Making Process Mining Green

Using Event Data in a Responsible Way

Wil van der Aalst

www.vdaalst.com *@wvdaalst*

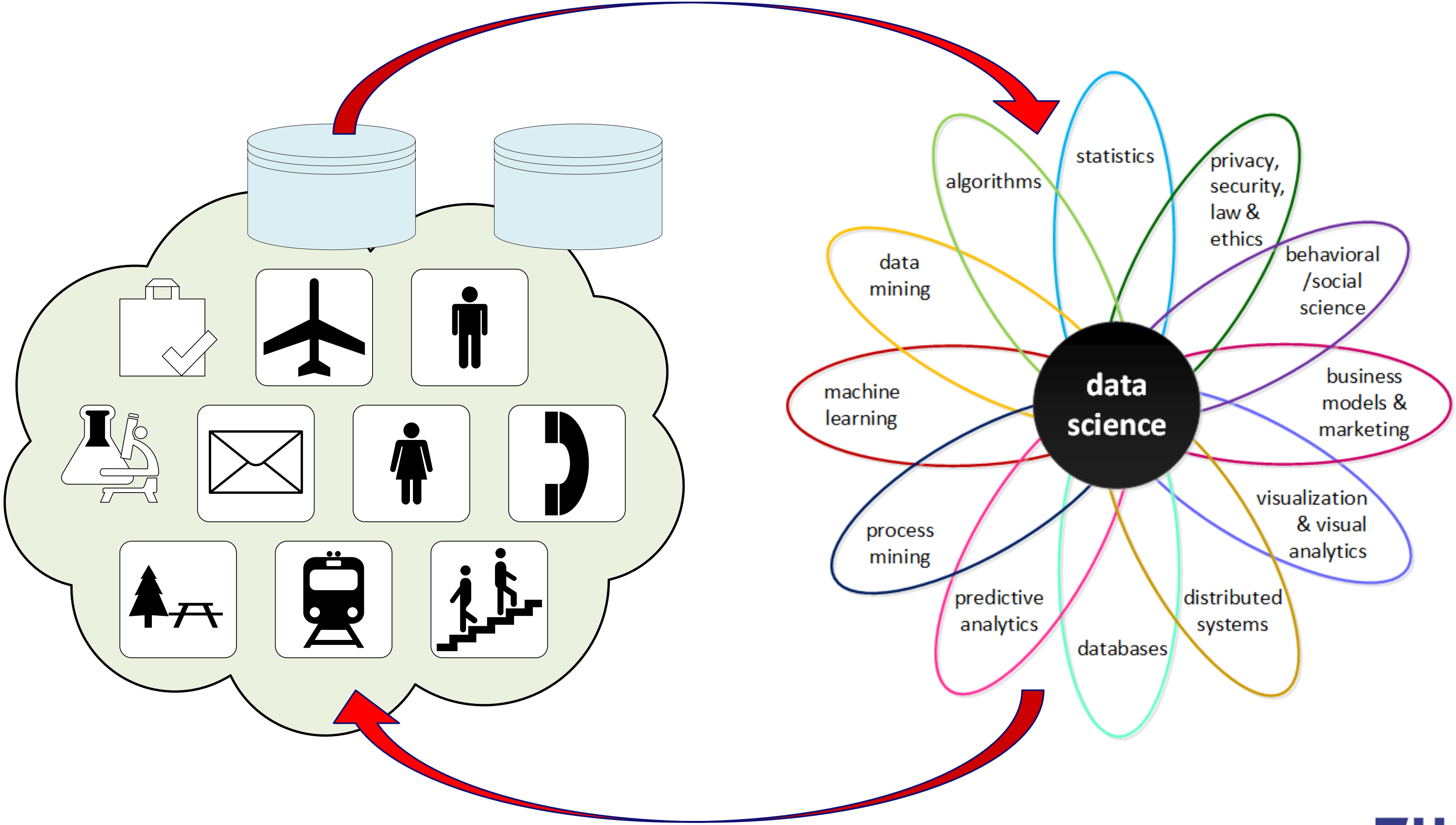
www.processmining.org

ICEIS 2016
18th International Conference on Enterprise Information Systems
Rome, Italy 25 - 28 April, 2016

ENASE 2016
11th International Conference on Evaluation of Novel Approaches to Software Engineering
ROME, ITALY | 27 - 28 APRIL, 2016

TU/e Technische Universiteit Eindhoven University of Technology

Where innovation starts


statistics
algorithms
privacy, security, law & ethics
data mining
behavioral /social science
machine learning
data science
business models & marketing
visualization & visual analytics
process mining
predictive analytics
distributed systems
databases

0101100
1001011
1101110
1011011
0111100
1001011
1101110
1001011
0101110
1001011
0100011
1000001
0101100
1001011
0101111
1001011
0101100

better, faster, more efficient, more effective, cheaper, …

With Great Power Comes Great Responsibility!!

If data is the new oil on which our society runs …

non-transparent

unfair use of data

spurious correlations

privacy violations

bogus conclusions

… then we should take care of data-related forms of pollution!

Green data science: separate the “pollution” from the actual purpose

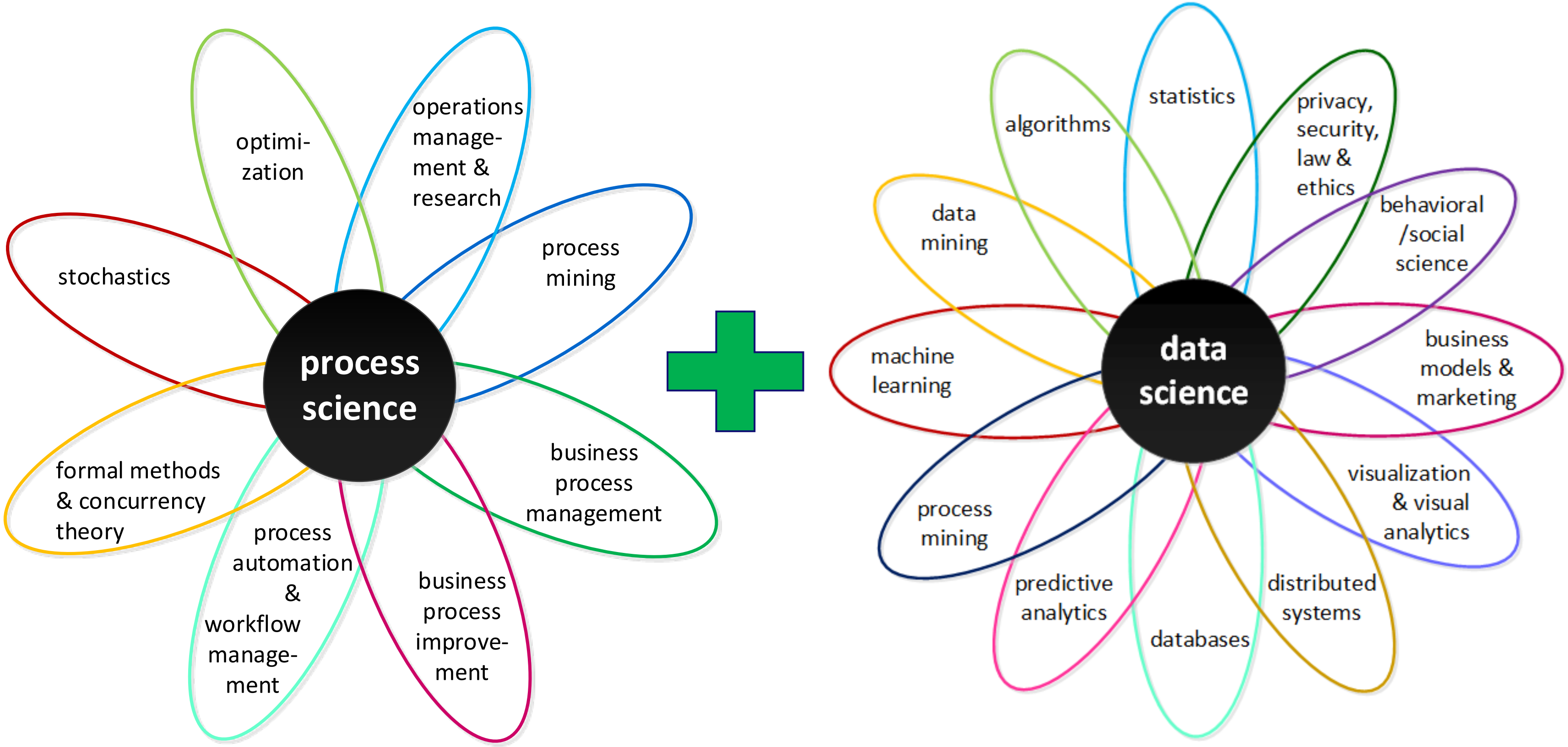
Two parts

2 responsible data science: our next big challenge

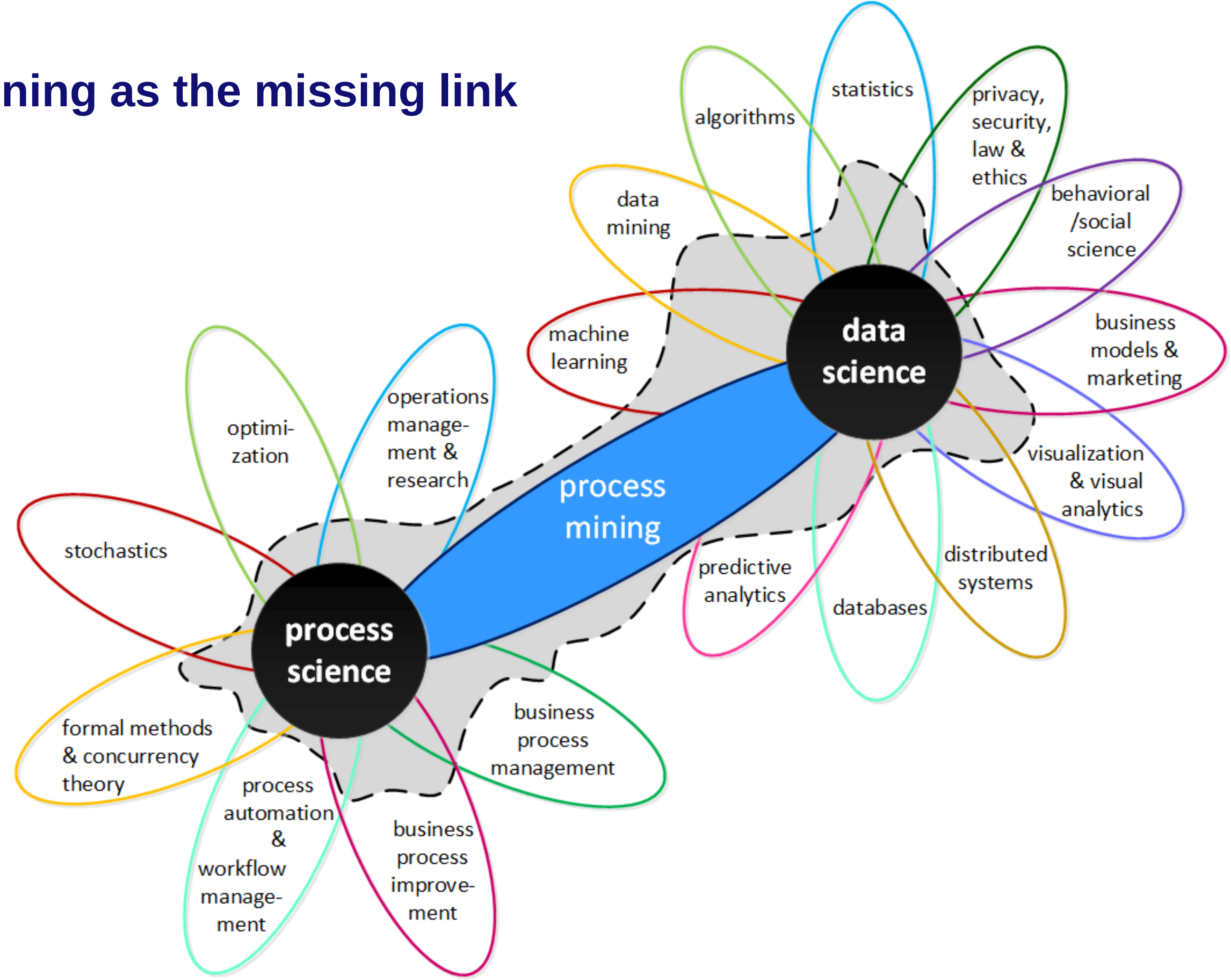
1 process mining: creating value from data

Part I

process mining: creating value from data

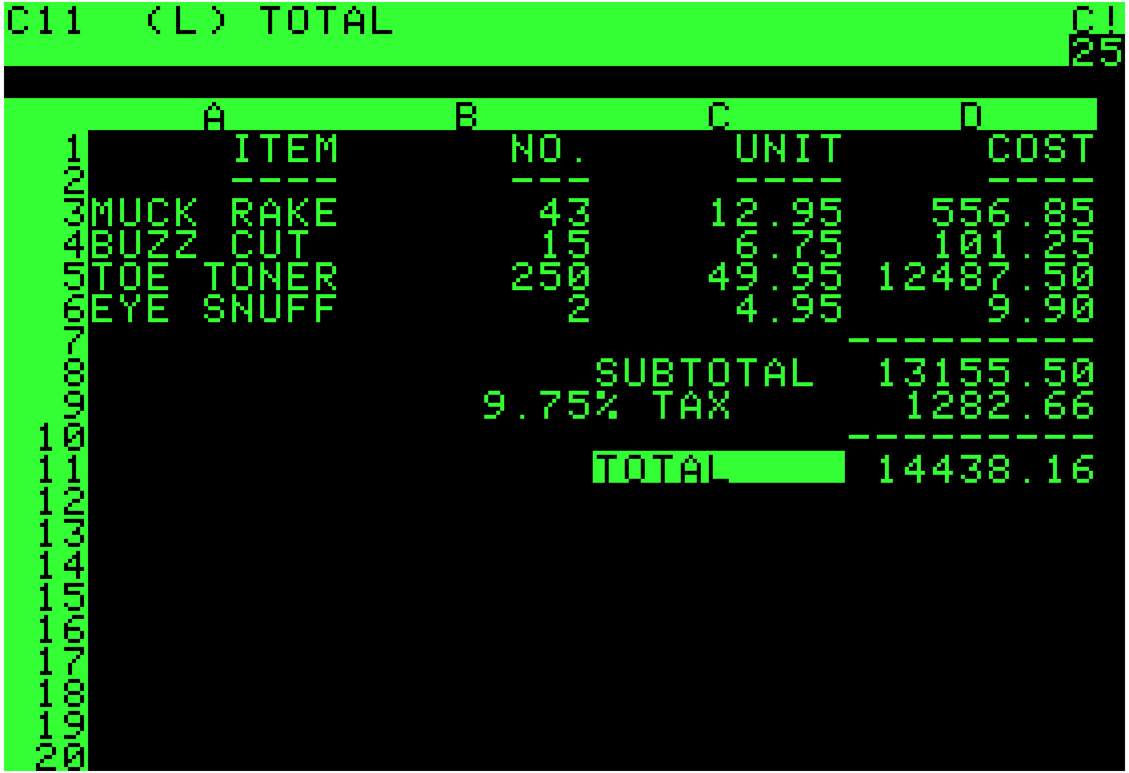

process science
optimi-zation
operations manage-ment & research
stochastics
process mining
formal methods & concurrency theory
business process management
process automation & workflow manage-ment
business process improve-ment
data science
statistics
algorithms
privacy, security, law & ethics
data mining
behavioral /social science
machine learning
business models & marketing
process mining
visualization & visual analytics
predictive analytics
distributed systems
databases

process mining as the missing link


algorithms
statistics
privacy, security, law & ethics
data mining
behavioral /social science
machine learning
data science
business models & marketing
optimi-zation
operations manage-ment & research
visualization & visual analytics
process mining
stochastics
predictive analytics
distributed systems
databases
process science
business process management
formal methods & concurrency theory
process automation & workflow manage-ment
business process improve-ment

