$	0.81	31878.2	1.1	exponential $\mu = 1912690$	0.99
LANL Sys 23	9288.2	1.3	weibull a = 509380 b = 0.846905	0.97	16272.3	1.2	weibull a = 895274 b = 0.851258	0.98

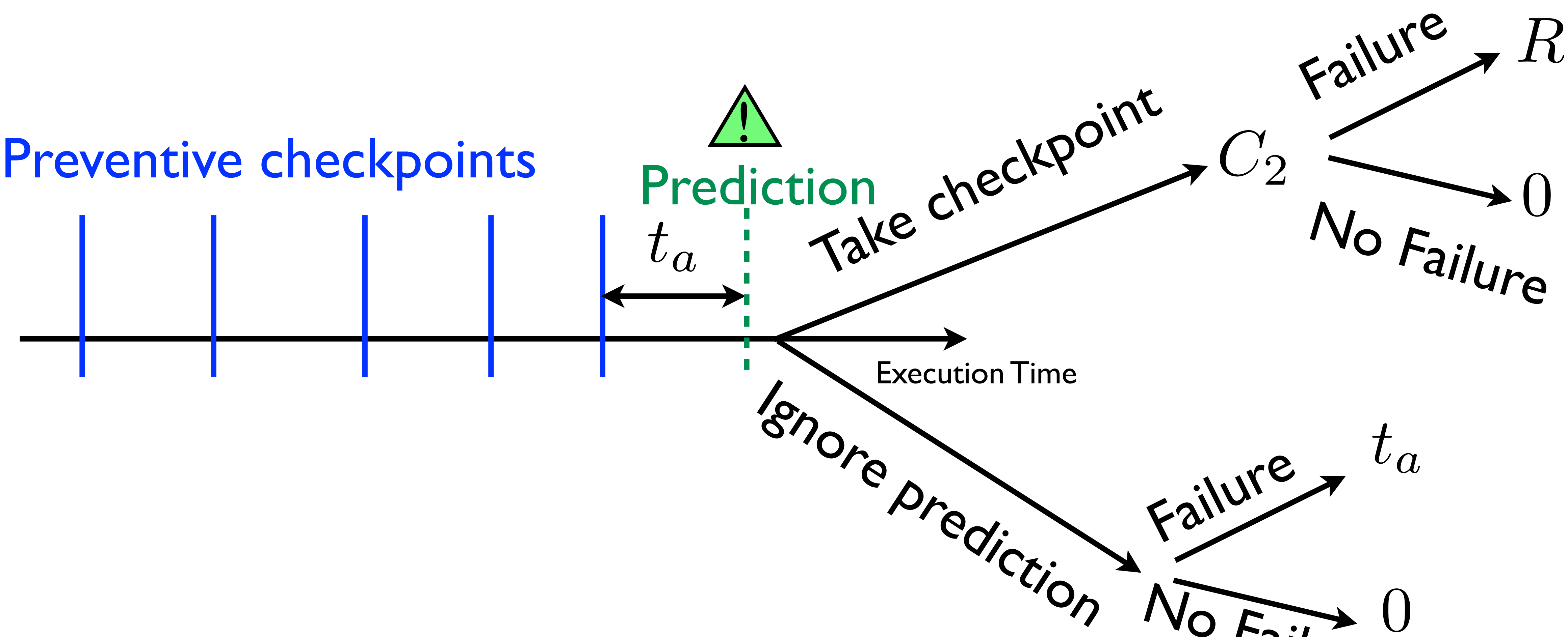
System name	False negative			
	Mean	CV	Best Fit	KS
LANL Sys 8	7859.6	1.4	weibull a = 401499 b = 0.767798	0.74
LANL Sys 10	8247.0	3.6	weibull a = 318087 b = 0.647838	0.29
LANL Sys 11	6353.5	3.0	weibull a = 232647 b = 0.609348	0.61
LANL Sys 13	8164.3	3.9	lognormal $\mu = 11.5257$ $\sigma = 1.87004$	0.14
LANL Sys 14	11351.0	2.5	weibull a = 391931 b = 0.559039	0.77
LANL Sys 15	12136.7	1.2	exponential $\mu = 728203$	0.17
LANL Sys 16	3430.6	1.3	weibull a = 182624 b = 0.810939	0.69
LANL Sys 18	818.6	1.5	lognormal $\mu = 10.1123$ $\sigma = 1.28677$	0.37
LANL Sys 19	863.6	1.4	exponential $\mu = 29000.5$	0.18
LANL Sys 21	1986.9	2.3	lognormal $\mu = 10.6382$ $\sigma = 1.46402$	0.85

