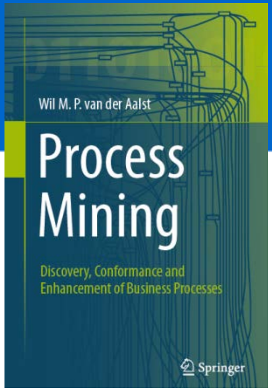
Process Mining: Data Science in Action

Four Quality Criteria for Process Discovery

prof.dr.ir. Wil van der Aalst
www.processmining.org

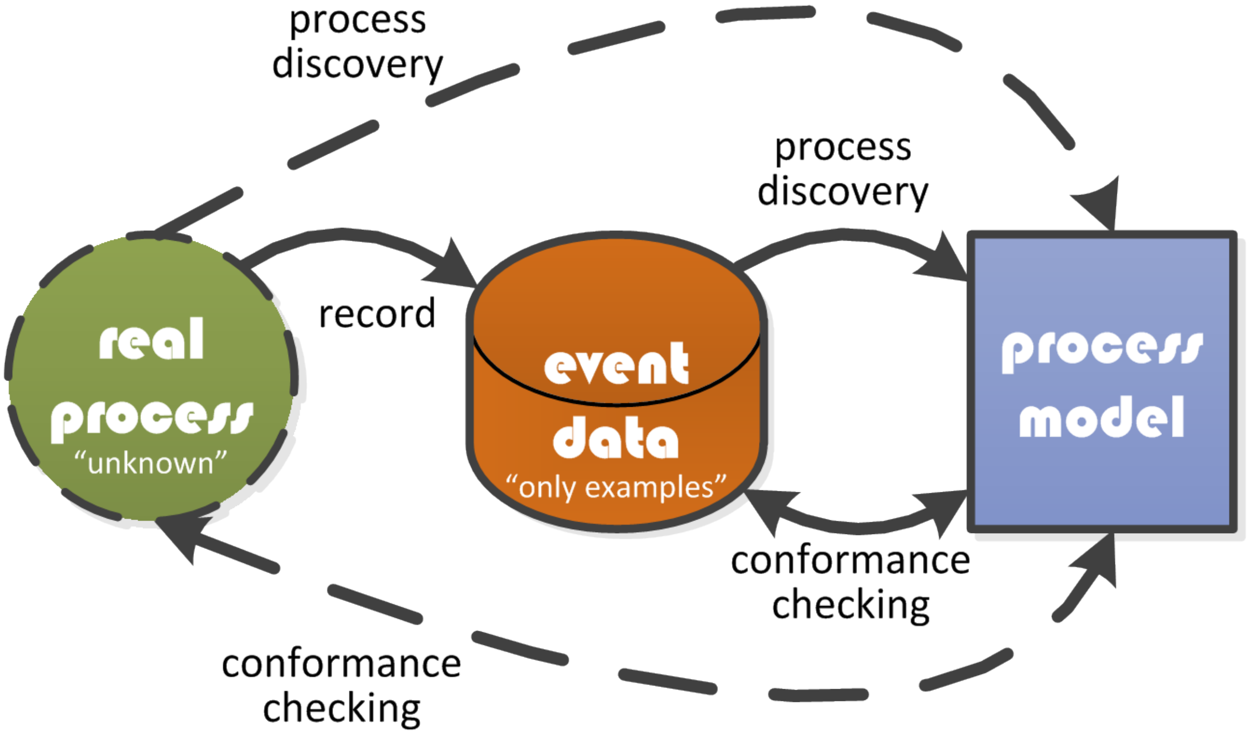


Technische Universiteit Eindhoven University of Technology

Where innovation starts


HIGH
QUALITY LEVEL
MED
LOW
fitness
lift
ability to explain observed behavior
precision
drag
avoiding underfitting
avoiding overfitting
thrust
generalization
Process Mining
Occam's Razor
gravity
simplicity

Overview



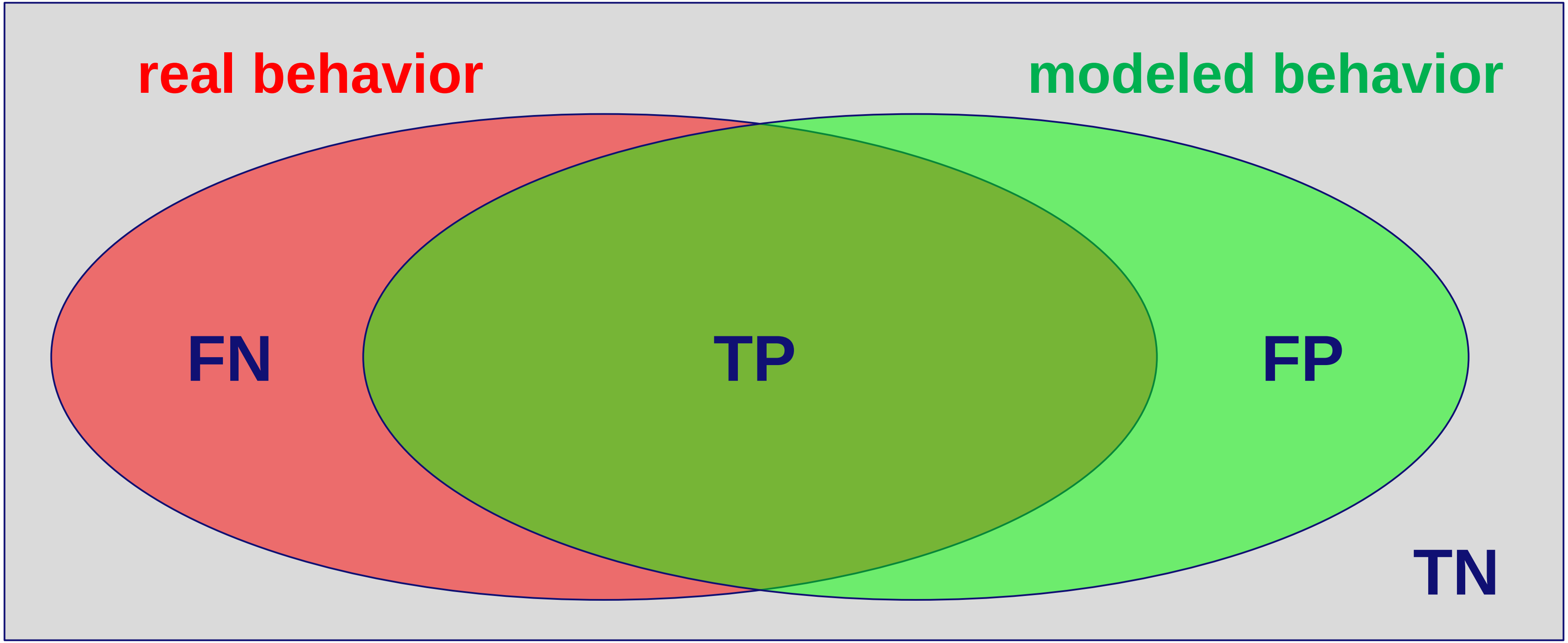
Is the process model a correct reflection of the real process?

Naïve approach based on classification

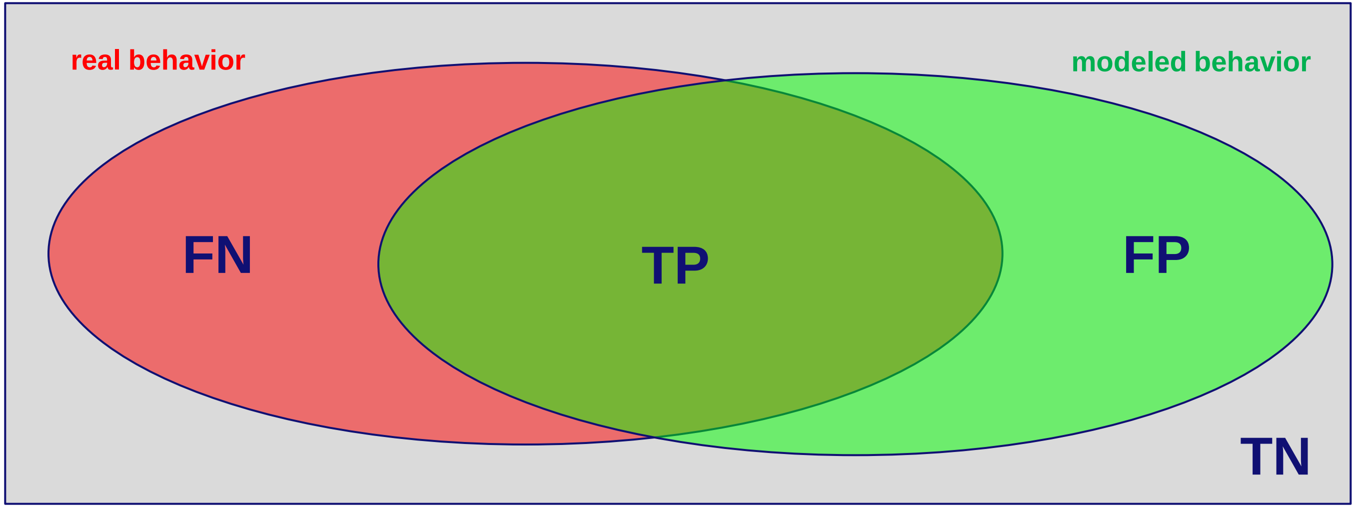
- **True Positives (TP): traces possible in model and also possible in real process.**
- **True Negatives (TN): traces not possible in model and also not possible in real process.**
- **False Positives (FP): traces possible in model but not possible in real process.**
- **False Negatives (FN): traces not possible model but possible in real process.**