Process Mining: On the interface between process science and data science

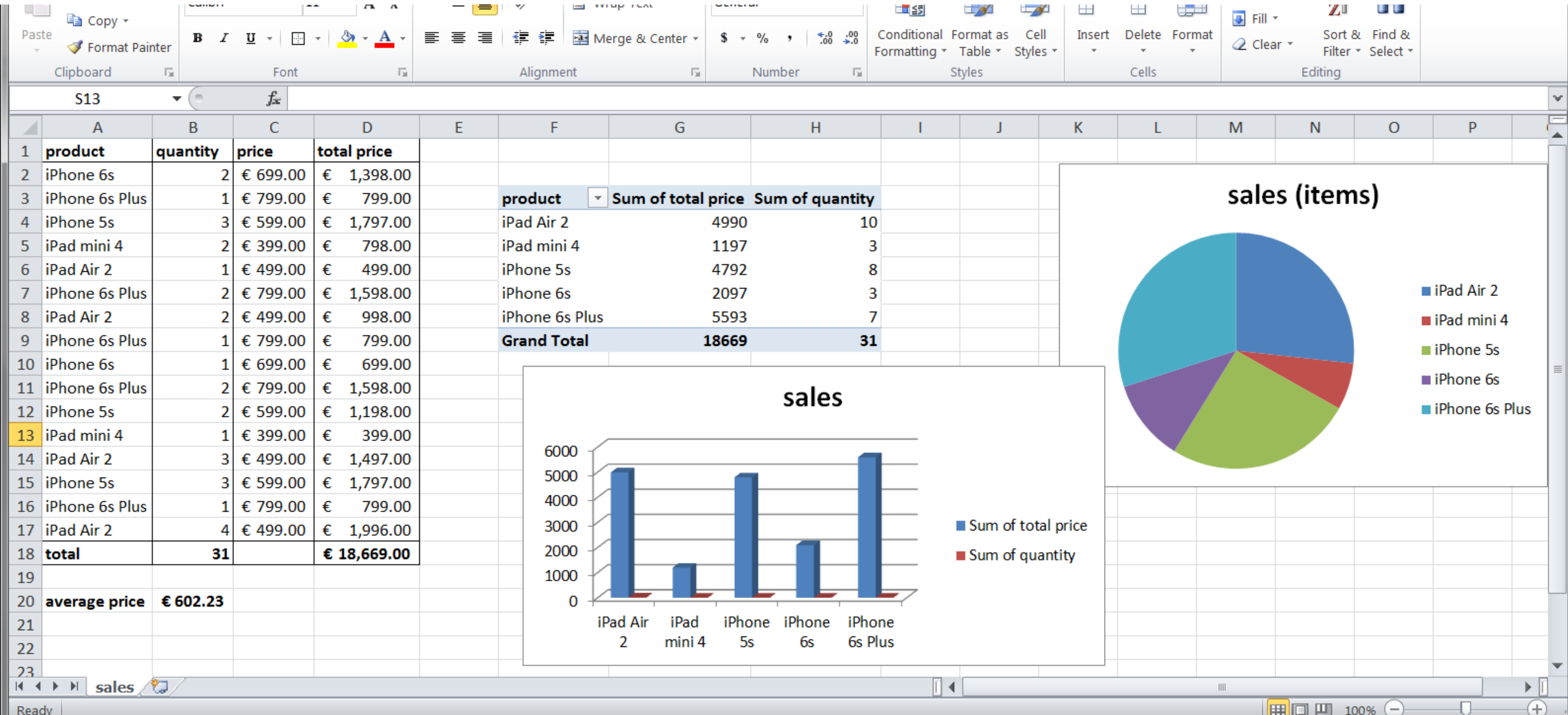
Spreadsheet: Killer App for early computers



```
C11   (L) TOTAL                            C!
                                           25
        A         B         C         D
 1     ITEM      NO.      UNIT      COST
 2     -----     ---      -----     -----
 3MUCK RAKE       43     12.95    556.85
 4BUZZ CUT        15      6.75    101.25
 5TOE TONER      250     49.95  12487.50
 6EYE SNUFF        2      4.95      9.90
 7                              ----------
 8                   SUBTOTAL   13155.50
 9             9.75% TAX         1282.66
10                              ----------
11                   TOTAL      14438.16
12
13
14
15
16
17
18
19
20
```

- **VisiCalc (killer app for Apple II, Oct. 1979)**
- **Lotus 1-2-3 (killer app for IBM PC 1983)**
- **Microsoft Excel (1985)**

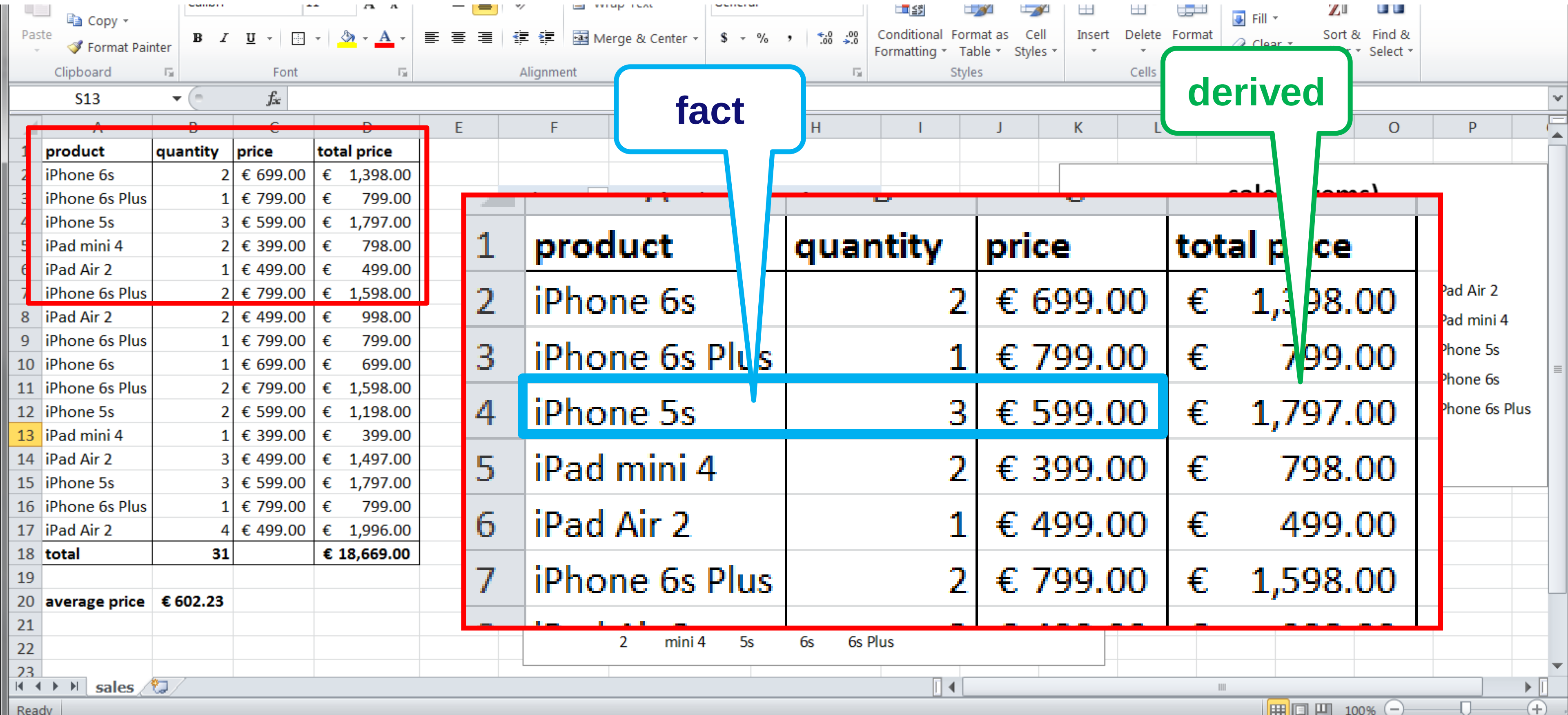
Spreadsheet: Static data



	A	B	C	D
1	**product**	**quantity**	**price**	**total price**
2	iPhone 6s	2	€ 699.00	€ 1,398.00
3	iPhone 6s Plus	1	€ 799.00	€ 799.00
4	iPhone 5s	3	€ 599.00	€ 1,797.00
5	iPad mini 4	2	€ 399.00	€ 798.00
6	iPad Air 2	1	€ 499.00	€ 499.00
7	iPhone 6s Plus	2	€ 799.00	€ 1,598.00
8	iPad Air 2	2	€ 499.00	€ 998.00
9	iPhone 6s Plus	1	€ 799.00	€ 799.00
10	iPhone 6s	1	€ 699.00	€ 699.00
11	iPhone 6s Plus	2	€ 799.00	€ 1,598.00
12	iPhone 5s	2	€ 599.00	€ 1,198.00
13	iPad mini 4	1	€ 399.00	€ 399.00
14	iPad Air 2	3	€ 499.00	€ 1,497.00
15	iPhone 5s	3	€ 599.00	€ 1,797.00
16	iPhone 6s Plus	1	€ 799.00	€ 799.00
17	iPad Air 2	4	€ 499.00	€ 1,996.00
18	**total**	**31**		**€ 18,669.00**
19				
20	**average price**	**€ 602.23**		

product	**Sum of total price**	**Sum of quantity**
iPad Air 2	4990	10
iPad mini 4	1197	3
iPhone 5s	4792	8
iPhone 6s	2097	3
iPhone 6s Plus	5593	7
Grand Total	**18669**	**31**

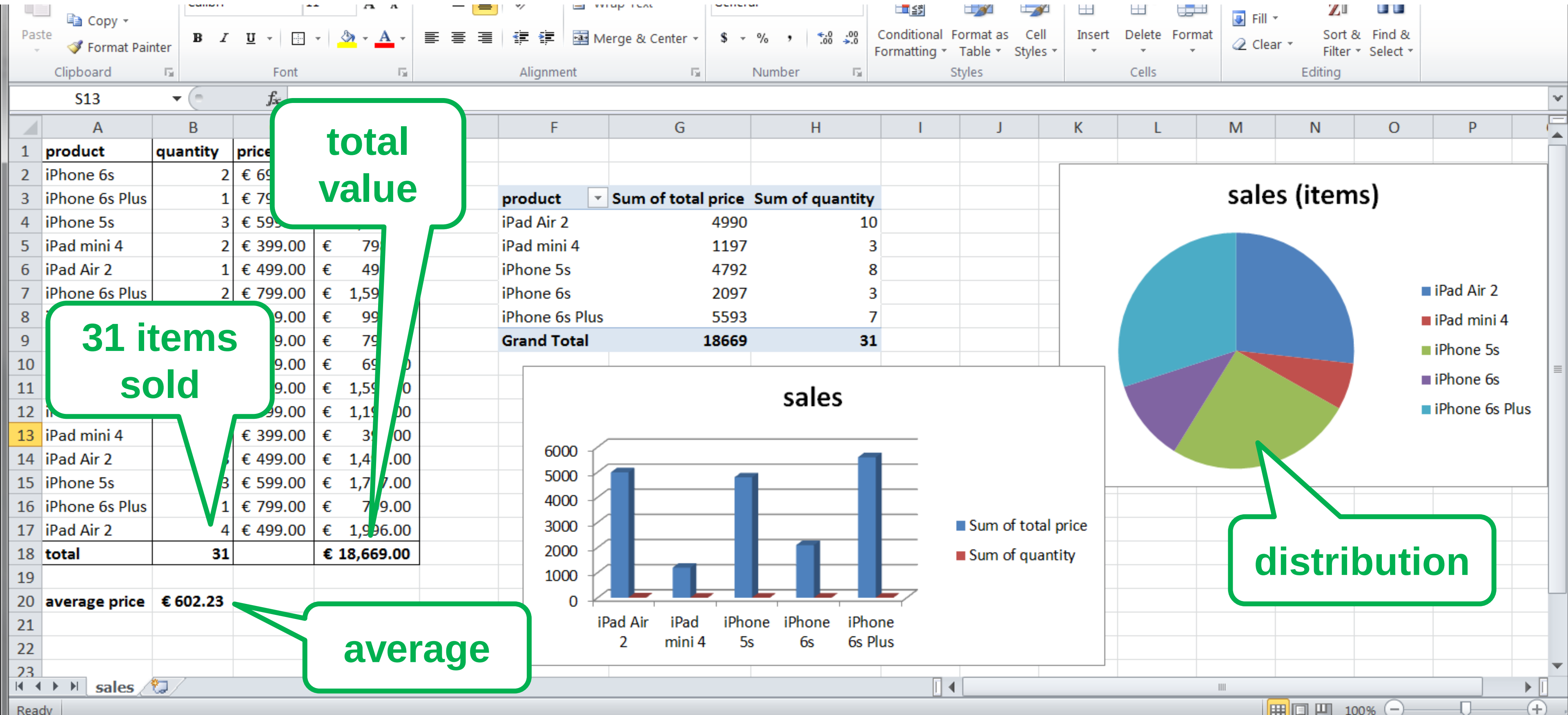
Spreadsheet: Static data



	A	B	C	D
1	**product**	**quantity**	**price**	**total price**
2	iPhone 6s	2	€ 699.00	€ 1,398.00
3	iPhone 6s Plus	1	€ 799.00	€ 799.00
4	iPhone 5s	3	€ 599.00	€ 1,797.00
5	iPad mini 4	2	€ 399.00	€ 798.00
6	iPad Air 2	1	€ 499.00	€ 499.00
7	iPhone 6s Plus	2	€ 799.00	€ 1,598.00
8	iPad Air 2	2	€ 499.00	€ 998.00
9	iPhone 6s Plus	1	€ 799.00	€ 799.00
10	iPhone 6s	1	€ 699.00	€ 699.00
11	iPhone 6s Plus	2	€ 799.00	€ 1,598.00
12	iPhone 5s	2	€ 599.00	€ 1,198.00
13	iPad mini 4	1	€ 399.00	€ 399.00
14	iPad Air 2	3	€ 499.00	€ 1,497.00
15	iPhone 5s	3	€ 599.00	€ 1,797.00
16	iPhone 6s Plus	1	€ 799.00	€ 799.00
17	iPad Air 2	4	€ 499.00	€ 1,996.00
18	**total**	**31**		**€ 18,669.00**
19				
20	**average price**	**€ 602.23**		
21				
22				
23				

1	**product**	**quantity**	**price**	**total price**
2	iPhone 6s	2	€ 699.00	€ 1,398.00
3	iPhone 6s Plus	1	€ 799.00	€ 799.00
4	iPhone 5s	3	€ 599.00	€ 1,797.00
5	iPad mini 4	2	€ 399.00	€ 798.00
6	iPad Air 2	1	€ 499.00	€ 499.00
7	iPhone 6s Plus	2	€ 799.00	€ 1,598.00