Outline

Mathematical Modeling:proposed combination

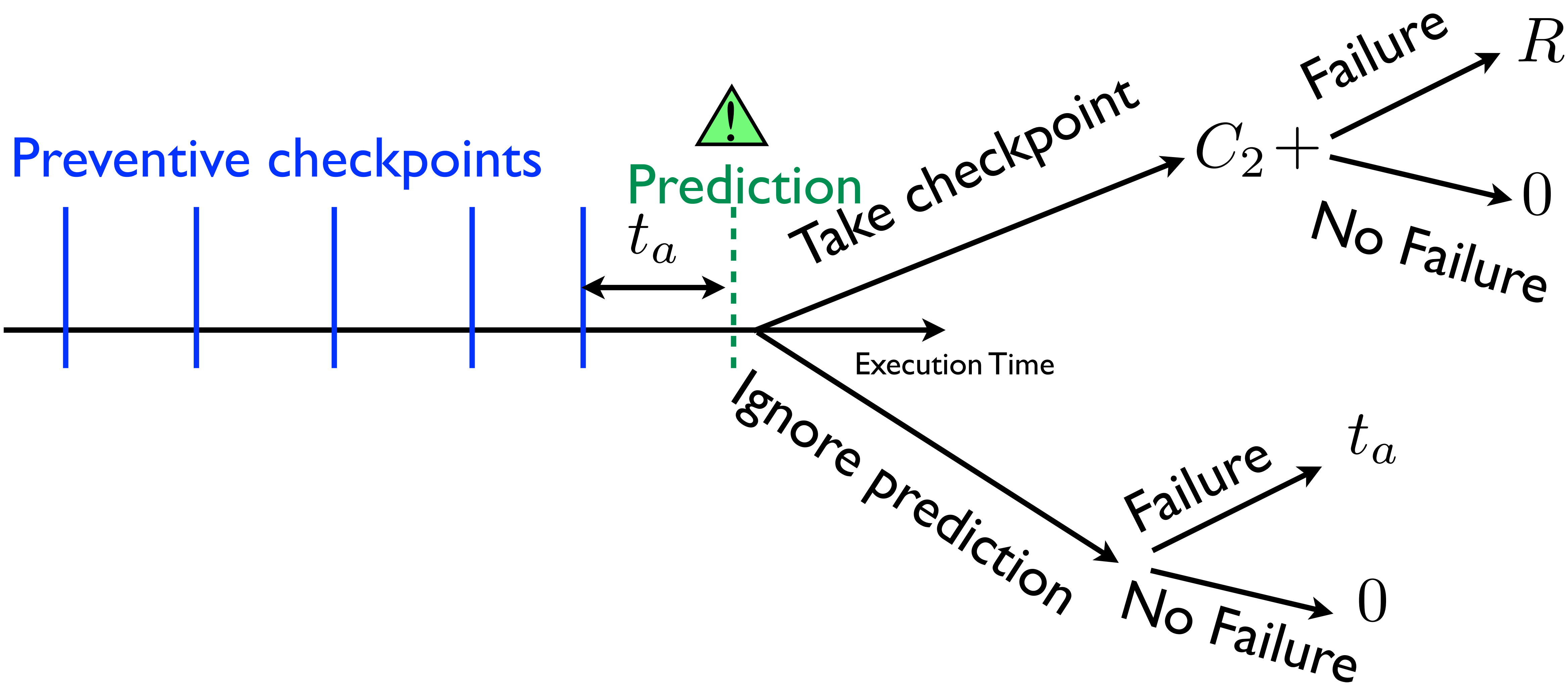


Mathematical Modeling:proposed combination



Failure happens with a probability p

Mathematical Modeling (Proactive decision)



The proactive action is performed iif

$$W_p \leq W_{np} \equiv \overline{p} c_2 / p \leq t_a$$

Mathematical Modeling

Preventive period

- Unpredicted failures are randomly distributed with a mean μ.
- The preventive checkpoint cost c_1.

Mathematical Modeling

Preventive period

- Unpredicted failures are randomly distributed with a mean μ.
- The preventive checkpoint cost c_1.