		predicted class	
		+	-
actual class	+	TP	FN
	-	FP	TN

Visualization of True/False Positives/Negatives



Metrics

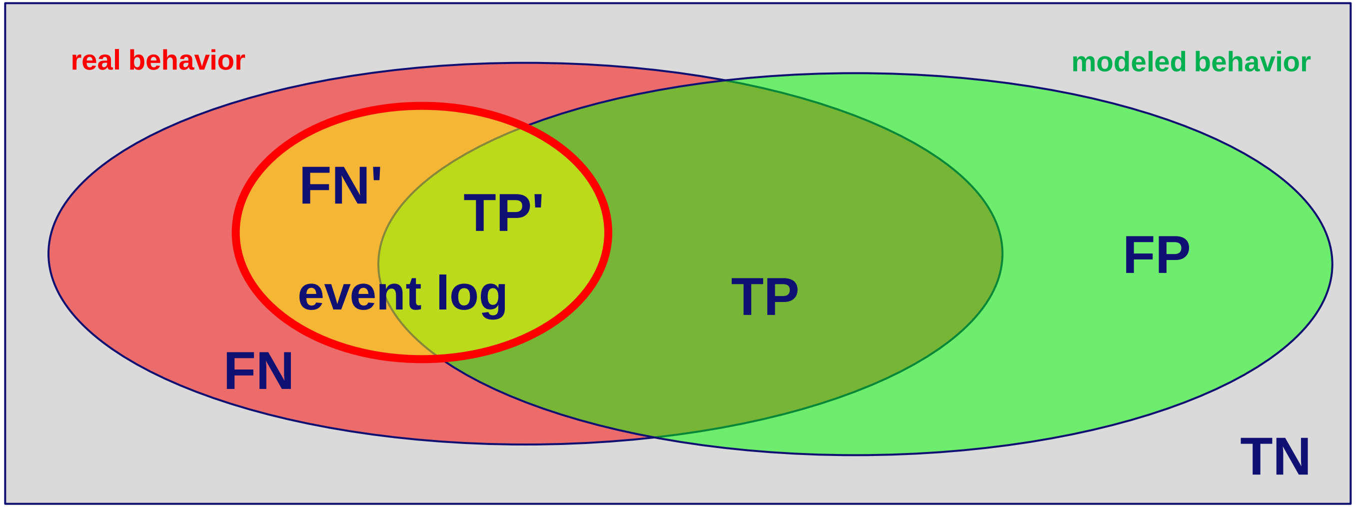


$$recall = \frac{TP}{TP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

Problem

Typically the event log only shows fraction of possible traces



$$\cancel{recall = \frac{TP}{TP + FN}}$$

$$\cancel{precision = \frac{TP}{TP + FP}}$$

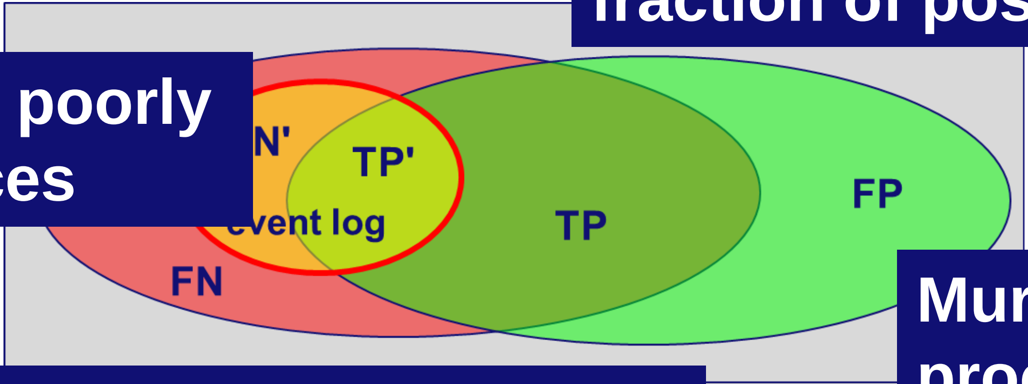
$$replay\_fitness = \frac{TP'}{TP' + FN'}$$

Challenges

No negative examples
(cannot see what cannot happen)

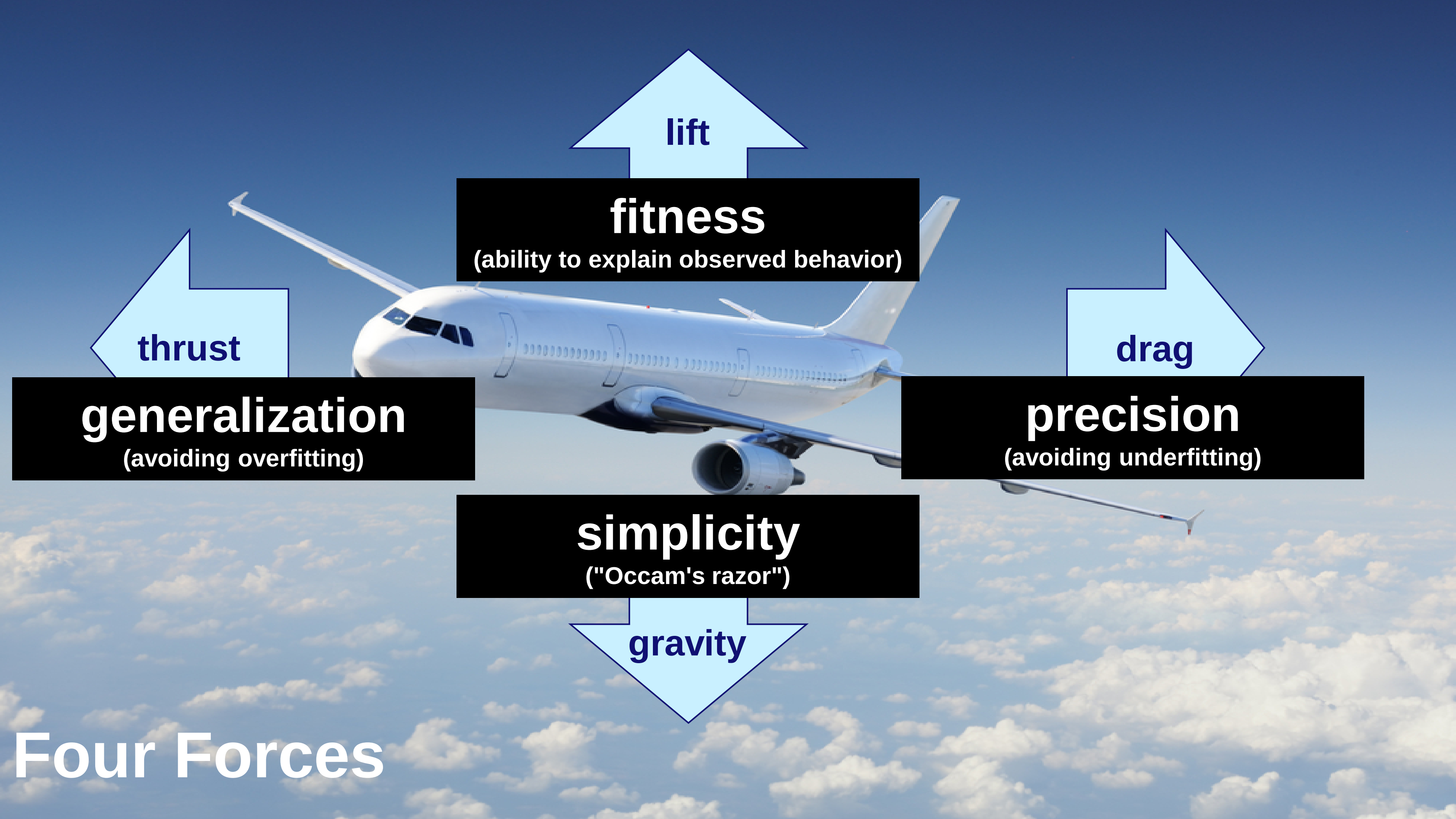
Log contains only a fraction of possible traces

Almost vs poorly fitting traces



Murphy's law for process mining
(anything is possible, so probabilities matter)