Spreadsheet: Static data



	A	B	C	D
1	product	quantity	price	
2	iPhone 6s	2	€ 69	
3	iPhone 6s Plus	1	€ 79	
4	iPhone 5s	3	€ 59	
5	iPad mini 4	2	€ 399.00	€ 79
6	iPad Air 2	1	€ 499.00	€ 49
7	iPhone 6s Plus	2	€ 799.00	€ 1,59
8			9.00	€ 99
9			9.00	€ 79
10			9.00	€ 69 0
11			9.00	€ 1,59 0
12	i		99.00	€ 1,19 00
13	iPad mini 4		€ 399.00	€ 3 00
14	iPad Air 2		€ 499.00	€ 1,4 .00
15	iPhone 5s	3	€ 599.00	€ 1,7 7.00
16	iPhone 6s Plus	1	€ 799.00	€ 7 9.00
17	iPad Air 2	4	€ 499.00	€ 1,996.00
18	total	31		€ 18,669.00
19				
20	average price	€ 602.23		

product	Sum of total price	Sum of quantity
iPad Air 2	4990	10
iPad mini 4	1197	3
iPhone 5s	4792	8
iPhone 6s	2097	3
iPhone 6s Plus	5593	7
Grand Total	18669	31

total value

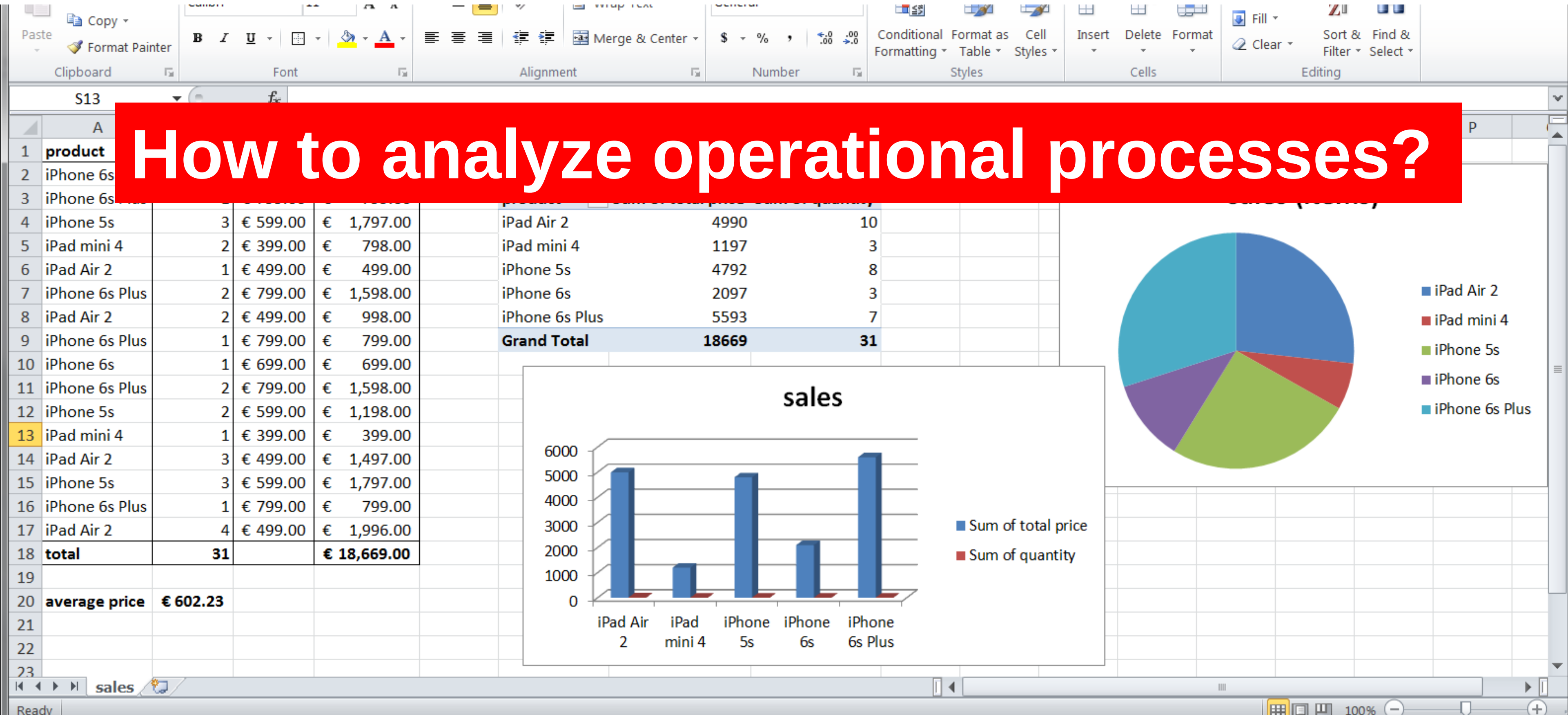
31 items sold

average

distribution

Spreadsheet: Static data

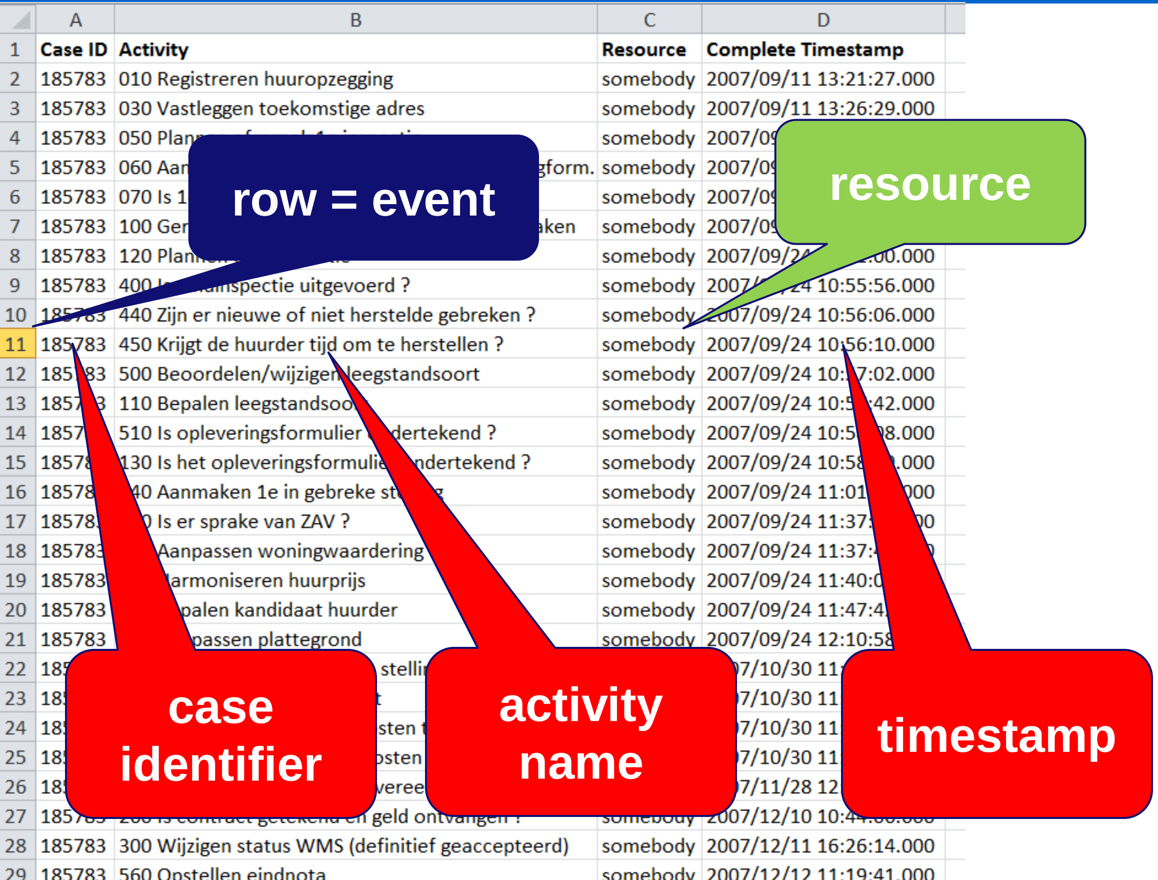
How to analyze operational processes?



	A			
1	product			
2	iPhone 6s	[illegible]	[illegible]	[illegible]
3	iPhone 6s Plus	[illegible]	[illegible]	[illegible]
4	iPhone 5s	3	€ 599.00	€ 1,797.00
5	iPad mini 4	2	€ 399.00	€ 798.00
6	iPad Air 2	1	€ 499.00	€ 499.00
7	iPhone 6s Plus	2	€ 799.00	€ 1,598.00
8	iPad Air 2	2	€ 499.00	€ 998.00
9	iPhone 6s Plus	1	€ 799.00	€ 799.00
10	iPhone 6s	1	€ 699.00	€ 699.00
11	iPhone 6s Plus	2	€ 799.00	€ 1,598.00
12	iPhone 5s	2	€ 599.00	€ 1,198.00
13	iPad mini 4	1	€ 399.00	€ 399.00
14	iPad Air 2	3	€ 499.00	€ 1,497.00
15	iPhone 5s	3	€ 599.00	€ 1,797.00
16	iPhone 6s Plus	1	€ 799.00	€ 799.00
17	iPad Air 2	4	€ 499.00	€ 1,996.00
18	**total**	**31**		**€ 18,669.00**
19				
20	**average price**	**€ 602.23**		
21				
22				
23				

product	Sum of total price	Sum of quantity
iPad Air 2	4990	10
iPad mini 4	1197	3
iPhone 5s	4792	8
iPhone 6s	2097	3
iPhone 6s Plus	5593	7
Grand Total	**18669**	**31**

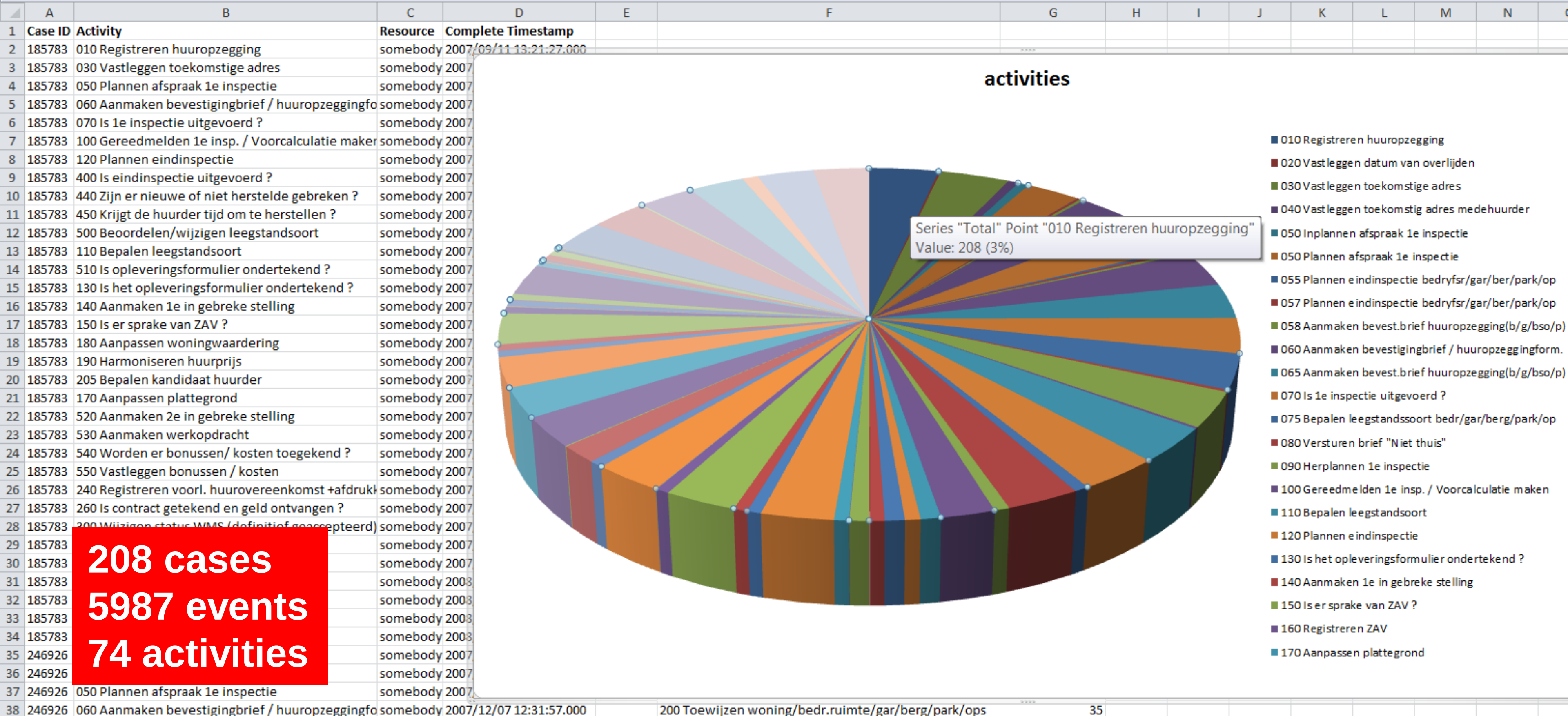
Process Mining: Spreadsheet for behavior



	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
3	185783	030 Vastleggen toekomstige adres	somebody	2007/09/11 13:26:29.000
4	185783	050 Plan	somebody	2007/09
5	185783	060 Aan gform.	somebody	2007/0
6	185783	070 Is 1	somebody	2007/0
7	185783	100 Ger aken	somebody	2007/0
8	185783	120 Plan	somebody	2007/09/2 .00.000
9	185783	400 inspectie uitgevoerd ?	somebody	2007/ 24 10:55:56.000
10	185783	440 Zijn er nieuwe of niet herstelde gebreken ?	somebody	2007/09/24 10:56:06.000
11	185783	450 Krijgt de huurder tijd om te herstellen ?	somebody	2007/09/24 10:56:10.000
12	185 83	500 Beoordelen/wijzigen leegstandsoort	somebody	2007/09/24 10: 7:02.000
13	1857 3	110 Bepalen leegstandsoo	somebody	2007/09/24 10:5 :42.000
14	1857	510 Is opleveringsformulier dertekend ?	somebody	2007/09/24 10:5 8.000
15	18578	130 Is het opleveringsformulie ndertekend ?	somebody	2007/09/24 10:58 .000
16	18578	40 Aanmaken 1e in gebreke st g	somebody	2007/09/24 11:01 000
17	18578	0 Is er sprake van ZAV ?	somebody	2007/09/24 11:37: 00
18	185783	Aanpassen woningwaardering	somebody	2007/09/24 11:37:
19	185783	larmoniseren huurprijs	somebody	2007/09/24 11:40:0
20	185783	palen kandidaat huurder	somebody	2007/09/24 11:47:4
21	185783	passen plattegrond	somebody	2007/09/24 12:10:58
22	18	stelli		7/10/30 11
23	18	t		7/10/30 11
24	18	sten t		7/10/30 11
25	18	osten		7/10/30 11
26	18	veree		7/11/28 12
27	185	geld ontvangen ?		2007/12/10 10:4
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

- Input: events (“things that have happened”)
- Mandatory per event:
  - case identifier
  - activity name
  - timestamp/date
- Optional
  - resource
  - transaction type
  - costs
  - …