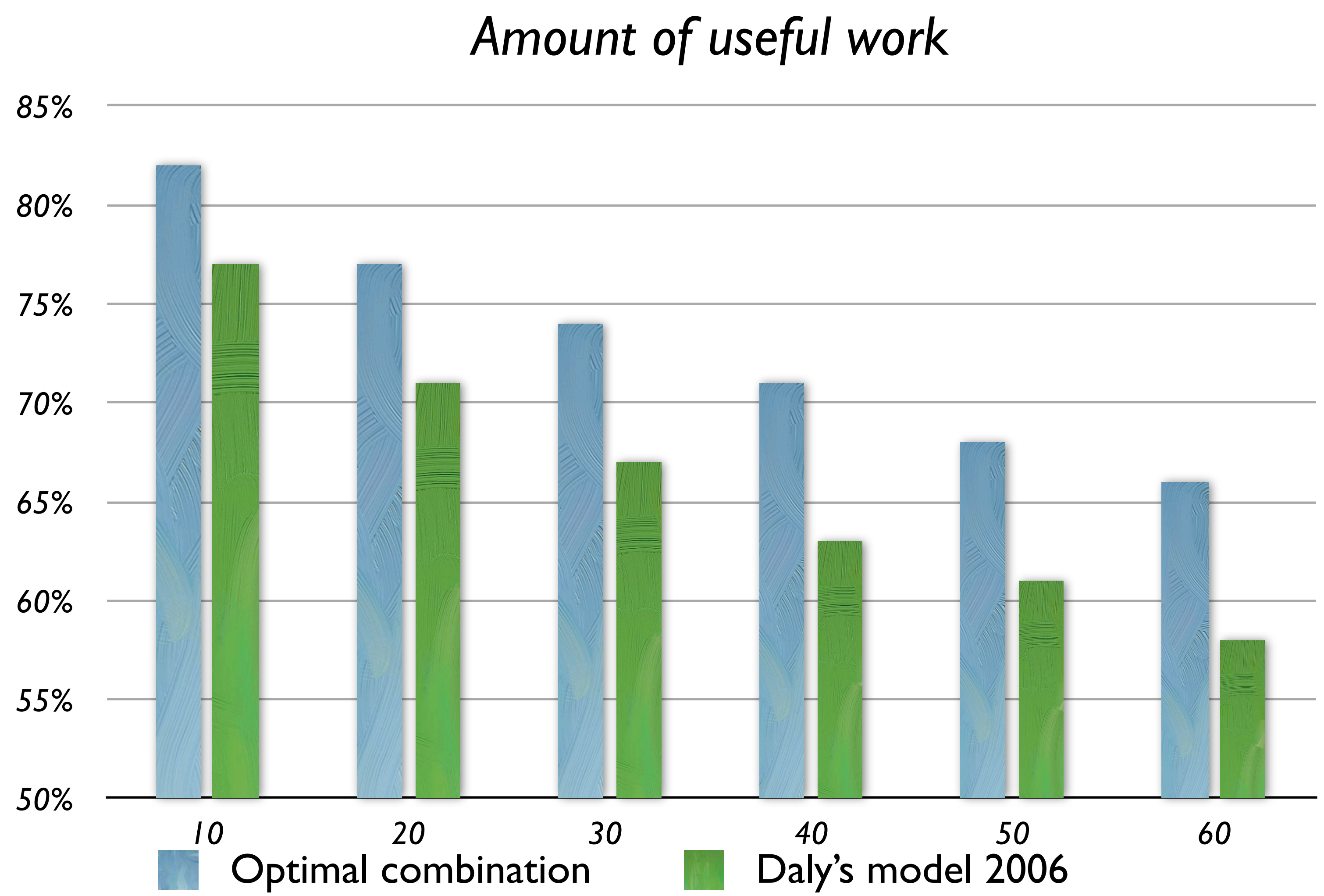
The first order approximation of the interval between preventive checkpoints:

$$\sqrt{2\mu c_1}$$

Simulation results

System 19 LANL actual failures data and prediction.