In case of loops often infinitely many possible traces

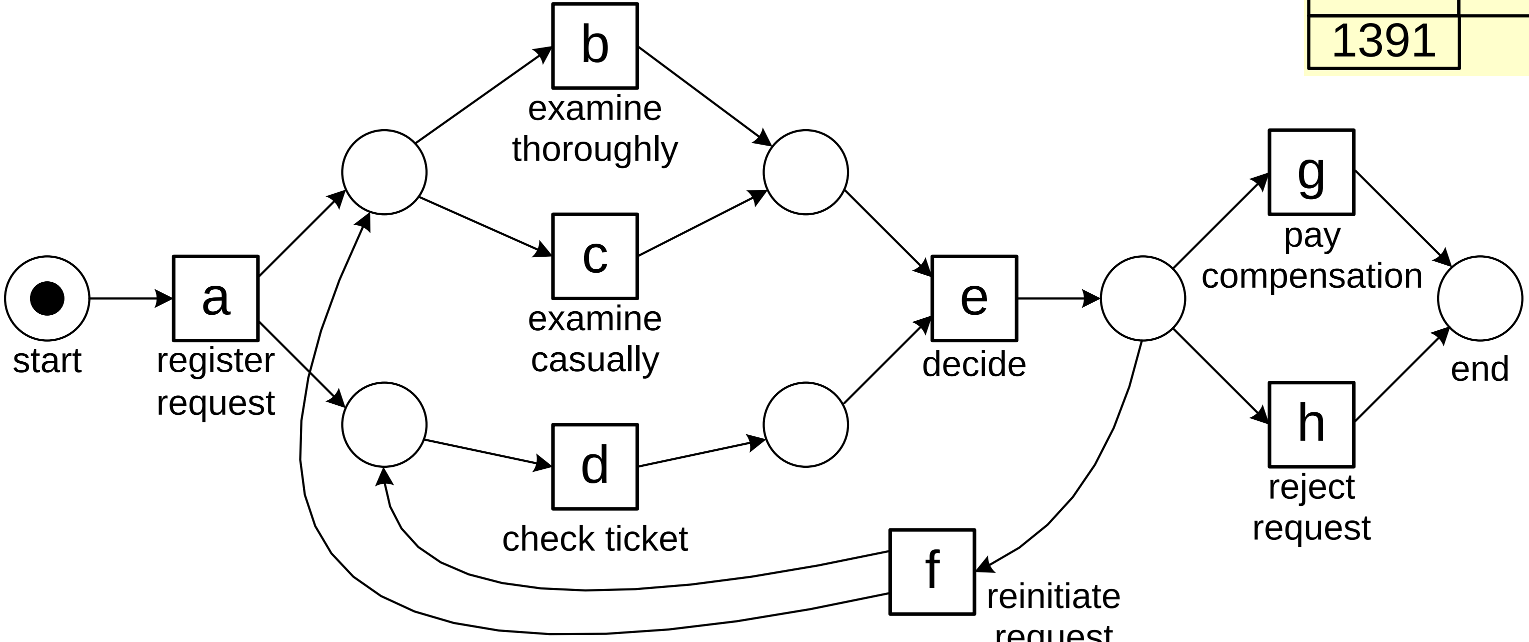

lift
fitness
(ability to explain observed behavior)
thrust
generalization
(avoiding overfitting)
drag
precision
(avoiding underfitting)
simplicity
("Occam's razor")
gravity
Four Forces

Example log

#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh

Model that seems to be OK ...

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
391	

fitness
(observed behavior fits)

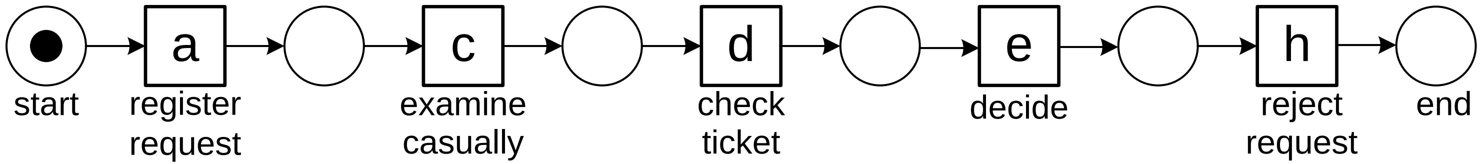
simplicity
("Occam's razor")

precision
(avoiding underfitting)

generalization
(avoiding overfitting)

Non-fitting model

#	trace
455	acdeh
191	abdeg
...	...
1	adcefdbefcdefdbeg
1391	



56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
391	

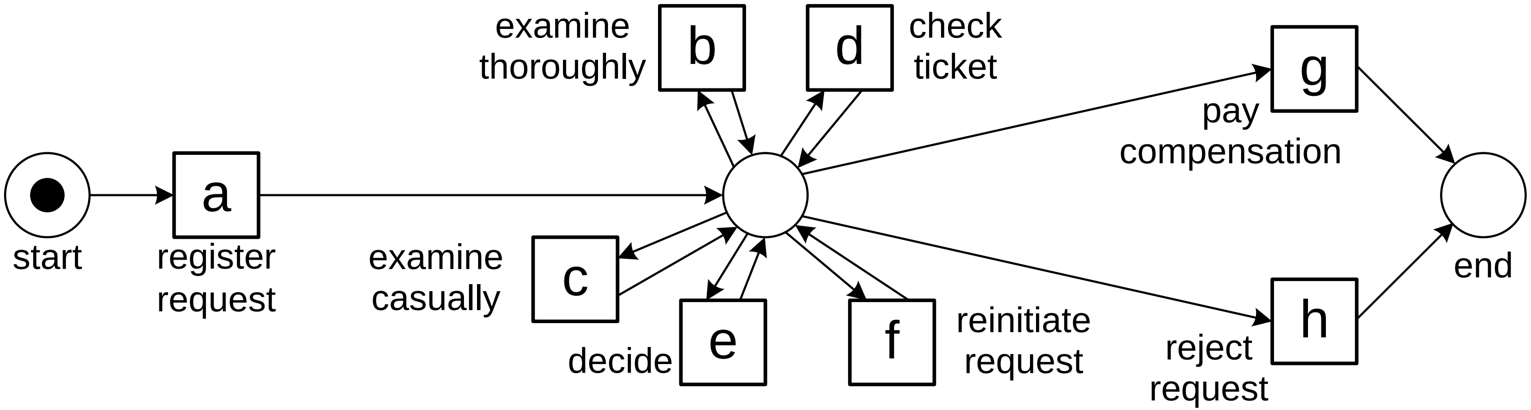
fitness
(observed behavior fits)

simplicity
("Occam's razor")

precision
(avoiding underfitting)

generalization
(avoiding overfitting)

Underfitting model



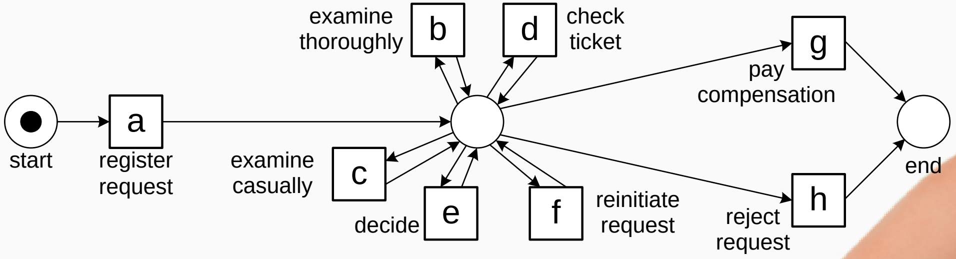
#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

fitness (observed behavior fits)

simplicity ("Occam's razor")

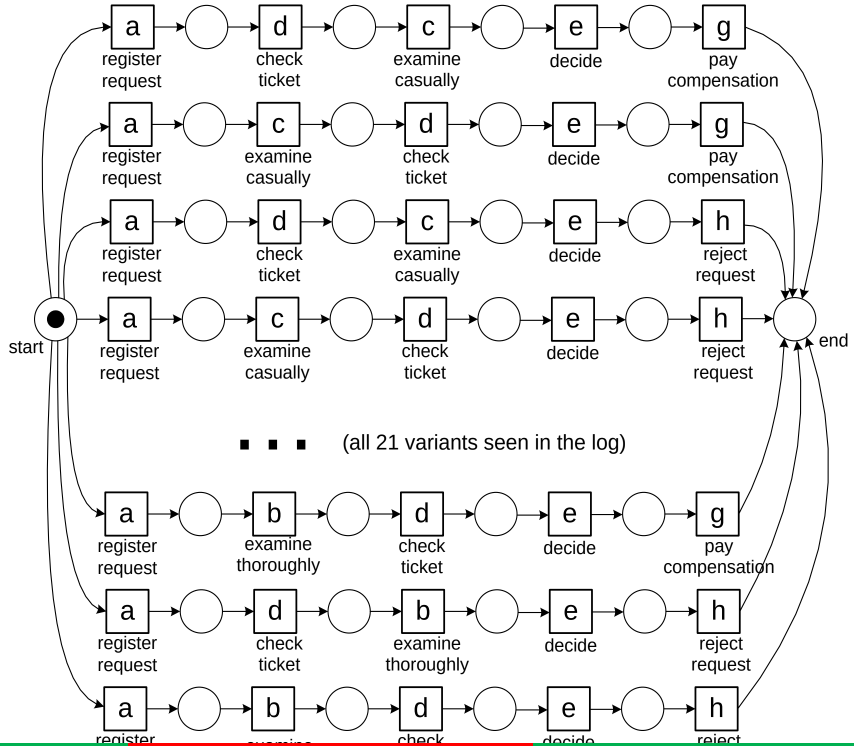
precision (avoiding underfitting)

generalization (avoiding overfitting)



underfitting

Overfitting model



#	trace
455	acdeh
191	abdeg
177	adceh
144	abdeh
111	acdeg
82	adceg
56	adbeh
47	acdefdbeh
38	adbeg
33	acdefbdeh
14	acdefbdeg
11	acdefdbeg
9	adcefcdeh
8	adcefdbeh
5	adcefbdeg
3	acdefbdefdbeg
2	adcefdbeg
2	adcefbdefbdeg
1	adcefdbefbdeh
1	adbefbdefdbeg
1	adcefdbefcdefdbeg
1391	