Process Mining: Spreadsheet for behavior

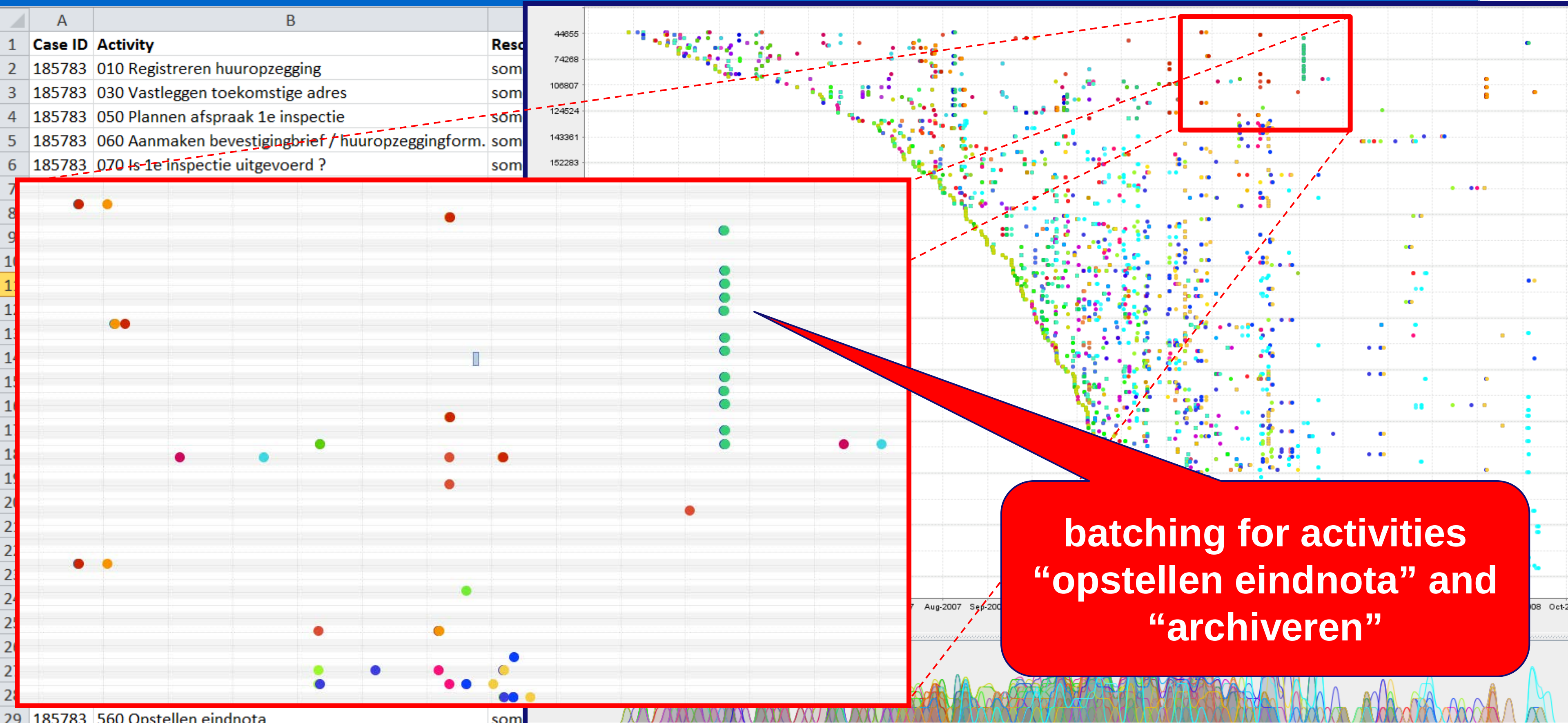


	A	B	C	D	E	F	G
1	Case ID	Activity	Resource	Complete Timestamp			
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000			
3	185783	030 Vastleggen toekomstige adres	somebody	2007			
4	185783	050 Plannen afspraak 1e inspectie	somebody	2007			
5	185783	060 Aanmaken bevestigingbrief / huuropzeggingfo	somebody	2007			
6	185783	070 Is 1e inspectie uitgevoerd ?	somebody	2007			
7	185783	100 Gereedmelden 1e insp. / Voorcalculatie maker	somebody	2007			
8	185783	120 Plannen eindinspectie	somebody	2007			
9	185783	400 Is eindinspectie uitgevoerd ?	somebody	2007			
10	185783	440 Zijn er nieuwe of niet herstelde gebreken ?	somebody	2007			
11	185783	450 Krijgt de huurder tijd om te herstellen ?	somebody	2007			
12	185783	500 Beoordelen/wijzigen leegstandsoort	somebody	2007			
13	185783	110 Bepalen leegstandsoort	somebody	2007			
14	185783	510 Is opleveringsformulier ondertekend ?	somebody	2007			
15	185783	130 Is het opleveringsformulier ondertekend ?	somebody	2007			
16	185783	140 Aanmaken 1e in gebreke stelling	somebody	2007			
17	185783	150 Is er sprake van ZAV ?	somebody	2007			
18	185783	180 Aanpassen woningwaardering	somebody	2007			
19	185783	190 Harmoniseren huurprijs	somebody	2007			
20	185783	205 Bepalen kandidaat huurder	somebody	2007			
21	185783	170 Aanpassen plattegrond	somebody	2007			
22	185783	520 Aanmaken 2e in gebreke stelling	somebody	2007			
23	185783	530 Aanmaken werkopdracht	somebody	2007			
24	185783	540 Worden er bonussen/ kosten toegekend ?	somebody	2007			
25	185783	550 Vastleggen bonussen / kosten	somebody	2007			
26	185783	240 Registreren voorl. huurovereenkomst +afdrukk	somebody	2007			
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007			
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007			
29	185783		somebody	2007			
30	185783		somebody	2007			
31	185783		somebody	2008			
32	185783		somebody	2008			
33	185783		somebody	2008			
34	185783		somebody	2008			
35	246926		somebody	2007			
36	246926		somebody	2007			
37	246926	050 Plannen afspraak 1e inspectie	somebody	2007			
38	246926	060 Aanmaken bevestigingbrief / huuropzeggingfo	somebody	2007/12/07 12:31:57.000		200 Toewijzen woning/bedr.ruimte/gar/berg/park/ops	35

activities
Series "Total" Point "010 Registreren huuropzegging"
Value: 208 (3%)
010 Registreren huuropzegging
020 Vastleggen datum van overlijden
030 Vastleggen toekomstige adres
040 Vastleggen toekomstig adres medehuurder
050 Inplannen afspraak 1e inspectie
050 Plannen afspraak 1e inspectie
055 Plannen eindinspectie bedryfsr/gar/ber/park/op
057 Plannen eindinspectie bedryfsr/gar/ber/park/op
058 Aanmaken bevest.brief huuropzegging(b/g/bso/p)
060 Aanmaken bevestigingbrief / huuropzeggingform.
065 Aanmaken bevest.brief huuropzegging(b/g/bso/p)
070 Is 1e inspectie uitgevoerd ?
075 Bepalen leegstandssoort bedr/gar/berg/park/op
080 Versturen brief "Niet thuis"
090 Herplannen 1e inspectie
100 Gereedmelden 1e insp. / Voorcalculatie maken
110 Bepalen leegstandsoort
120 Plannen eindinspectie
130 Is het opleveringsformulier ondertekend ?
140 Aanmaken 1e in gebreke stelling
150 Is er sprake van ZAV ?
160 Registreren ZAV
170 Aanpassen plattegrond

208 cases
5987 events
74 activities

Process Mining: Spreadsheet for behavior



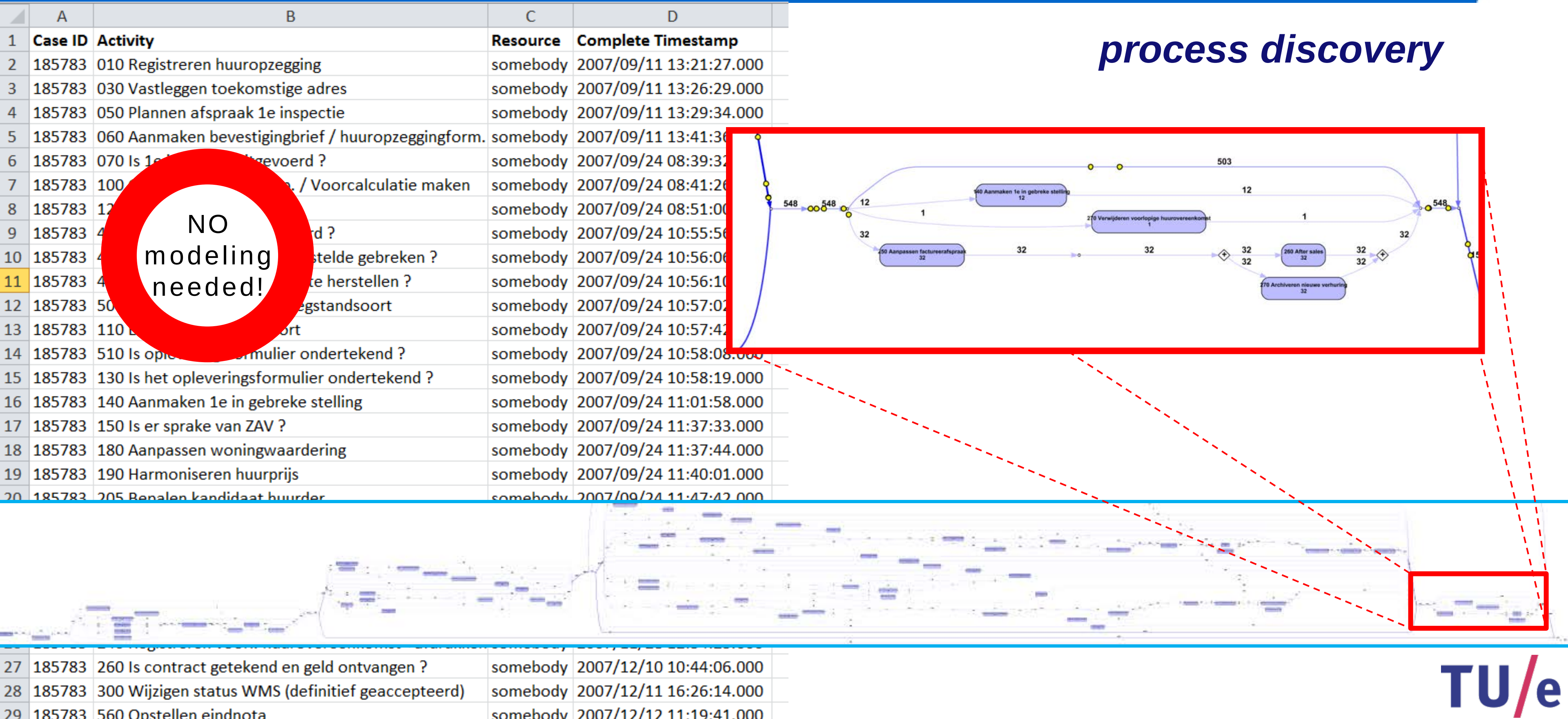
	A	B	
1	Case ID	Activity	Reso
2	185783	010 Registreren huuropzegging	som
3	185783	030 Vastleggen toekomstige adres	som
4	185783	050 Plannen afspraak 1e inspectie	som
5	185783	060 Aanmaken bevestigingbrief / huuropzeggingform.	som
6	185783	070 Is 1e inspectie uitgevoerd ?	som
29	185783	560 Opstellen eindnota	som

batching for activities “opstellen eindnota” and “archiveren”

Process Discovery

Process Mining: Spreadsheet for behavior

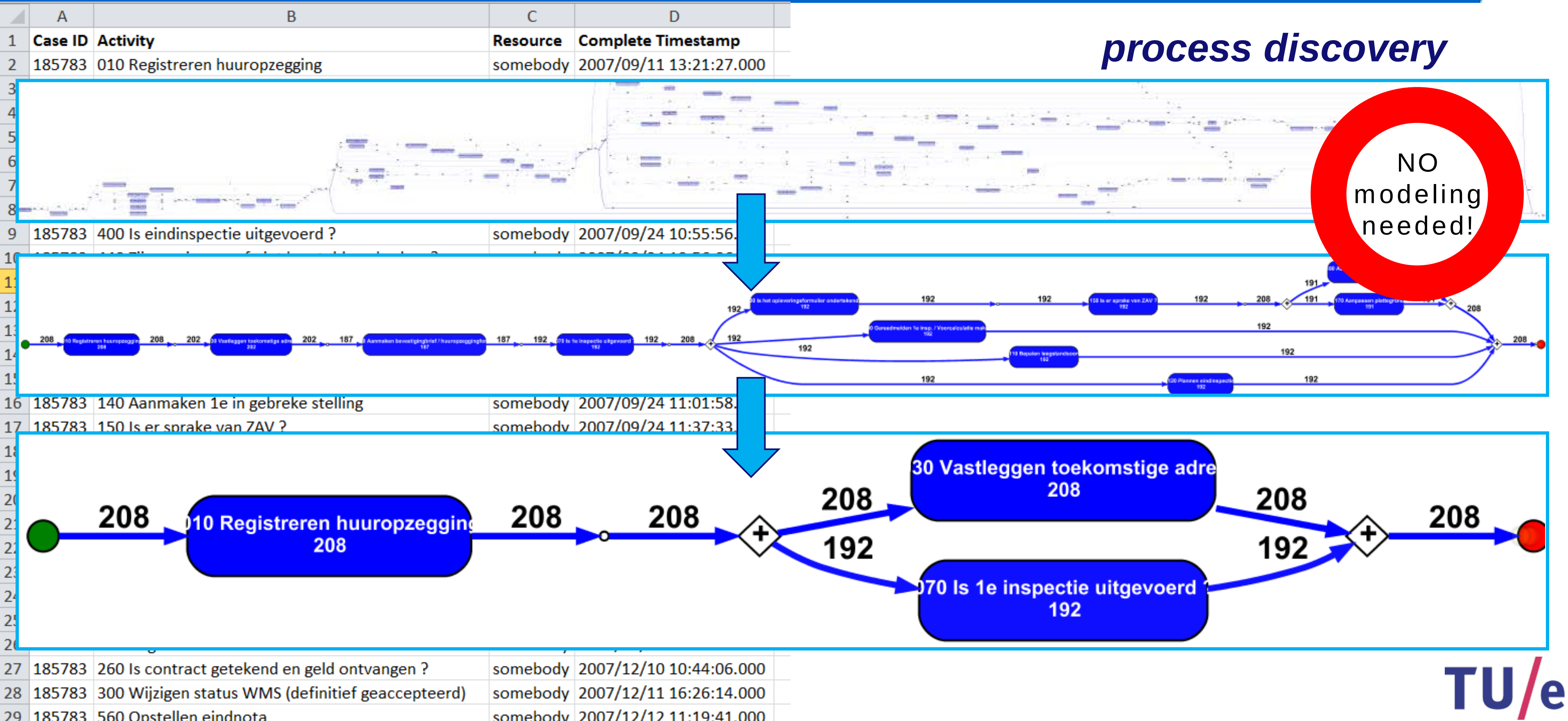
process discovery



NO modeling needed!
503
548
548
12
12
1
1
32
32
32
32
32
32
32
32
548
140 Aanmaken 1e in gebreke stelling
12
270 Verwijderen voorlopige huurovereenkomst
1
250 Aanpassen factureerafspraak
32
260 After sales
32
270 Archiveren nieuwe verhuring
32

	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
3	185783	030 Vastleggen toekomstige adres	somebody	2007/09/11 13:26:29.000
4	185783	050 Plannen afspraak 1e inspectie	somebody	2007/09/11 13:29:34.000
5	185783	060 Aanmaken bevestigingbrief / huuropzeggingform.	somebody	2007/09/11 13:41:36
6	185783	070 Is 1e itgevoerd ?	somebody	2007/09/24 08:39:32
7	185783	100 . / Voorcalculatie maken	somebody	2007/09/24 08:41:26
8	185783	12	somebody	2007/09/24 08:51:00
9	185783	4 rd ?	somebody	2007/09/24 10:55:56
10	185783	4 stelde gebreken ?	somebody	2007/09/24 10:56:06
11	185783	4 te herstellen ?	somebody	2007/09/24 10:56:10
12	185783	50 egstandsoort	somebody	2007/09/24 10:57:02
13	185783	110 L ort	somebody	2007/09/24 10:57:42
14	185783	510 Is ople ormulier ondertekend ?	somebody	2007/09/24 10:58:08.000
15	185783	130 Is het opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:19.000
16	185783	140 Aanmaken 1e in gebreke stelling	somebody	2007/09/24 11:01:58.000
17	185783	150 Is er sprake van ZAV ?	somebody	2007/09/24 11:37:33.000
18	185783	180 Aanpassen woningwaardering	somebody	2007/09/24 11:37:44.000
19	185783	190 Harmoniseren huurprijs	somebody	2007/09/24 11:40:01.000
20	185783	205 Benalen kandidaat huurder	somebody	2007/09/24 11:47:42.000
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