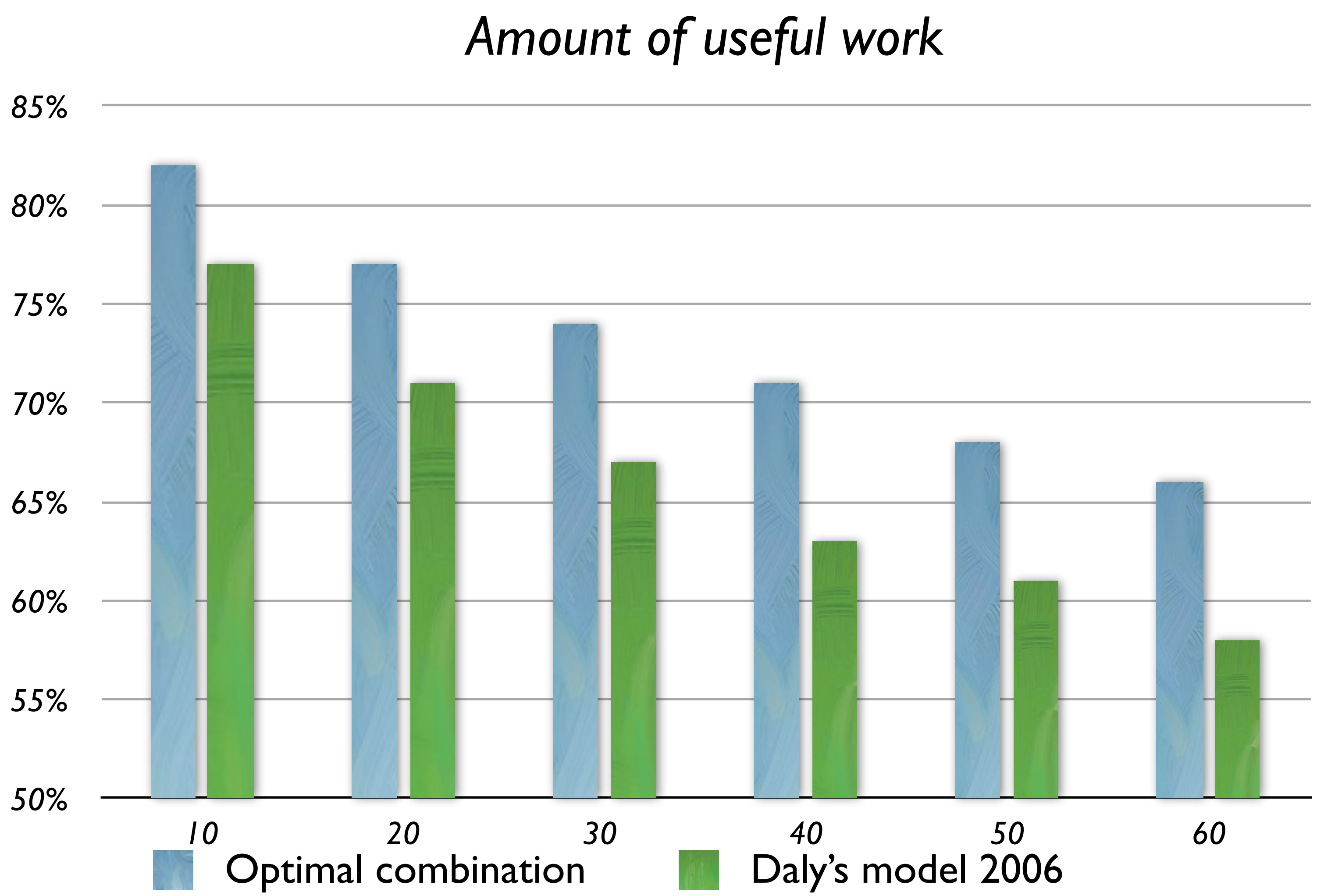
- More than 3,000 failures and 1,700 unpredicted failures.
- 45% recall and 90% precision.



Simulation results

System 19 LANL actual failures data and prediction.

- More than 3,000 failures and 1,700 unpredicted failures.
- 45% recall and 90% precision.



13% of improvement which is the theoretical peak for such configuration.

Outline

Conclusion

- Classification based on the randomness tests (iid vs non-iid)

Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)

Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
- Failure prediction mechanism catches the non-randomness and correlation.

Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
- Failure prediction mechanism catches the non-randomness and correlation.
- Failure prediction mechanism acts as a scale function and it affects only the scale parameter.

Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
- Failure prediction mechanism catches the non-randomness and correlation.
- Failure prediction mechanism acts as a scale function and it affects only the scale parameter.
- The peak of correlation on the initial traces has an important impact on the prediction results, specifically on the recall value

Future Work

- Analyze more deeply the set of systems with a high correlation like system 2 or 20 and isolate sources of non-randomness.

Future Work

- Analyze more deeply the set of systems with a high correlation like system 2 or 20 and isolate sources of non-randomness.
- Investigate if a cross-correlation of different time scale has an impact of the prediction mechanism.

Future Work

- Analyze more deeply the set of systems with a high correlation like system 2 or 20 and isolate sources of non-randomness.
- Investigate if a cross-correlation of different time scale has an impact of the prediction mechanism.
- Manage the tradeoff between the precision and the recall.

Questions ?