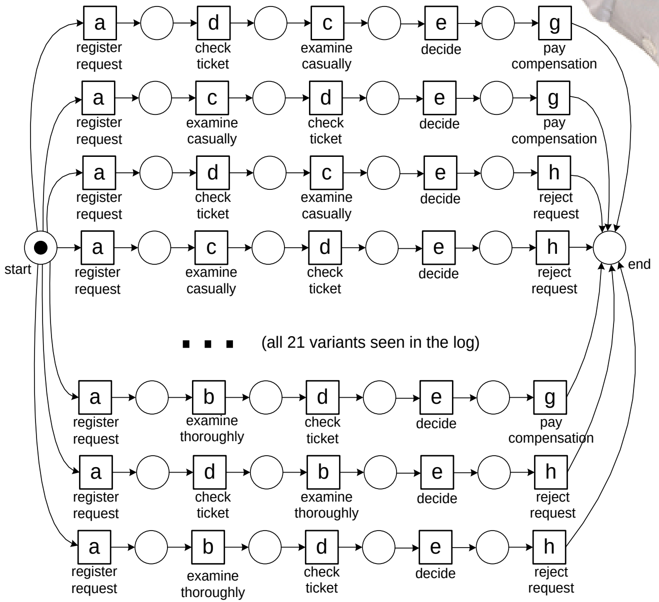
fitness
(observed behavior fits)

simplicity
("Occam's razor")

precision
(avoiding underfitting)

generalization
(avoiding overfitting)

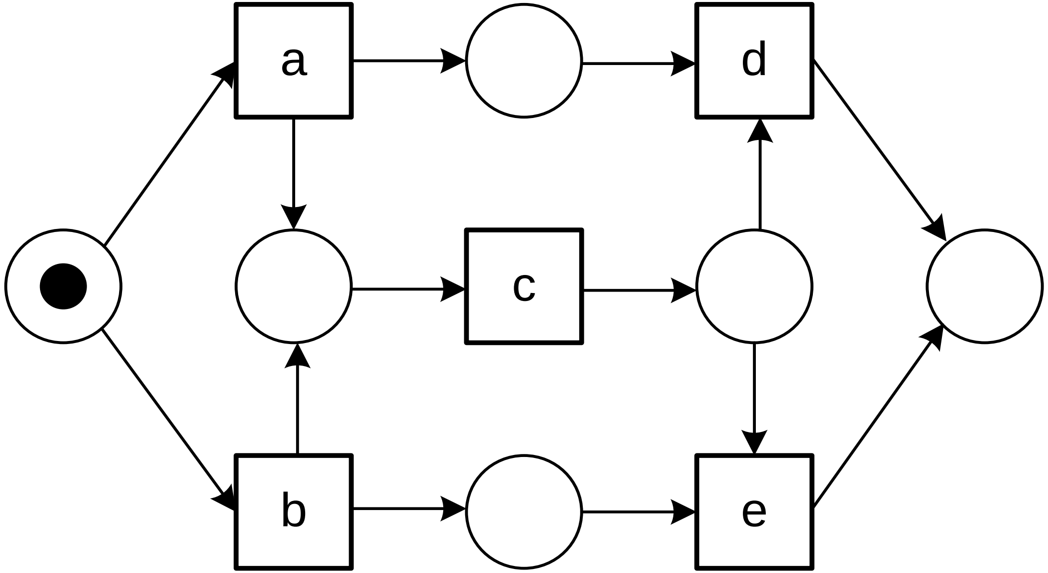
overfitting



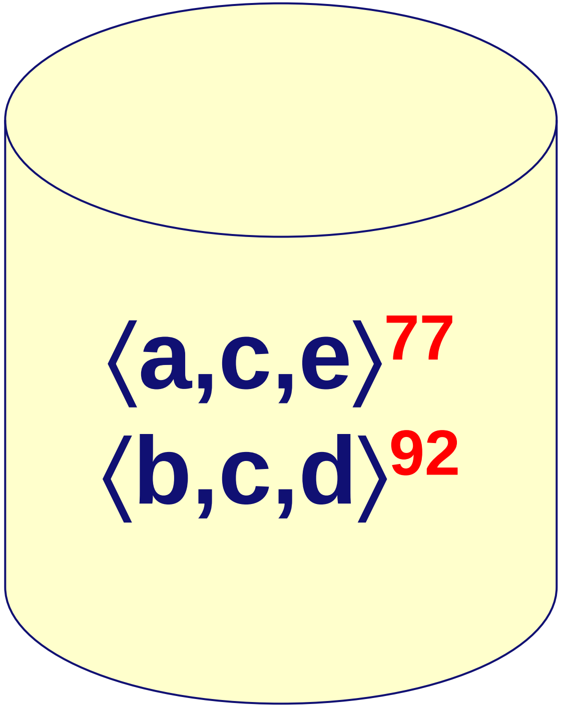
Fitness: good or bad?

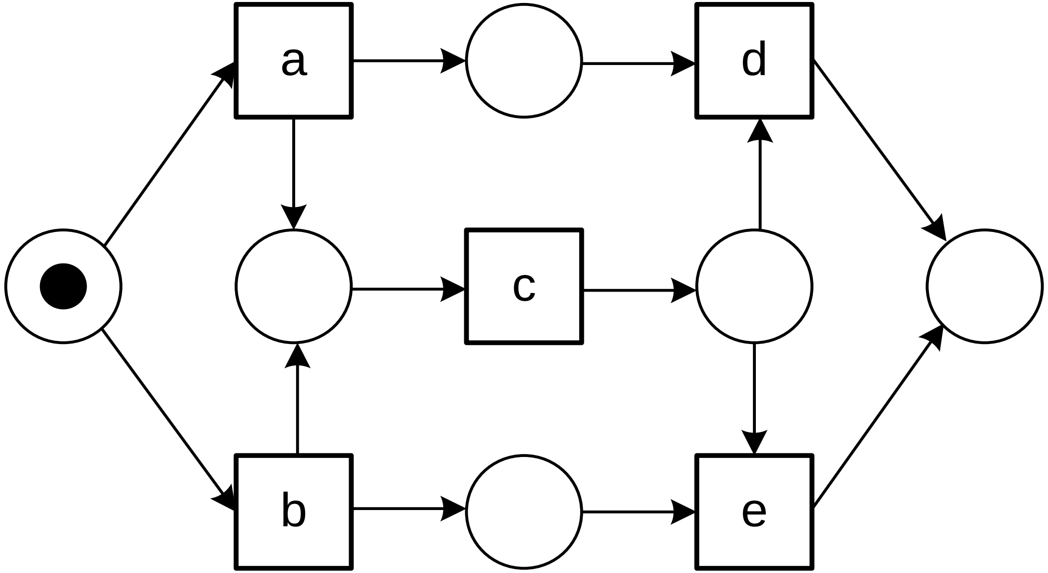
$\langle a,c,e \rangle^{77}$

$\langle b,c,d \rangle^{92}$



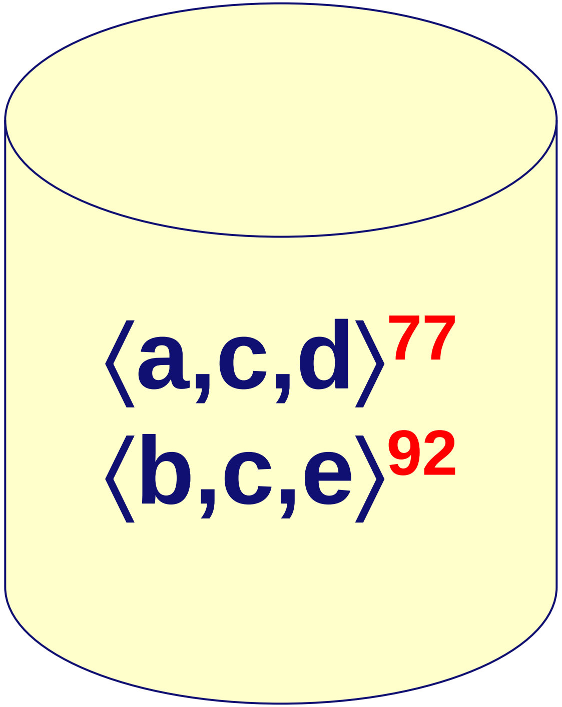
Fitness: bad!