Process Mining: Spreadsheet for behavior

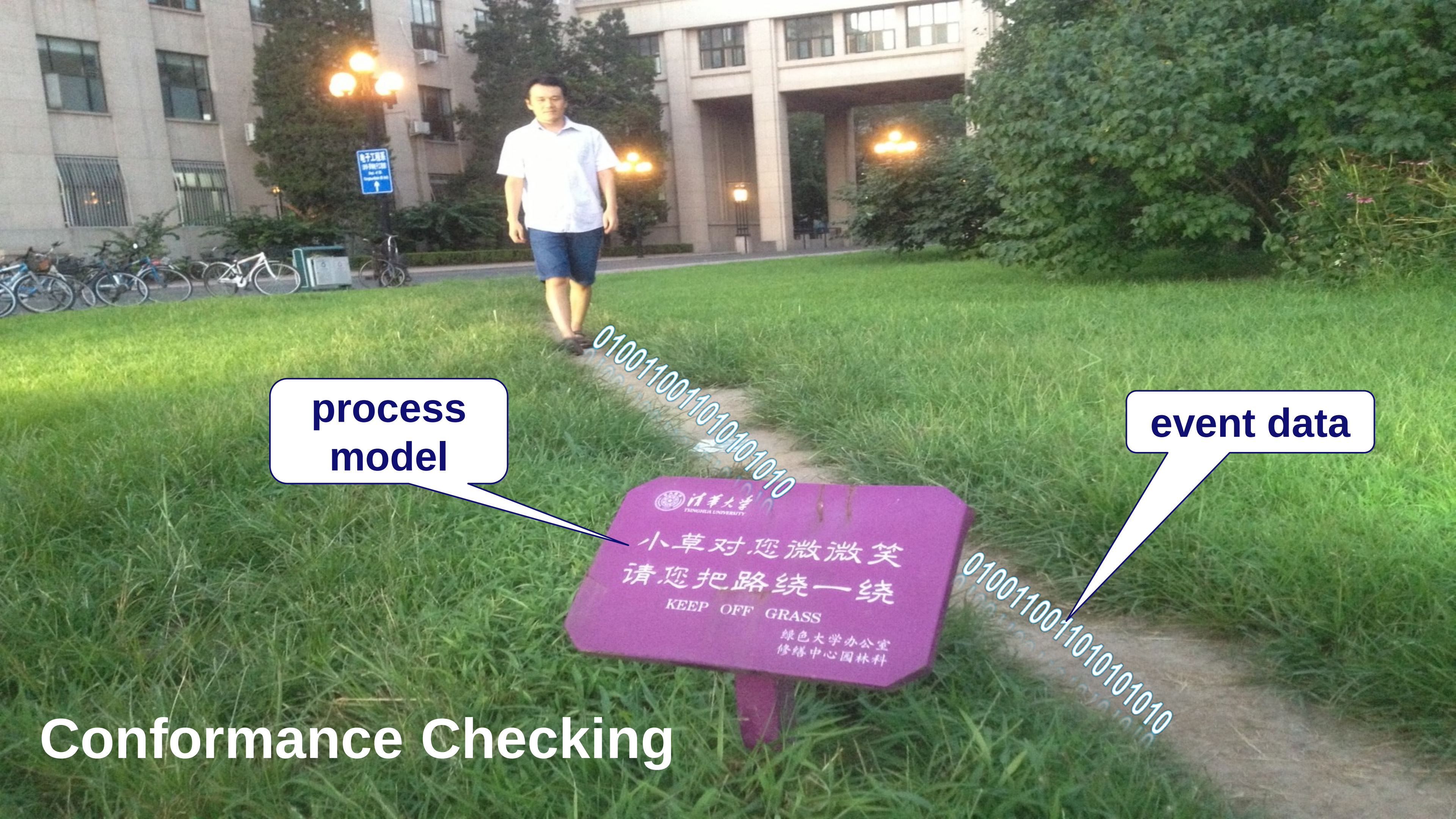


	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
9	185783	400 Is eindinspectie uitgevoerd ?	somebody	2007/09/24 10:55:56.
16	185783	140 Aanmaken 1e in gebreke stelling	somebody	2007/09/24 11:01:58.
17	185783	150 Is er sprake van ZAV ?	somebody	2007/09/24 11:37:33.
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

208
10 Registreren huuropzeggin 208
202
30 Vastleggen toekomstige adre 202
187
Aanmaken bevestigingbrief / huuropzeggingfor 187
192
070 Is 1e inspectie uitgevoerd 192
Is het opleveringsformulier ondertekend 192
Gereedmelden 1e insp. / Voorcalculatie ma 192
110 Bepalen leegstandsoor 192
120 Plannen eindinspectie 192
150 Is er sprake van ZAV ? 192
170 Aanpassen plattegrond 191
191
208
010 Registreren huuropzegging 208
208
208
208
30 Vastleggen toekomstige adre 208
192
070 Is 1e inspectie uitgevoerd 192
208
192
208

process discovery

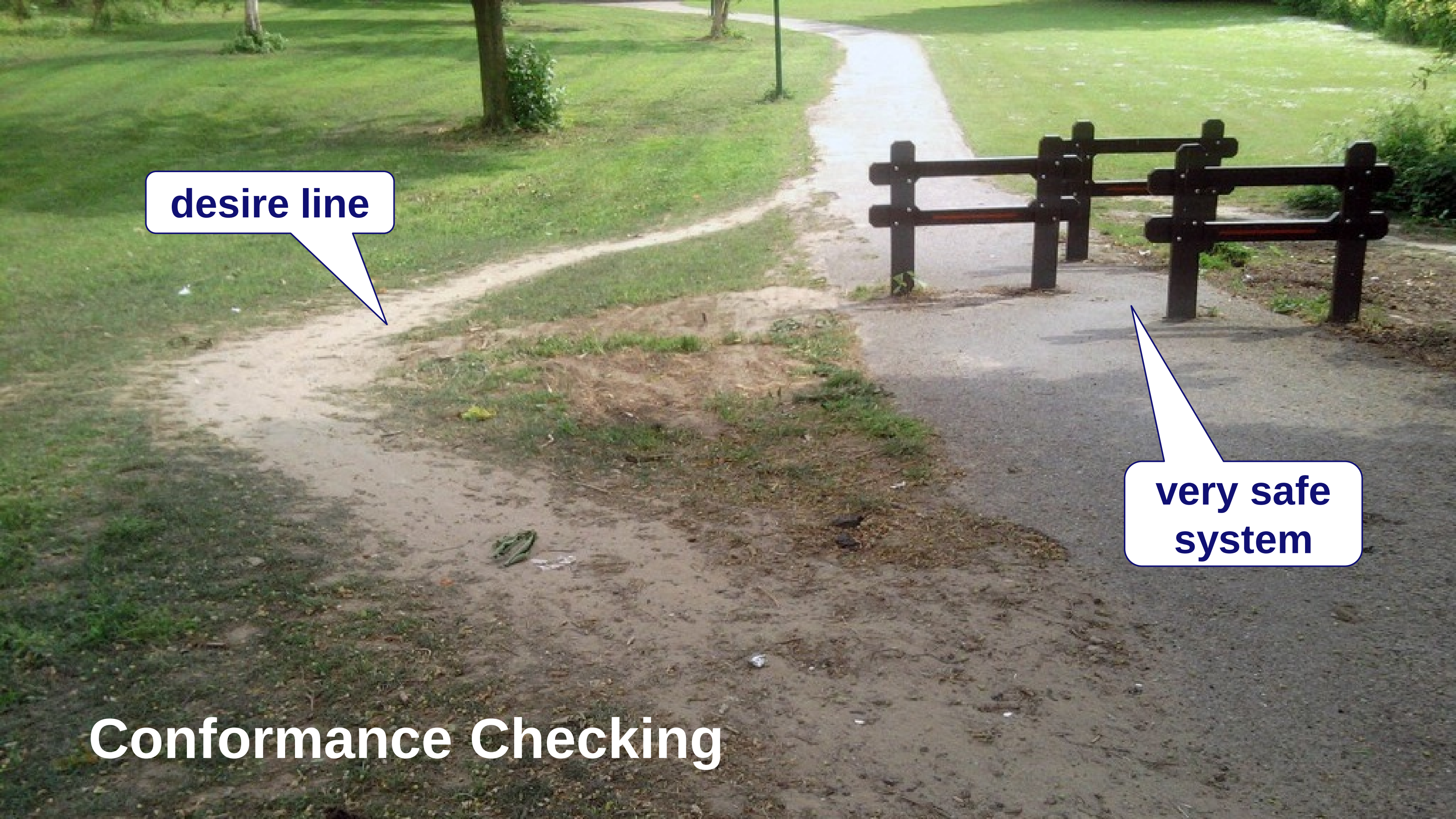
NO modeling needed!



process model

event data

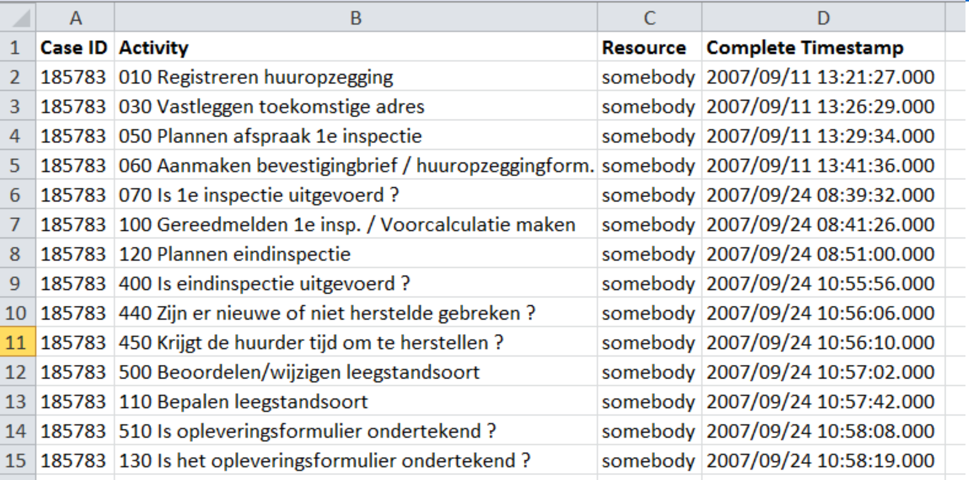
Conformance Checking



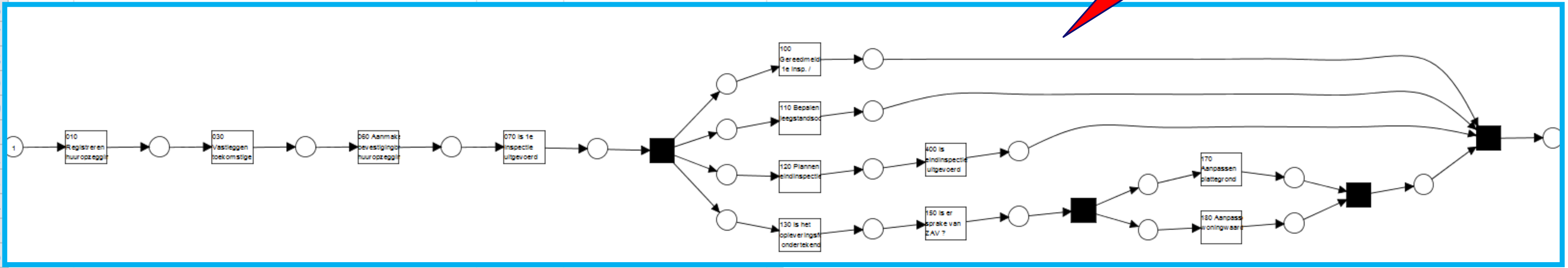
Conformance Checking

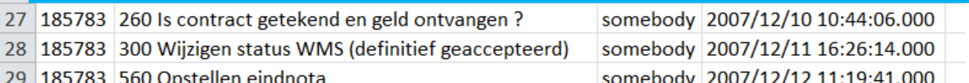
Process Mining: Spreadsheet for behavior

conformance checking



discovered or hand-made

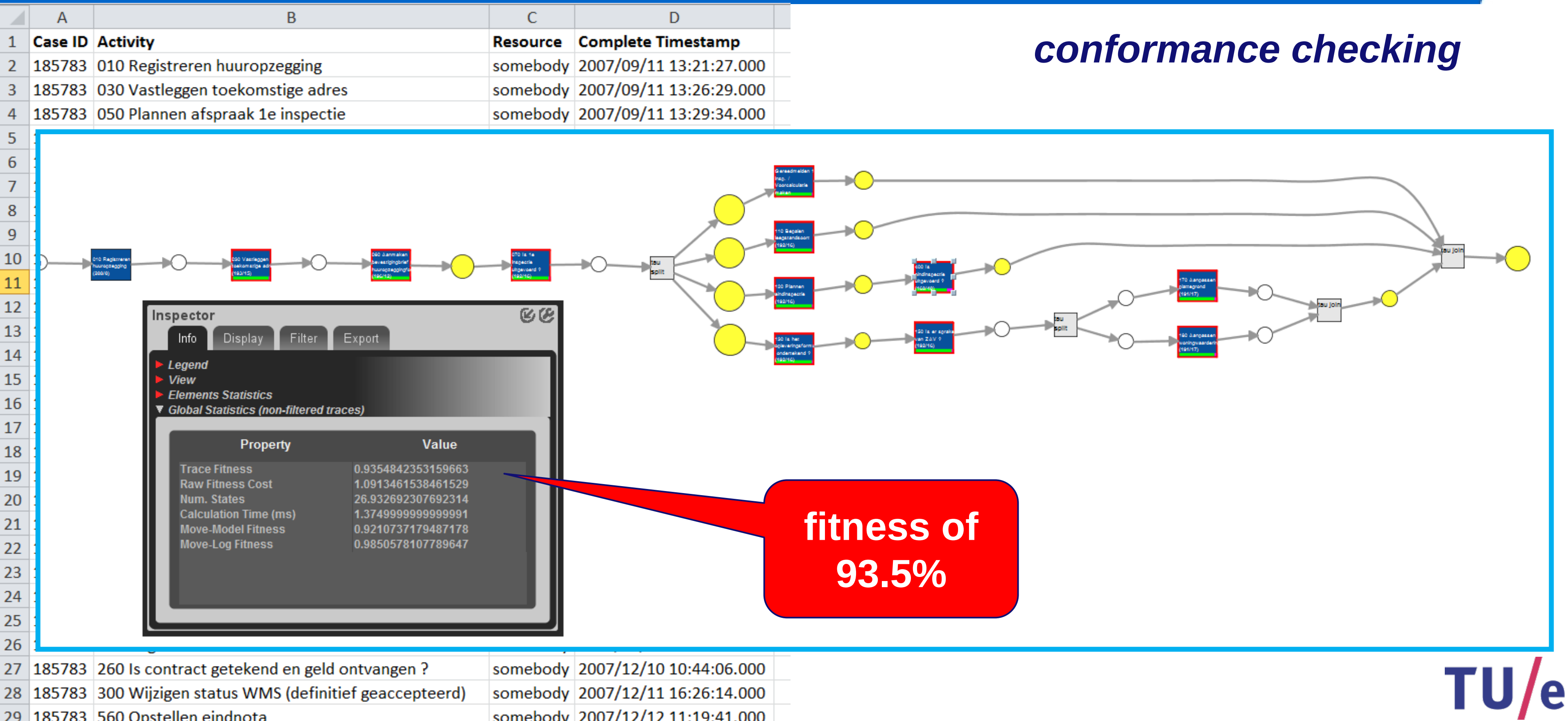

1
010 Registreren huuropzeggir
030 Vastleggen toekomstige
060 Aanmake bevestigingb huuropzeggir
070 Is 1e inspectie uitgevoerd
100 Gereedmeld 1e insp. /
110 Bepalen leegstandsoc
120 Plannen eindinspectie
130 Is het opleveringsf ondertekend
400 Is eindinspectie uitgevoerd
150 Is er sprake van ZAV ?
170 Aanpassen plattegrond
180 Aanpass woningwaard