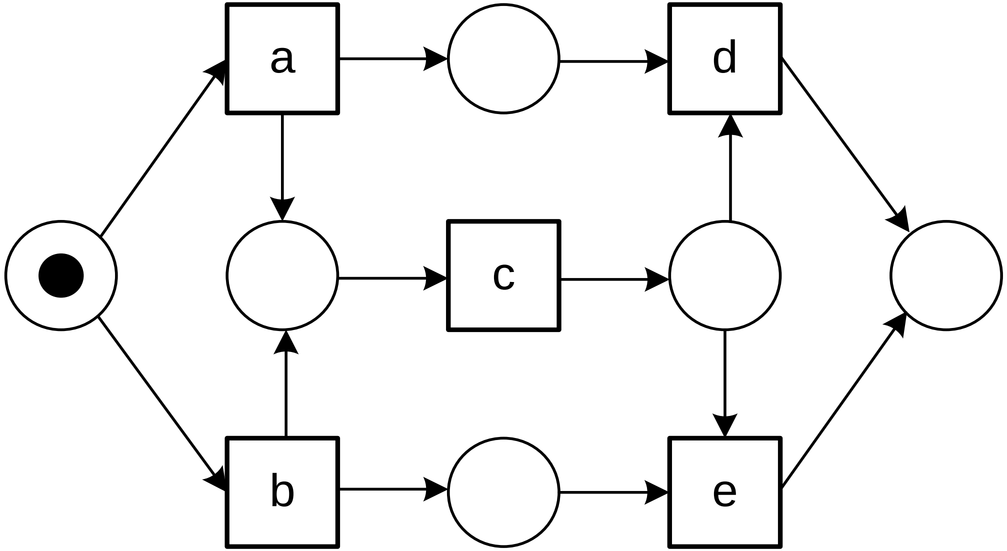




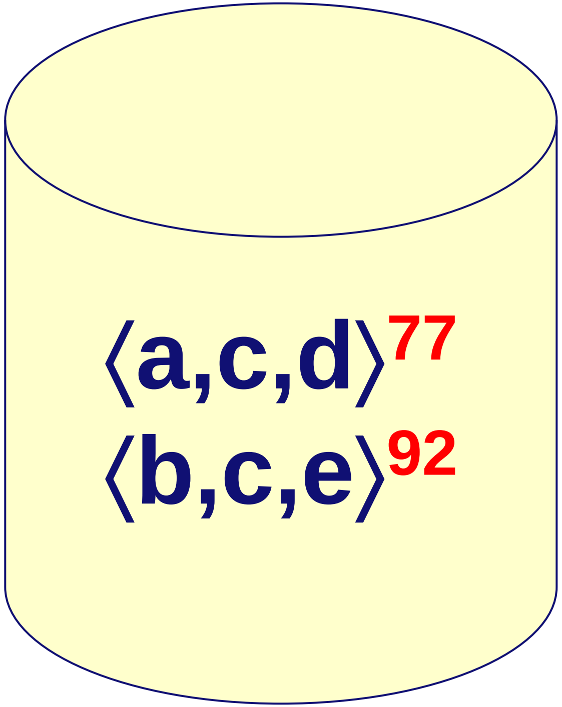
both traces do not fit ...

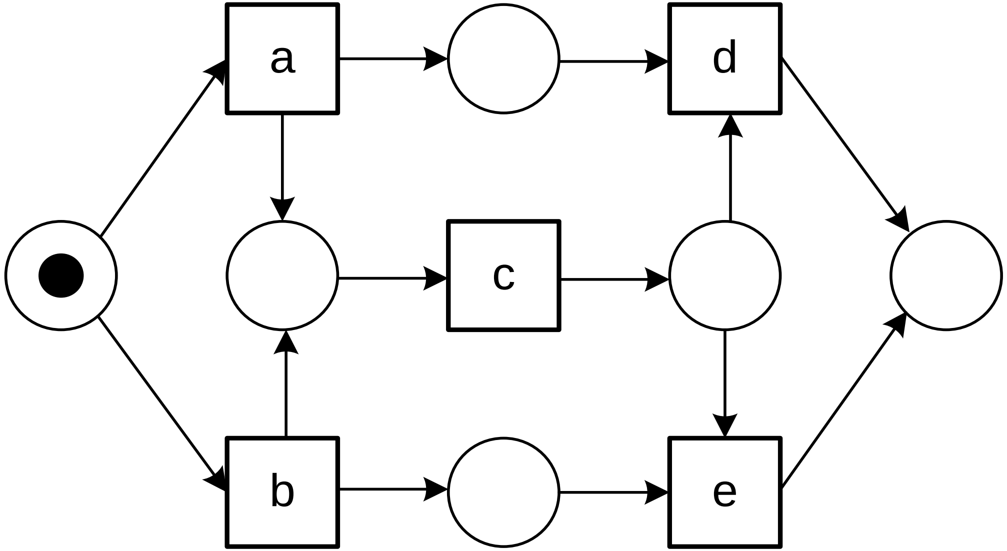
Precision: good or bad?





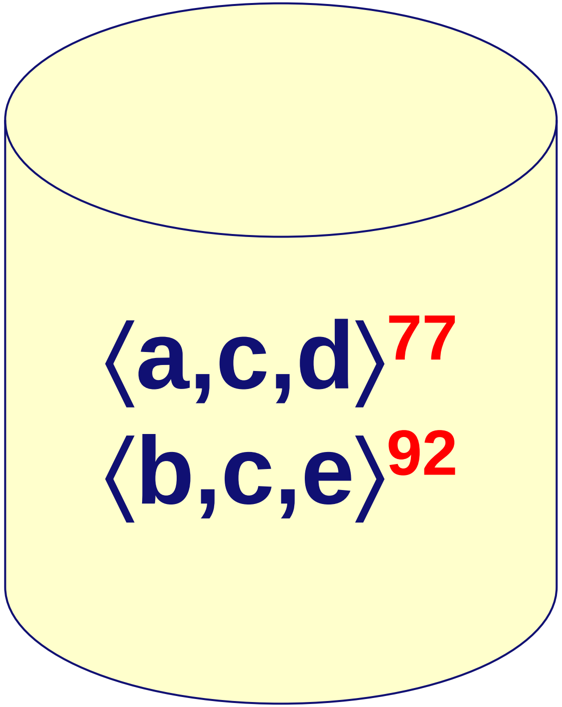
Precision: good!