	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
3	185783	030 Vastleggen toekomstige adres	somebody	2007/09/11 13:26:29.000
4	185783	050 Plannen afspraak 1e inspectie	somebody	2007/09/11 13:29:34.000
5	185783	060 Aanmaken bevestigingbrief / huuropzeggingform.	somebody	2007/09/11 13:41:36.000
6	185783	070 Is 1e inspectie uitgevoerd ?	somebody	2007/09/24 08:39:32.000
7	185783	100 Gereedmelden 1e insp. / Voorcalculatie maken	somebody	2007/09/24 08:41:26.000
8	185783	120 Plannen eindinspectie	somebody	2007/09/24 08:51:00.000
9	185783	400 Is eindinspectie uitgevoerd ?	somebody	2007/09/24 10:55:56.000
10	185783	440 Zijn er nieuwe of niet herstelde gebreken ?	somebody	2007/09/24 10:56:06.000
11	185783	450 Krijgt de huurder tijd om te herstellen ?	somebody	2007/09/24 10:56:10.000
12	185783	500 Beoordelen/wijzigen leegstandsoort	somebody	2007/09/24 10:57:02.000
13	185783	110 Bepalen leegstandsoort	somebody	2007/09/24 10:57:42.000
14	185783	510 Is opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:08.000
15	185783	130 Is het opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:19.000
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

Process Mining: Spreadsheet for behavior

conformance checking



	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
3	185783	030 Vastleggen toekomstige adres	somebody	2007/09/11 13:26:29.000
4	185783	050 Plannen afspraak 1e inspectie	somebody	2007/09/11 13:29:34.000
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				
21				
22				
23				
24				
25				
26				
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

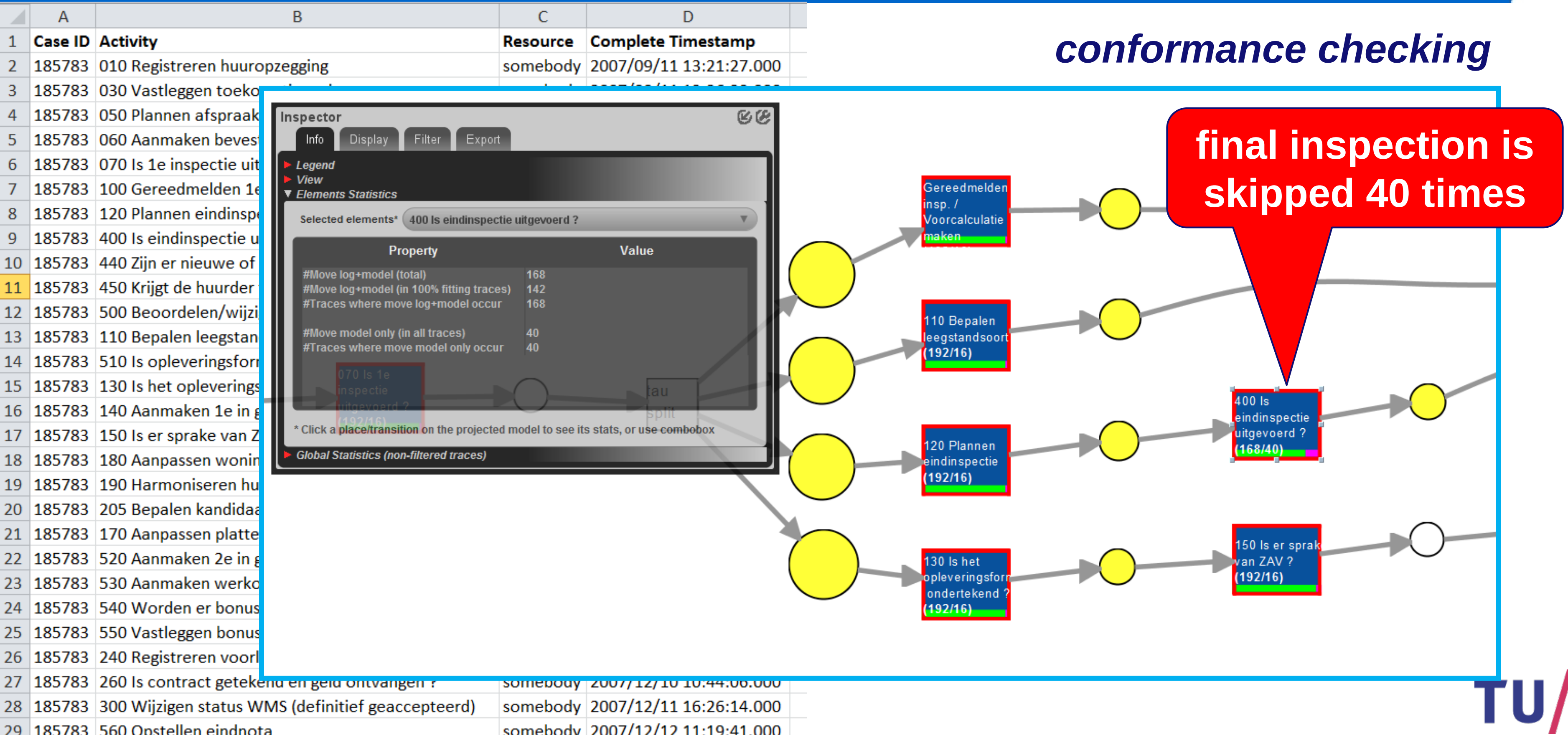
010 Registreren huuropzegging (208/0)
030 Vastleggen toekomstige adr (193/15)
050 Aanmaken bevestigingbrief huuropzeggingfo (196/12)
070 Is 1e inspectie uitgevoerd ? (192/16)
tau split
Gereedmelden 1 insp. / Voorcalculatie maken
110 Bepalen leegstandsoort (192/16)
120 Plannen eindinspectie (192/16)
130 Is het opleveringsform ondertekend ? (192/16)
400 Is eindinspectie uitgevoerd ?
150 Is er sprake van ZAV ? (192/16)
tau split
170 Aangassen plattegrond (191/17)
180 Aangassen woningwaardering (191/17)
tau join
tau join
Inspector
Info
Display
Filter
Export
Legend
View
Elements Statistics
Global Statistics (non-filtered traces)

Property	Value
Trace Fitness	0.9354842353159663
Raw Fitness Cost	1.0913461538461529
Num. States	26.932692307692314
Calculation Time (ms)	1.3749999999999991
Move-Model Fitness	0.9210737179487178
Move-Log Fitness	0.9850578107789647

fitness of 93.5%

Process Mining: Spreadsheet for behavior

conformance checking



	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
3	185783	030 Vastleggen toeko		
4	185783	050 Plannen afspraak		
5	185783	060 Aanmaken beves		
6	185783	070 Is 1e inspectie uit		
7	185783	100 Gereedmelden 1e		
8	185783	120 Plannen eindinsp		
9	185783	400 Is eindinspectie u		
10	185783	440 Zijn er nieuwe of		
11	185783	450 Krijgt de huurder		
12	185783	500 Beoordelen/wijzi		
13	185783	110 Bepalen leegstan		
14	185783	510 Is opleveringsfor		
15	185783	130 Is het opleverings		
16	185783	140 Aanmaken 1e in		
17	185783	150 Is er sprake van Z		
18	185783	180 Aanpassen wonin		
19	185783	190 Harmoniseren hu		
20	185783	205 Bepalen kandida		
21	185783	170 Aanpassen platte		
22	185783	520 Aanmaken 2e in		
23	185783	530 Aanmaken werko		
24	185783	540 Worden er bonus		
25	185783	550 Vastleggen bonus		
26	185783	240 Registreren voor		
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

Inspector

Selected elements*: 400 Is eindinspectie uitgevoerd ?

Property	Value
#Move log+model (total)	168
#Move log+model (in 100% fitting traces)	142
#Traces where move log+model occur	168
#Move model only (in all traces)	40
#Traces where move model only occur	40

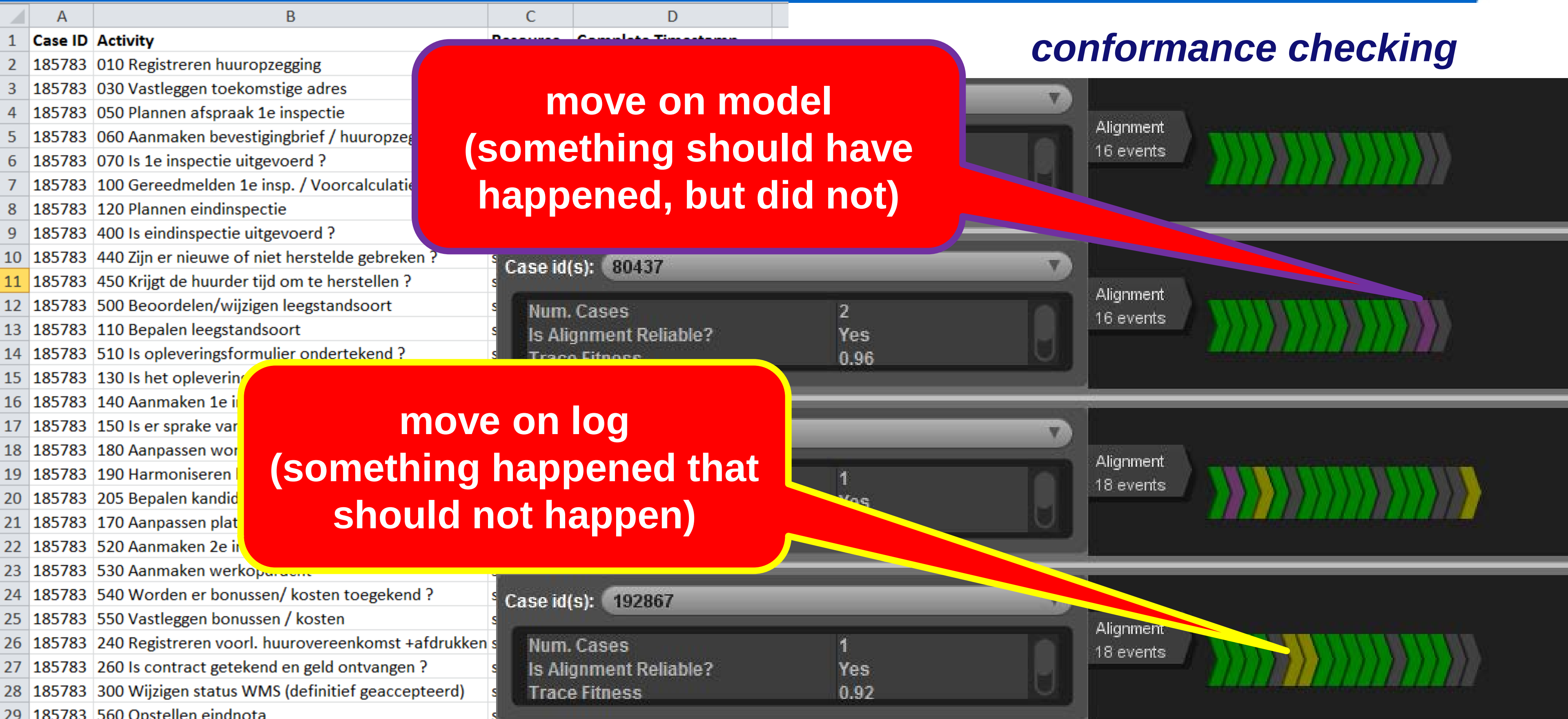
* Click a place/transition on the projected model to see its stats, or use combobox

Global Statistics (non-filtered traces)
070 Is 1e inspectie uitgevoerd ? (192/16)
tau split
Gereedmelden insp. / Voorcalculatie maken
110 Bepalen leegstandsoort (192/16)
120 Plannen eindinspectie (192/16)
130 Is het opleveringsform ondertekend ? (192/16)
400 Is eindinspectie uitgevoerd ? (168/40)
150 Is er sprake van ZAV ? (192/16)

final inspection is skipped 40 times

Process Mining: Spreadsheet for behavior

conformance checking



	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging		
3	185783	030 Vastleggen toekomstige adres		
4	185783	050 Plannen afspraak 1e inspectie		
5	185783	060 Aanmaken bevestigingbrief / huuropzeg		
6	185783	070 Is 1e inspectie uitgevoerd ?		
7	185783	100 Gereedmelden 1e insp. / Voorcalculatie		
8	185783	120 Plannen eindinspectie		
9	185783	400 Is eindinspectie uitgevoerd ?		
10	185783	440 Zijn er nieuwe of niet herstelde gebreken ?	s	
11	185783	450 Krijgt de huurder tijd om te herstellen ?	s	
12	185783	500 Beoordelen/wijzigen leegstandsoort	s	
13	185783	110 Bepalen leegstandsoort	s	
14	185783	510 Is opleveringsformulier ondertekend ?	s	
15	185783	130 Is het opleverin		
16	185783	140 Aanmaken 1e i		
17	185783	150 Is er sprake var		
18	185783	180 Aanpassen wor		
19	185783	190 Harmoniseren I		
20	185783	205 Bepalen kandid		
21	185783	170 Aanpassen plat		
22	185783	520 Aanmaken 2e i		
23	185783	530 Aanmaken werko		
24	185783	540 Worden er bonussen/ kosten toegekend ?	s	
25	185783	550 Vastleggen bonussen / kosten	s	
26	185783	240 Registreren voorl. huurovereenkomst +afdrukken	s	
27	185783	260 Is contract getekend en geld ontvangen ?	s	
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	s	
29	185783	560 Opstellen eindnota	s	

Alignment
16 events

Case id(s): 80437

Num. Cases 2
Is Alignment Reliable? Yes
Trace Fitness 0.96

Alignment
16 events

1
Yes

Alignment
18 events

Case id(s): 192867

Num. Cases 1
Is Alignment Reliable? Yes
Trace Fitness 0.92

Alignment
18 events

move on model (something should have happened, but did not)

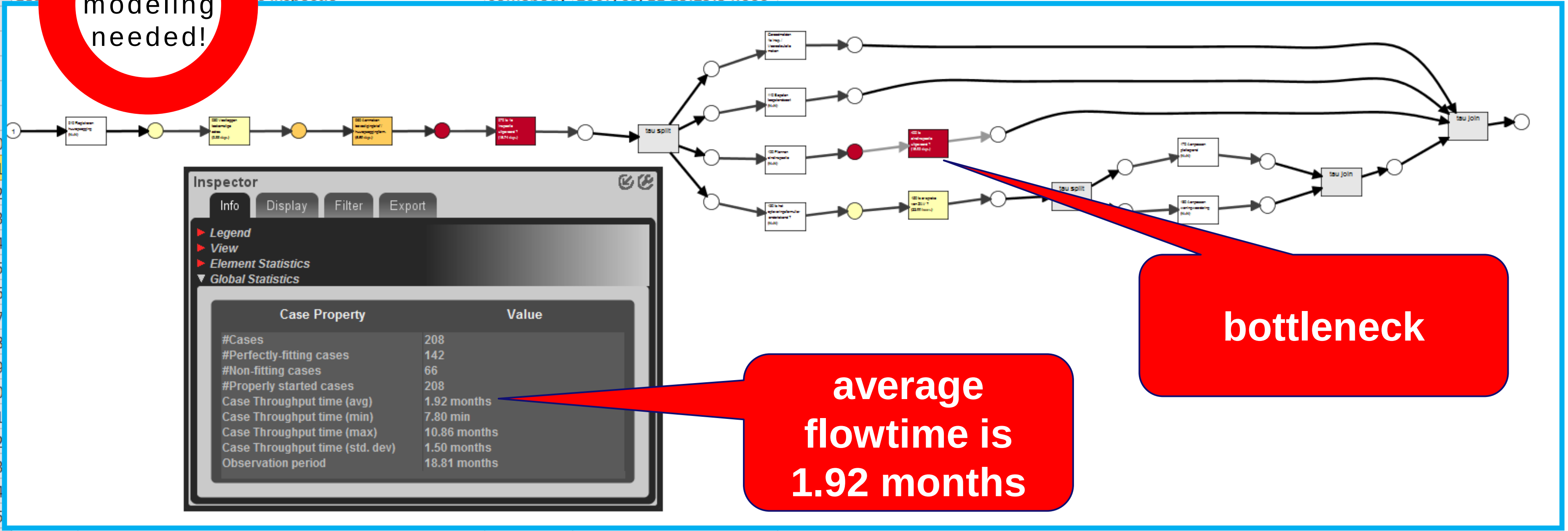
move on log (something happened that should not happen)