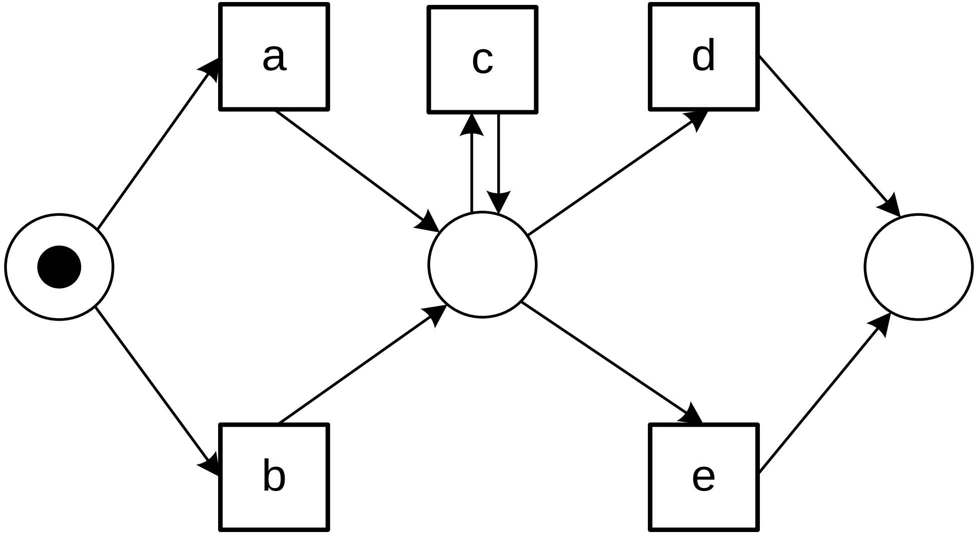




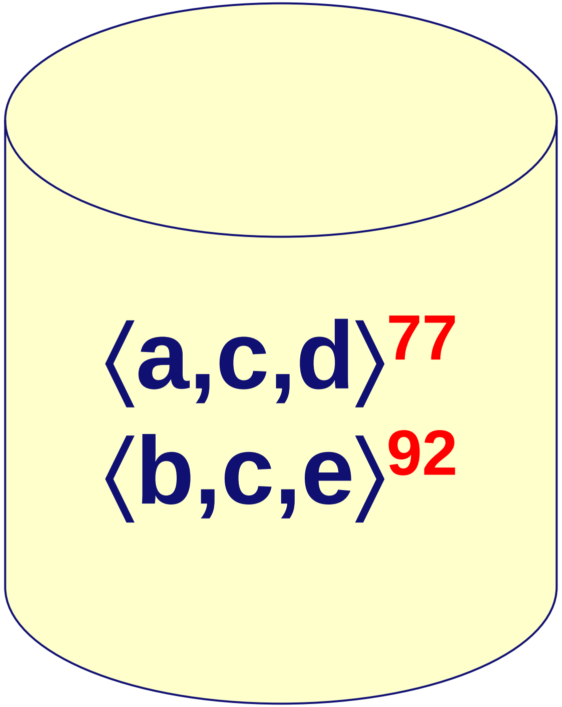
not underfitting…

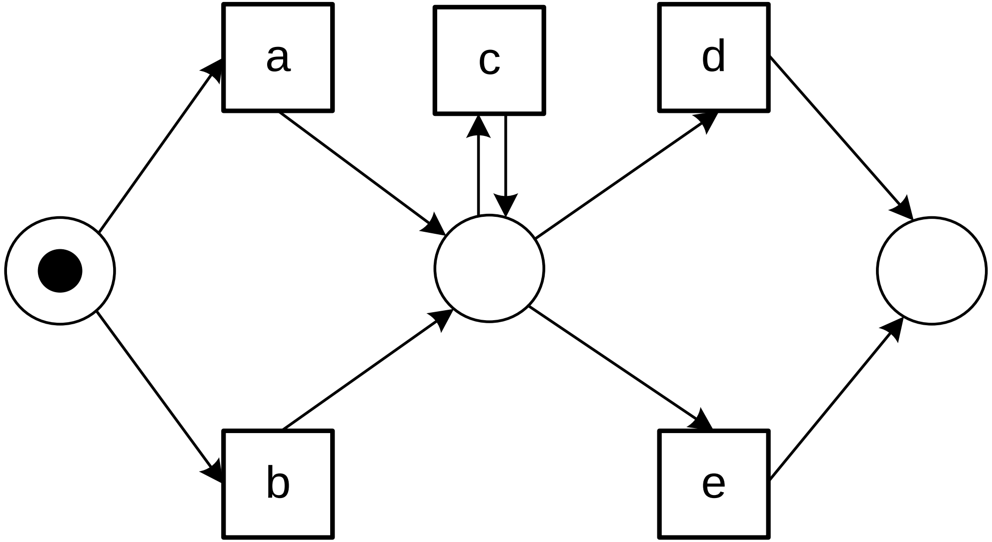
Precision: good or bad?





Precision: bad!

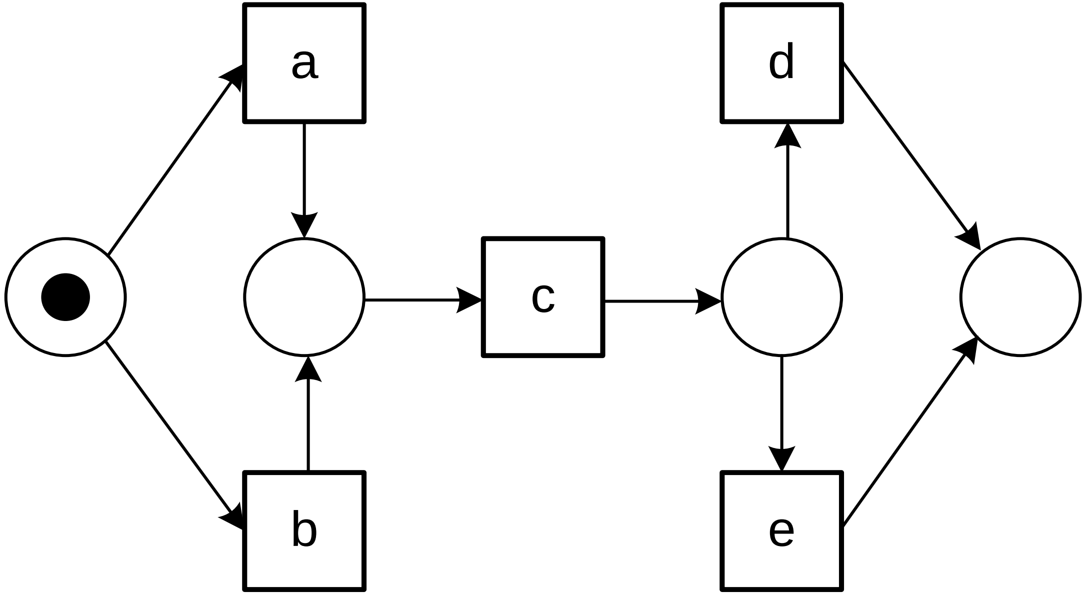




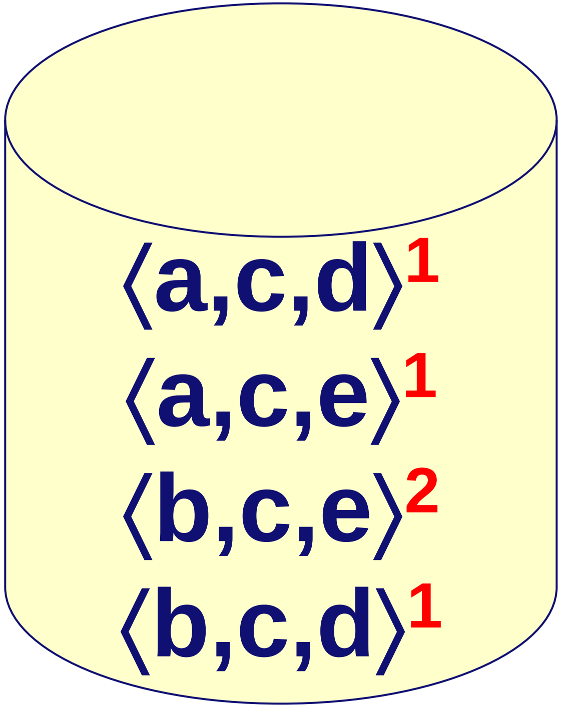
underfitting (allows for highly unlikely behavior) …

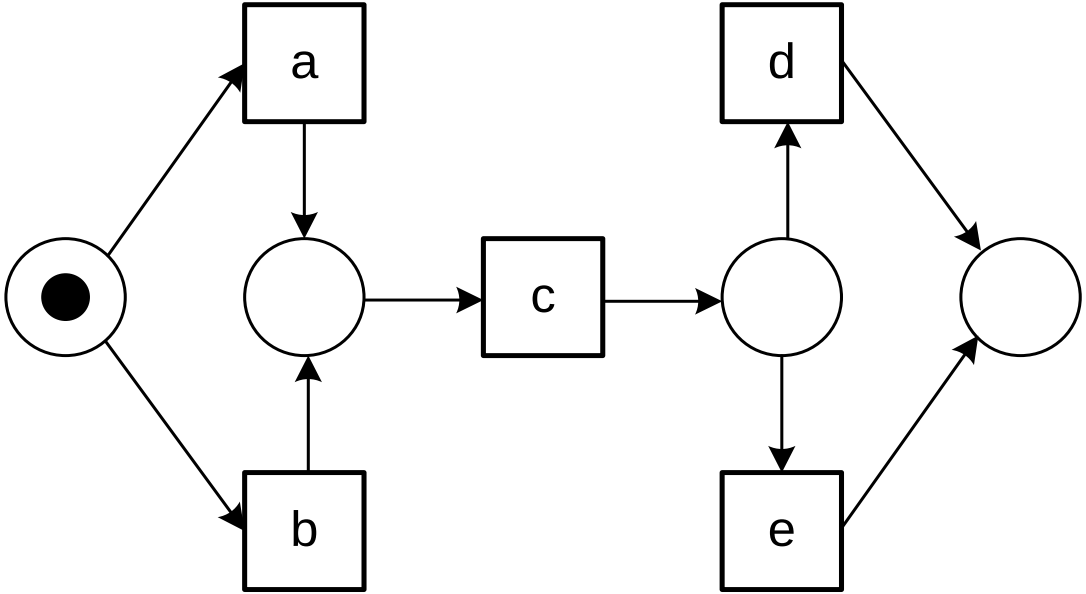
Generalization: good or bad?

$\langle a,c,d\rangle^{1}$
$\langle a,c,e\rangle^{1}$
$\langle b,c,e\rangle^{2}$
$\langle b,c,d\rangle^{1}$



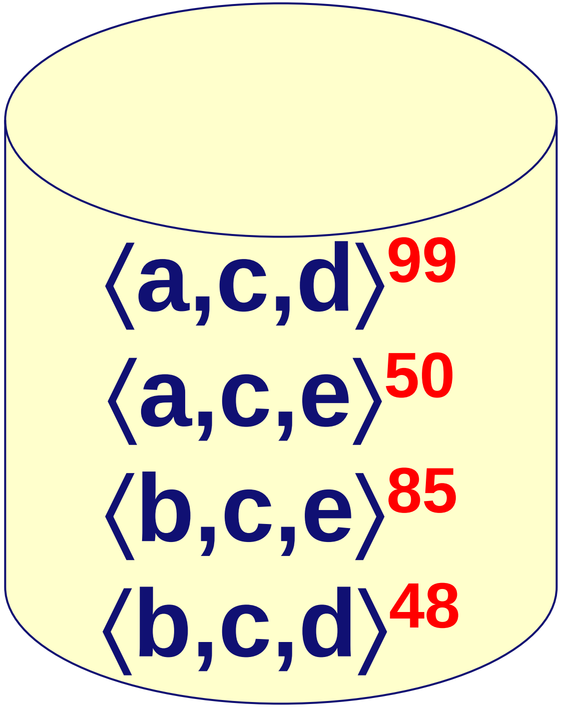
Generalization: bad!

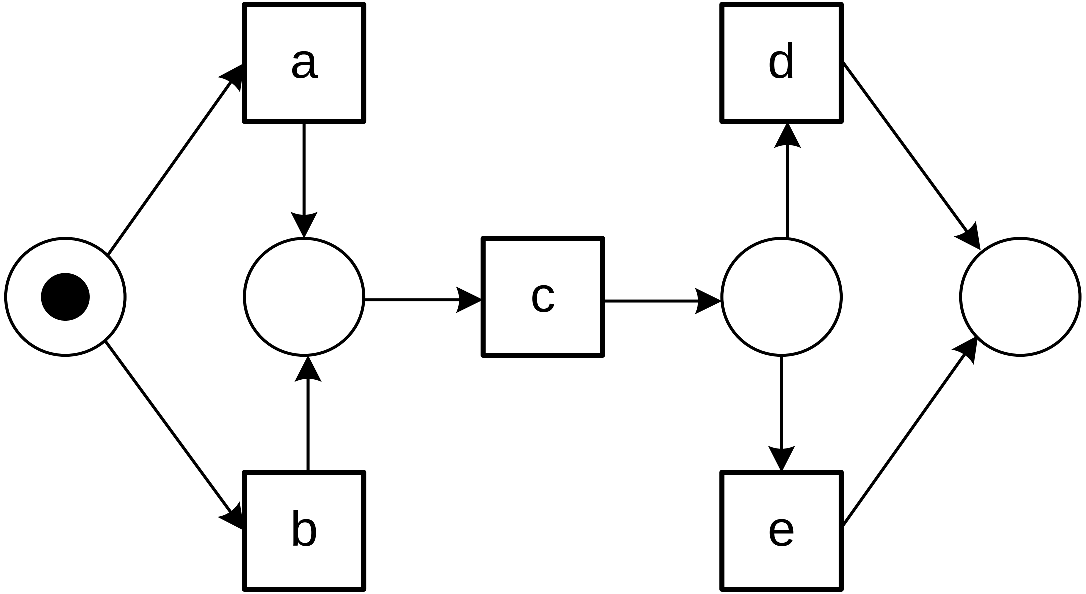




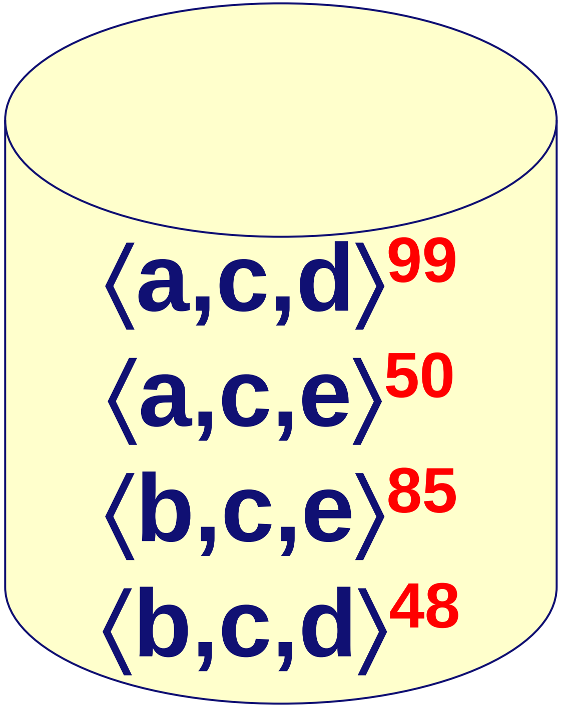
risk of overfitting on 5 example traces …

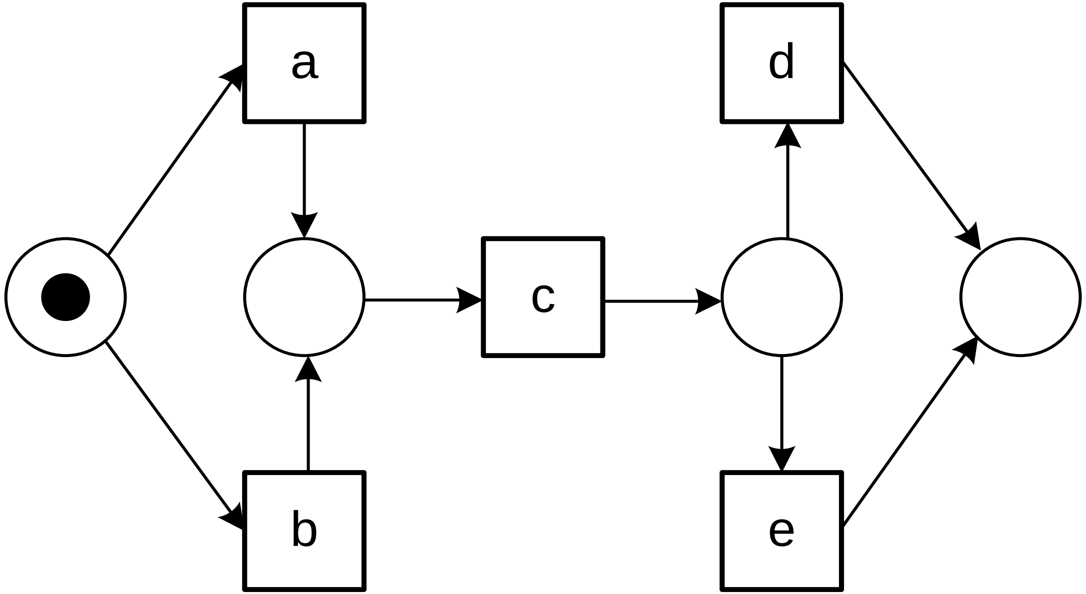
Generalization: good or bad?





Generalization: good!





not overfitting...

Simplicity: good or bad?

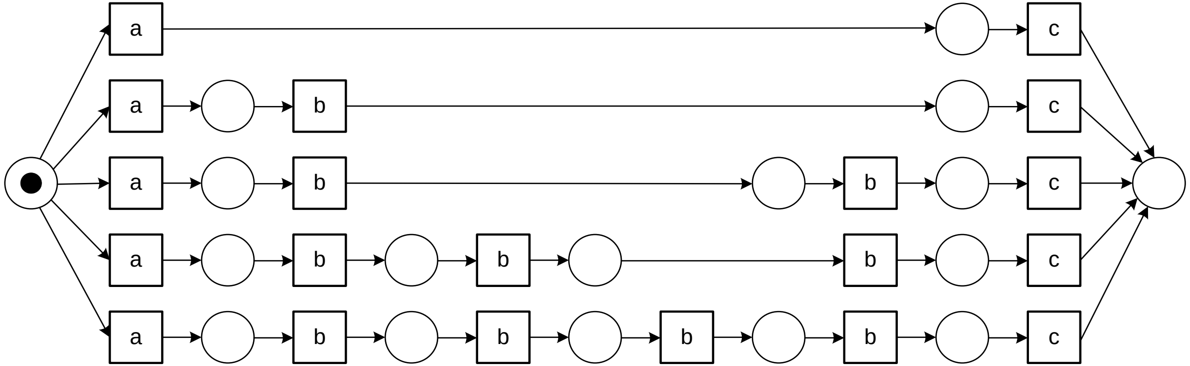
$\langle a,c \rangle^{16}$

$\langle a,b,c \rangle^{8}$

$\langle a,b,b,c \rangle^{4}$

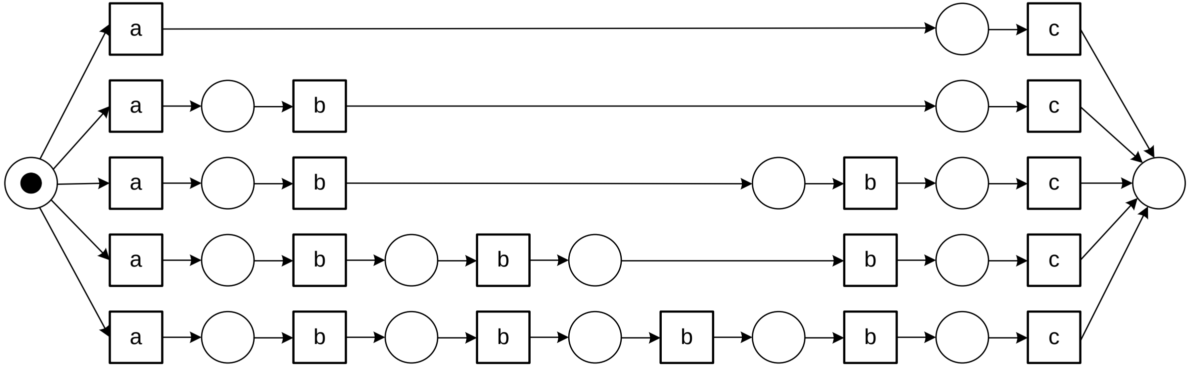
$\langle a,b,b,b,c \rangle^{2}$

$\langle a,b,b,b,b,c \rangle^{1}$



Simplicity: bad!

$\langle a,c\rangle^{16}$
$\langle a,b,c\rangle^{8}$
$\langle a,b,b,c\rangle^{4}$
$\langle a,b,b,b,c\rangle^{2}$
$\langle a,b,b,b,b,c\rangle^{1}$



too complex/specific…

Simplicity: good or bad?