Process Mining: Spreadsheet for behavior

	A	B	C	D
1	Case ID		Resource	Complete Timestamp
2	1857	opzegging	somebody	2007/09/11 13:21:27.000
3	185	nstige adres	somebody	2007/09/11 13:26:29.000
4	185	e inspectie	somebody	2007/09/11 13:29:34.000
26	185783	240 Registreren voorl. huurovereenkomst +afdrukken	somebody	2007/11/28 12:34:23.000
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

NO modeling needed!

performance analysis



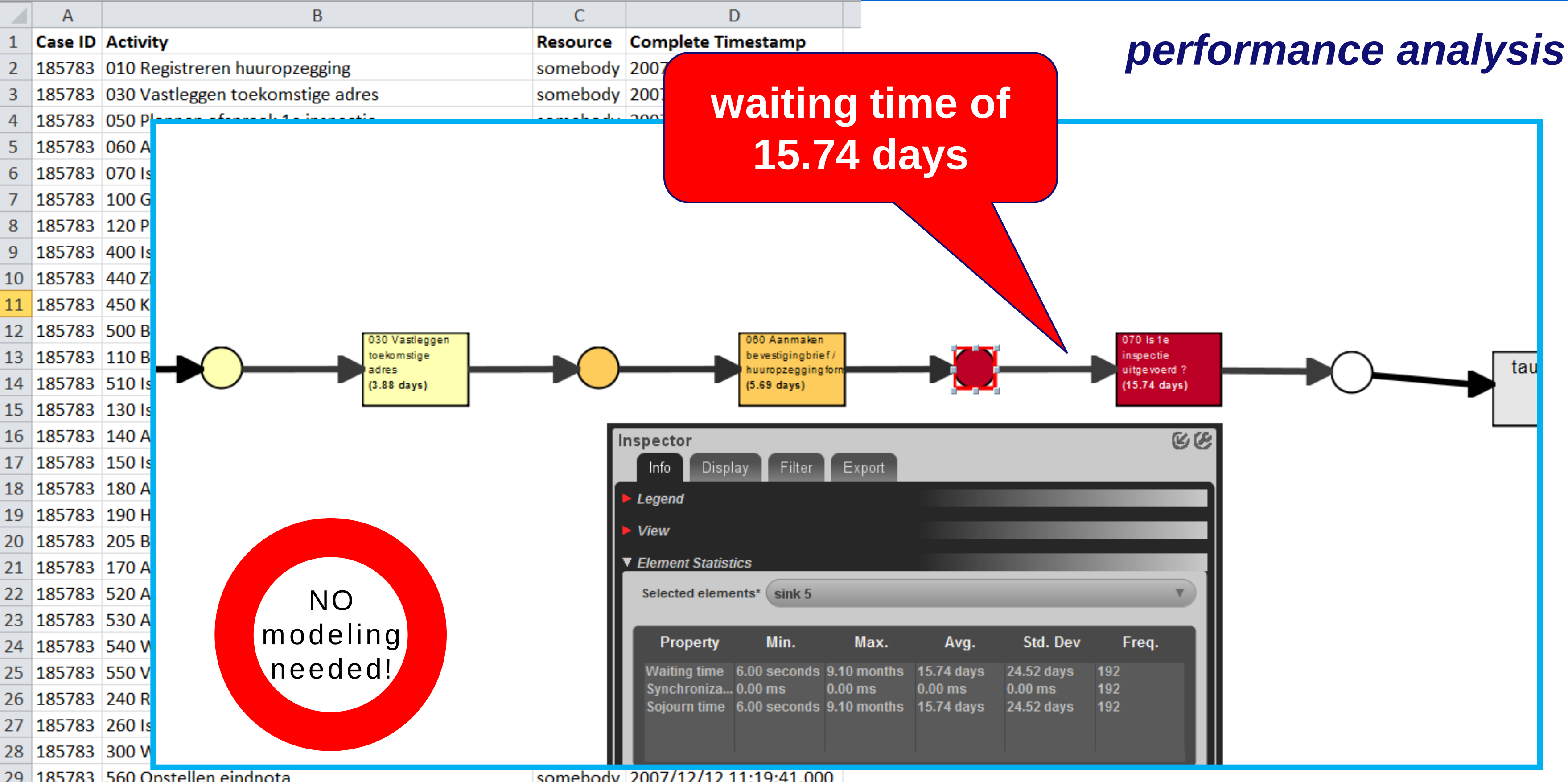
Case Property	Value
#Cases	208
#Perfectly-fitting cases	142
#Non-fitting cases	66
#Properly started cases	208
Case Throughput time (avg)	1.92 months
Case Throughput time (min)	7.80 min
Case Throughput time (max)	10.86 months
Case Throughput time (std. dev)	1.50 months
Observation period	18.81 months

bottleneck

average flowtime is 1.92 months

Process Mining: Spreadsheet for behavior

performance analysis



	A	B	C	D
1	Case ID	Activity	Resource	Complete Timestamp
2	185783	010 Registreren huuropzegging	somebody	2007
3	185783	030 Vastleggen toekomstige adres	somebody	2007
4	185783	050 Plannen afspraak 1e inspectie	somebody	2007
5	185783	060 A		
6	185783	070 Is		
7	185783	100 G		
8	185783	120 P		
9	185783	400 Is		
10	185783	440 Zi		
11	185783	450 K		
12	185783	500 B		
13	185783	110 B		
14	185783	510 Is		
15	185783	130 Is		
16	185783	140 A		
17	185783	150 Is		
18	185783	180 A		
19	185783	190 H		
20	185783	205 B		
21	185783	170 A		
22	185783	520 A		
23	185783	530 A		
24	185783	540 W		
25	185783	550 V		
26	185783	240 R		
27	185783	260 Is		
28	185783	300 W		
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

030 Vastleggen toekomstige adres (3.88 days)
060 Aanmaken bevestigingbrief / huuropzegging form (5.69 days)
070 Is 1e inspectie uitgevoerd ? (15.74 days)
tau

waiting time of 15.74 days

Inspector

Selected elements*: sink 5

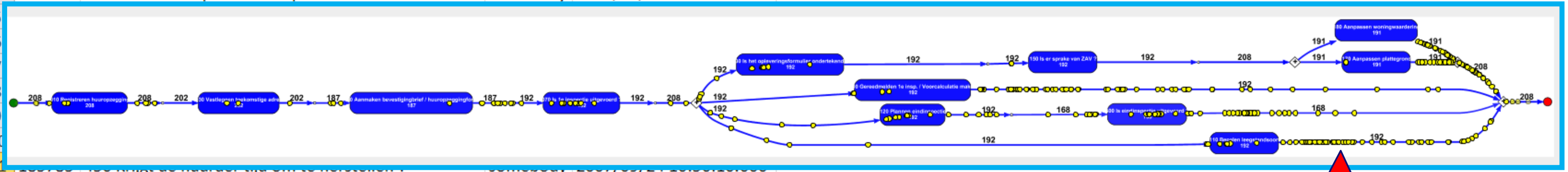
Property	Min.	Max.	Avg.	Std. Dev	Freq.
Waiting time	6.00 seconds	9.10 months	15.74 days	24.52 days	192
Synchroniza...	0.00 ms	0.00 ms	0.00 ms	0.00 ms	192
Sojourn time	6.00 seconds	9.10 months	15.74 days	24.52 days	192

NO modeling needed!

Process Mining: Spreadsheet for behavior

	A	B	C	D
1	**Case ID**	**Activity**	**Resource**	**Complete Timestamp**
2	185783	010 Registreren huuropzegging	somebody	2007/09/11 13:21:27.000
3	185783	030 Vastleggen toekomstige adres	somebody	2007/09/11 13:26:29.000
4	185783	050 Plannen afspraak 1e inspectie	somebody	2007/09/11 13:29:34.000
11	[illegible]	[illegible] Krijgt de huurder tijd om te herstellen ?	[illegible]	[illegible]
12	185783	500 Beoordelen/wijzigen leegstandsoort	somebody	2007/09/24 10:57:02.000
13	185783	110 Bepalen leegstandsoort	somebody	2007/09/24 10:57:42.000
14	185783	510 Is opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:08.000
15	185783	130 Is het opleveringsformulier ondertekend ?	somebody	2007/09/24 10:58:19.000
16	185783	140 Aanmaken 1e in gebreke stelling	somebody	2007/09/24 11:01:58.000
17	185783	150 Is er sprake van ZAV ?	somebody	2007/09/24 11:37:33.000
18	185783	180 Aanpassen woningwaardering	somebody	2007/09/24 11:37:44.000
19	185783	190 Harmoniseren huurprijs	somebody	2007/09/24 11:40:01.000
20	185783	205 Bepalen kandidaat huurder	somebody	2007/09/24 11:47:42.000
21	185783	170 Aanpassen plattegrond	somebody	2007/09/24 12:10:58.000
22	185783	520 Aanmaken 2e in gebreke stelling	somebody	2007/10/30 11:45:53.000
23	185783	530 Aanmaken werkopdracht	somebody	2007/10/30 11:46:09.000
24	185783	540 Worden er bonussen/ kosten toegekend ?	somebody	2007/10/30 11:46:36.000
25	185783	550 Vastleggen bonussen / kosten	somebody	2007/10/30 11:53:00.000
26	185783	240 Registreren voorl. huurovereenkomst +afdrukken	somebody	2007/11/28 12:34:23.000
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

animating reality

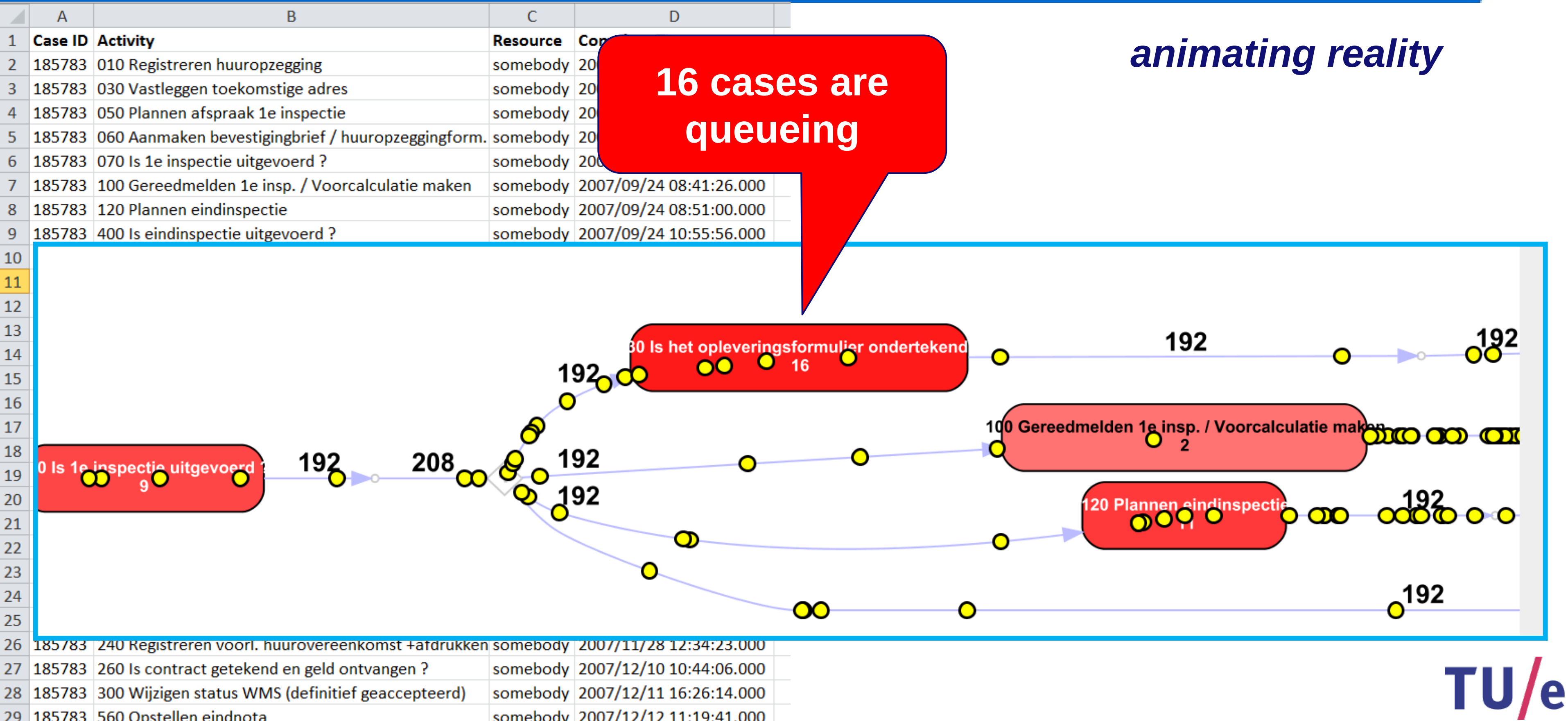

208
010 Registreren huuropzegging
208
202
030 Vastleggen toekomstige adres
202
187
Aanmaken bevestigingbrief / huuropzeggingsformulier
187
192
070 Is 1e inspectie uitgevoerd
192
208
Is het opleveringsformulier ondertekend
192
150 Is er sprake van ZAV ?
208
180 Aanpassen woningwaardering
191
170 Aanpassen plattegrond
191
Gereedmelden 1e insp. / Voorcalculatie maken
120 Plannen eindinspectie
168
Is eindinspectie uitgevoerd
110 Bepalen leegstandsoort
208

real cases

NO modeling needed!

Process Mining: Spreadsheet for behavior

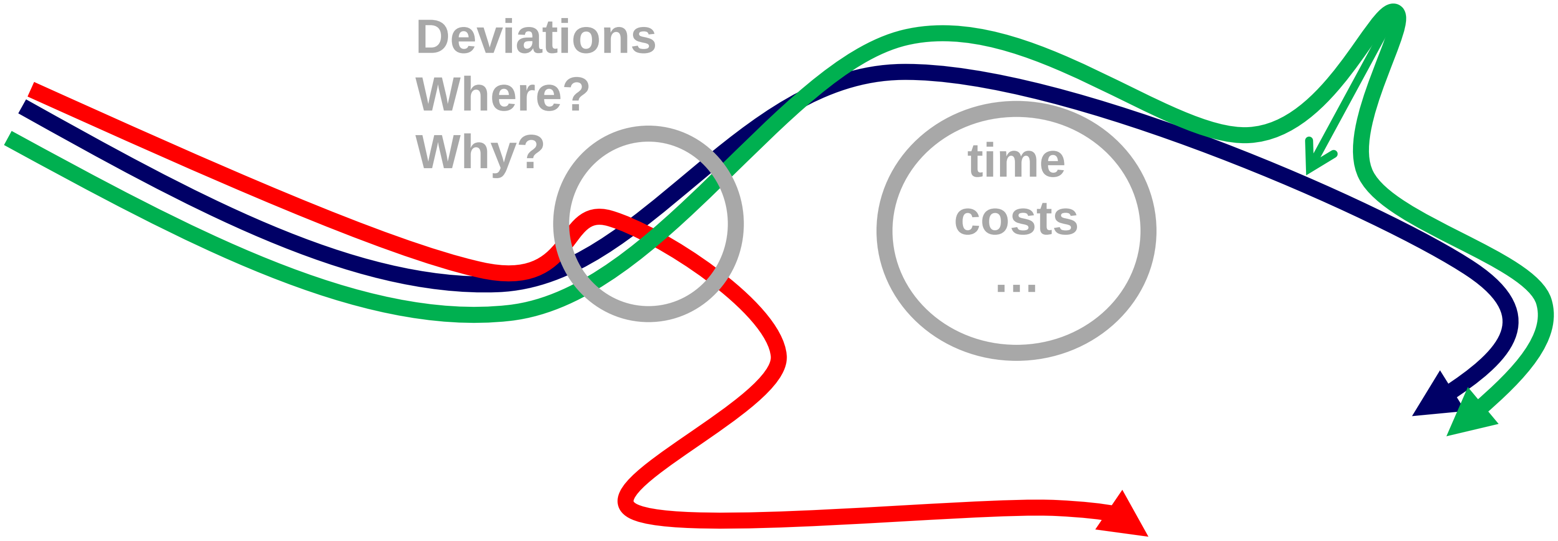
animating reality



	A	B	C	D
1	**Case ID**	**Activity**	**Resource**	**Co**
2	185783	010 Registreren huuropzegging	somebody	20
3	185783	030 Vastleggen toekomstige adres	somebody	20
4	185783	050 Plannen afspraak 1e inspectie	somebody	20
5	185783	060 Aanmaken bevestigingbrief / huuropzeggingform.	somebody	20
6	185783	070 Is 1e inspectie uitgevoerd ?	somebody	200
7	185783	100 Gereedmelden 1e insp. / Voorcalculatie maken	somebody	2007/09/24 08:41:26.000
8	185783	120 Plannen eindinspectie	somebody	2007/09/24 08:51:00.000
9	185783	400 Is eindinspectie uitgevoerd ?	somebody	2007/09/24 10:55:56.000
26	185783	240 Registreren voorl. huurovereenkomst +afdrukken	somebody	2007/11/28 12:34:23.000
27	185783	260 Is contract getekend en geld ontvangen ?	somebody	2007/12/10 10:44:06.000
28	185783	300 Wijzigen status WMS (definitief geaccepteerd)	somebody	2007/12/11 16:26:14.000
29	185783	560 Opstellen eindnota	somebody	2007/12/12 11:19:41.000

0 Is 1e inspectie uitgevoerd ?
9
192
208
192
192
30 Is het opleveringsformulier ondertekend
16
192
192
100 Gereedmelden 1e insp. / Voorcalculatie maken
2
120 Plannen eindinspectie
11
192
192

16 cases are queueing

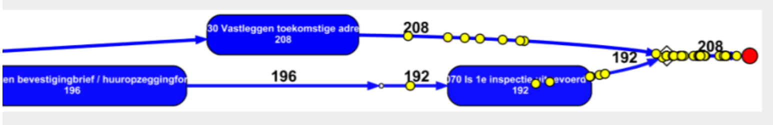


Deviations
Where?
Why?