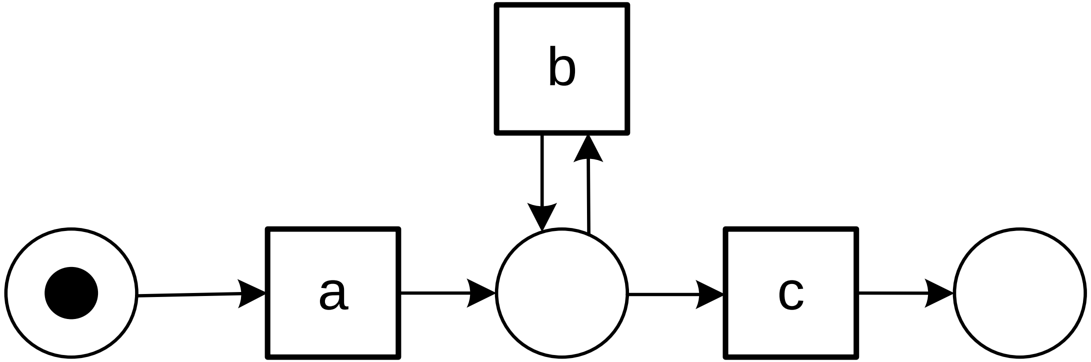
$\langle a,c\rangle^{16}$

$\langle a,b,c\rangle^{8}$

$\langle a,b,b,c\rangle^{4}$

$\langle a,b,b,b,c\rangle^{2}$

$\langle a,b,b,b,b,c\rangle^{1}$



Simplicity: good!

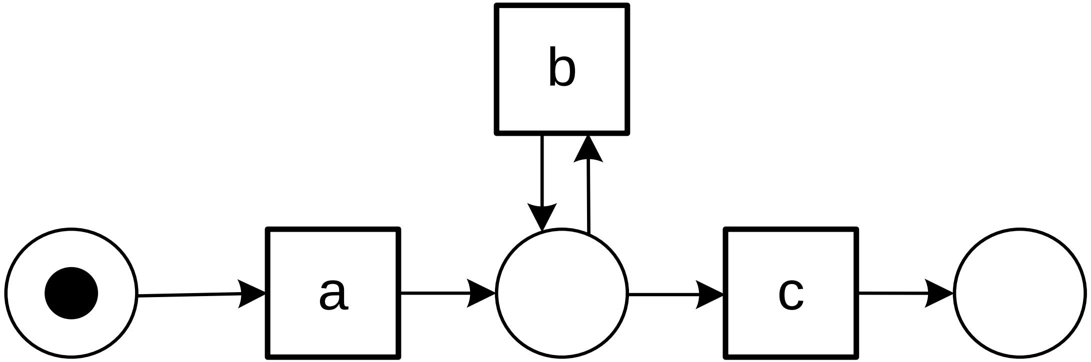
$\langle a,c\rangle^{16}$

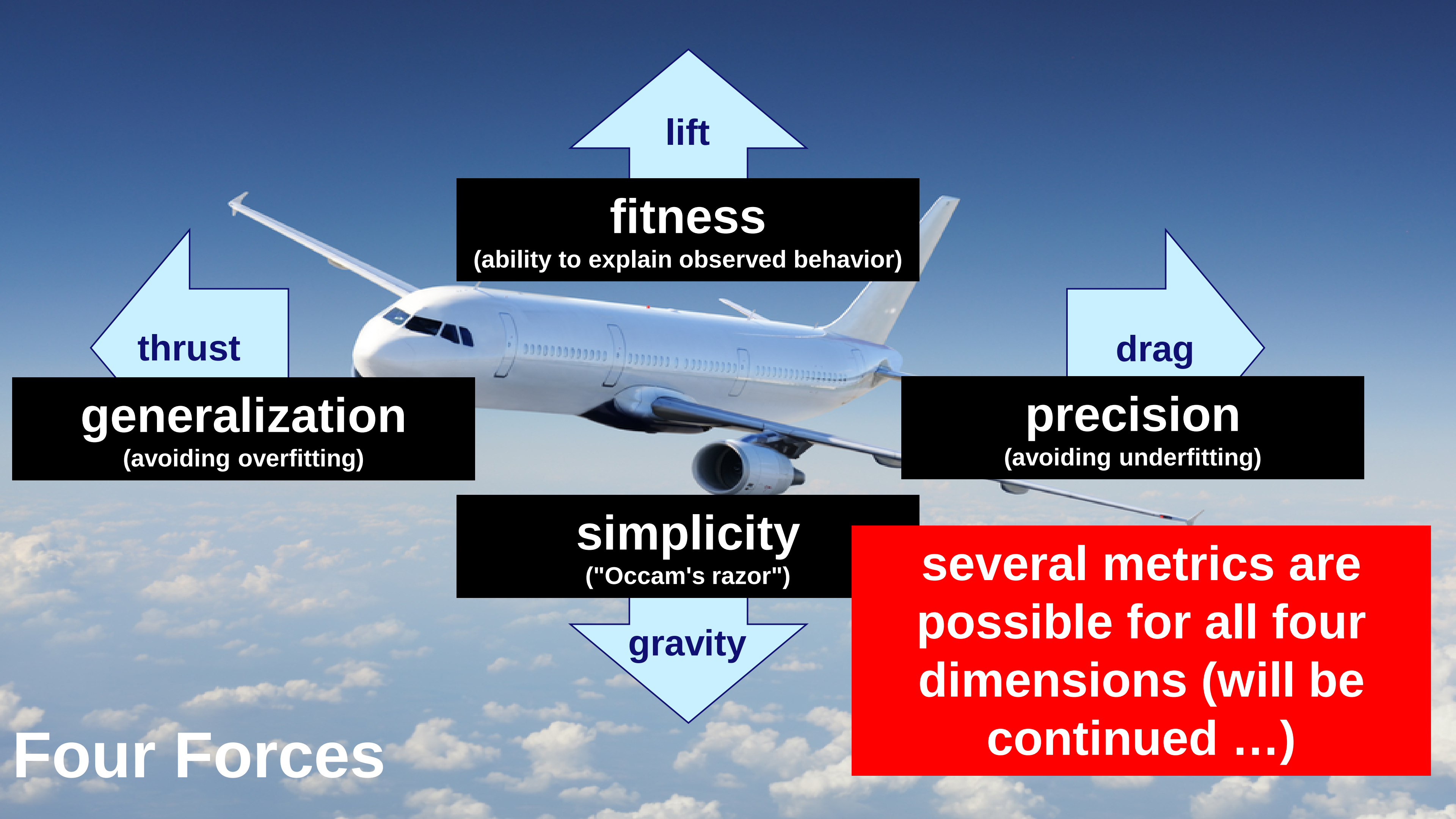
$\langle a,b,c\rangle^{8}$

$\langle a,b,b,c\rangle^{4}$

$\langle a,b,b,b,c\rangle^{2}$

$\langle a,b,b,b,b,c\rangle^{1}$




lift
fitness
(ability to explain observed behavior)
thrust
generalization
(avoiding overfitting)
drag
precision
(avoiding underfitting)
simplicity
("Occam's razor")
gravity
several metrics are possible for all four dimensions (will be continued ...)
Four Forces

modeling language provides a bias

concurrency

cancellation

OR-joins

priorities

duplicate activities

Part I: Preliminaries

Chapter 1
Introduction

Chapter 2
Process Modeling and Analysis

Chapter 3
Data Mining

Part II: From Event Logs to Process Models

Chapter 4
Getting the Data

Chapter 5
Process Discovery: An Introduction

Chapter 6
Advanced Process Discovery Techniques

Part III: Beyond Process Discovery

Chapter 7
Conformance Checking

Chapter 8
Mining Additional Perspectives

Chapter 9
Operational Support

Part IV: Putting Process Mining to Work

Chapter 10
Tool Support

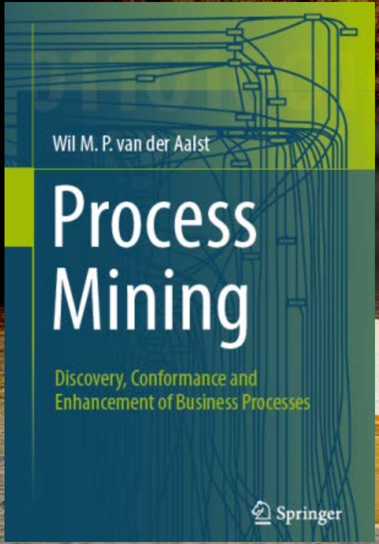
Chapter 11
Analyzing “Lasagna Processes”

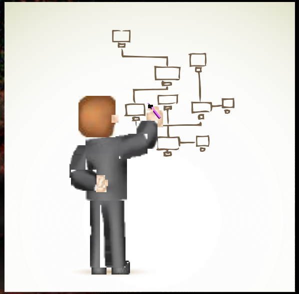
Chapter 12
Analyzing “Spaghetti Processes”

Part V: Reflection

Chapter 13
Cartography and Navigation

Chapter 14
Epilogue





TU/e