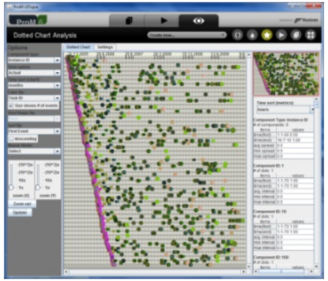
time
costs
...

Process Mining Software

ProM
process mining workbench


30 Vastleggen toekomstige adre
208
208
192
208
en bevestigingbrief / huuropzeggingfor
196
196
192
070 Is 1e inspectie uitgevoerd
192

1500+ plug-ins available covering the whole process mining spectrum



>128k downloads

Merge two Event Logs using AIS algorith

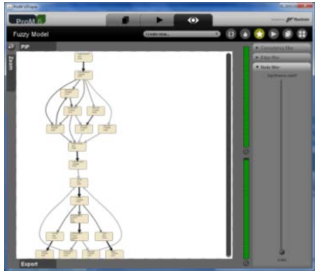
Mine for a Fuzzy Model

Mine for a Handover-of-Work Social Netw

Mine for a Petri Net using ILP

Mine for a Reassignment Social Network

Mine for a Similar-Task Social Network

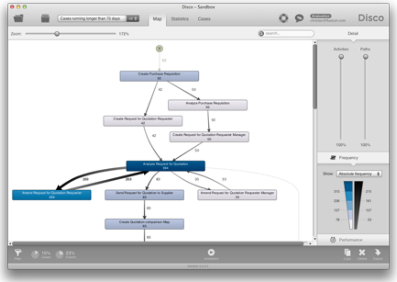


ProM

process mining workbench

Disco

perceptivesoftware
a Lexmark company

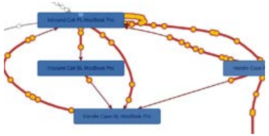


minit

Rialto by exeura

QPR
Quality. Processes. Results.

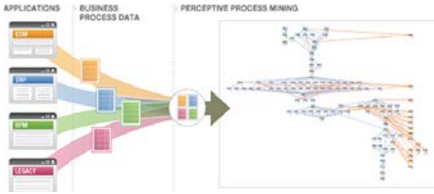
celonis
process mining



XMPRO
GET BETTER AT GETTING WORK DONE

stereoLogic

Software AG
Process Performance Manager
Discover, analyze and monitor:
the road to process improvement



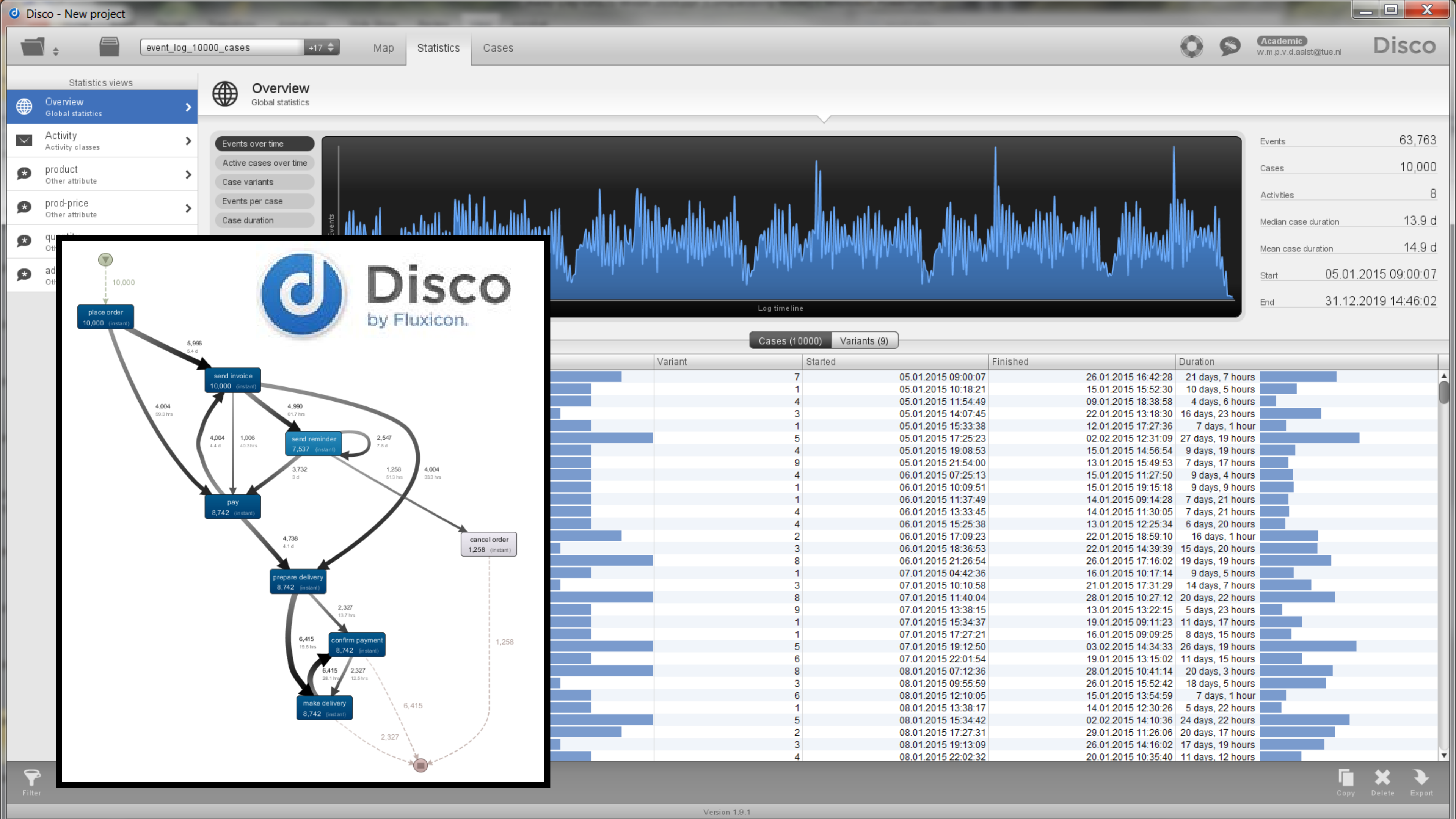
myinvenio

FUJITSU

SNP The Transformation Company

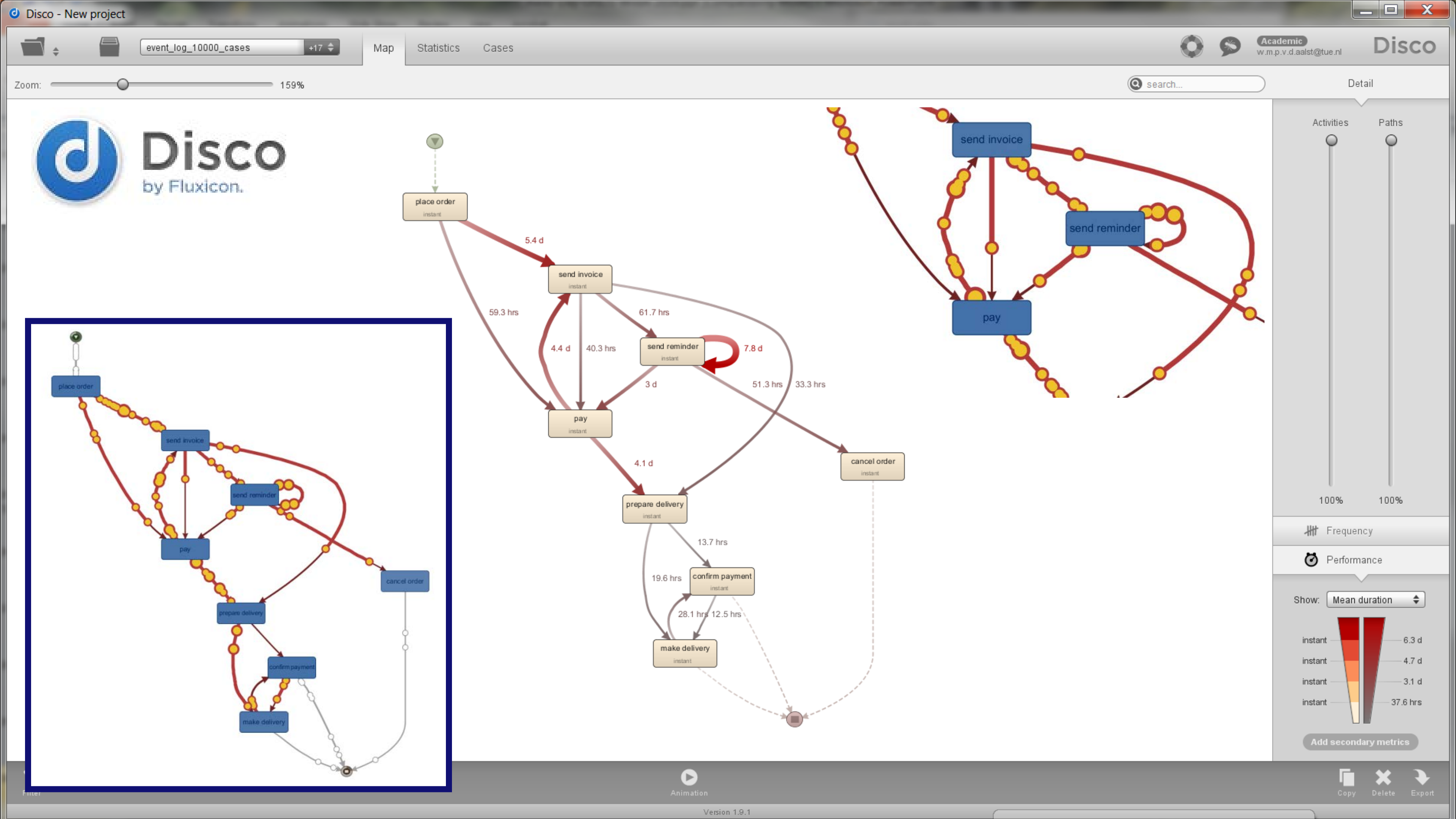
TU/e

©Wil van der Aalst & TU/e (use only with permission & acknowledgements)

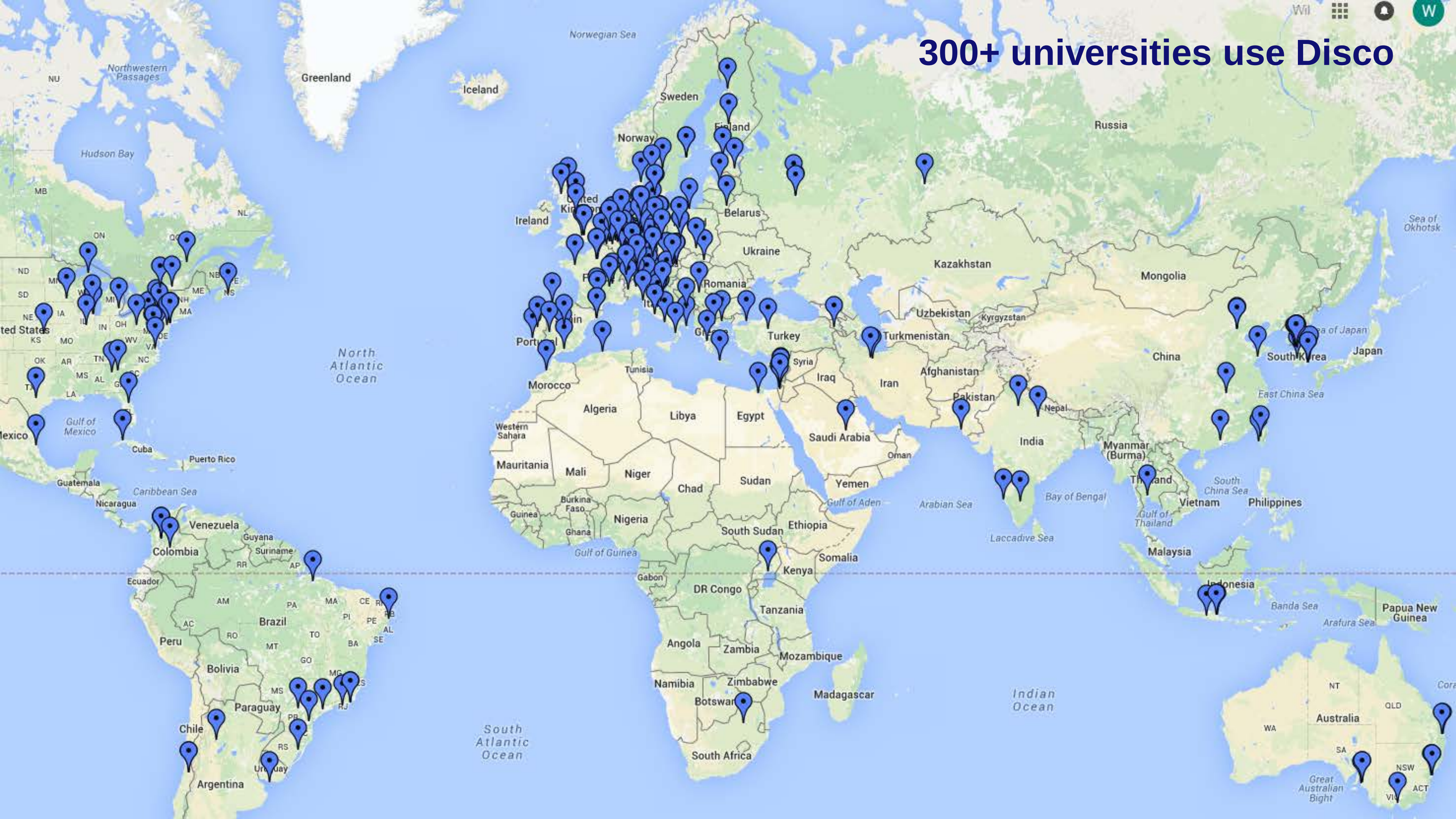


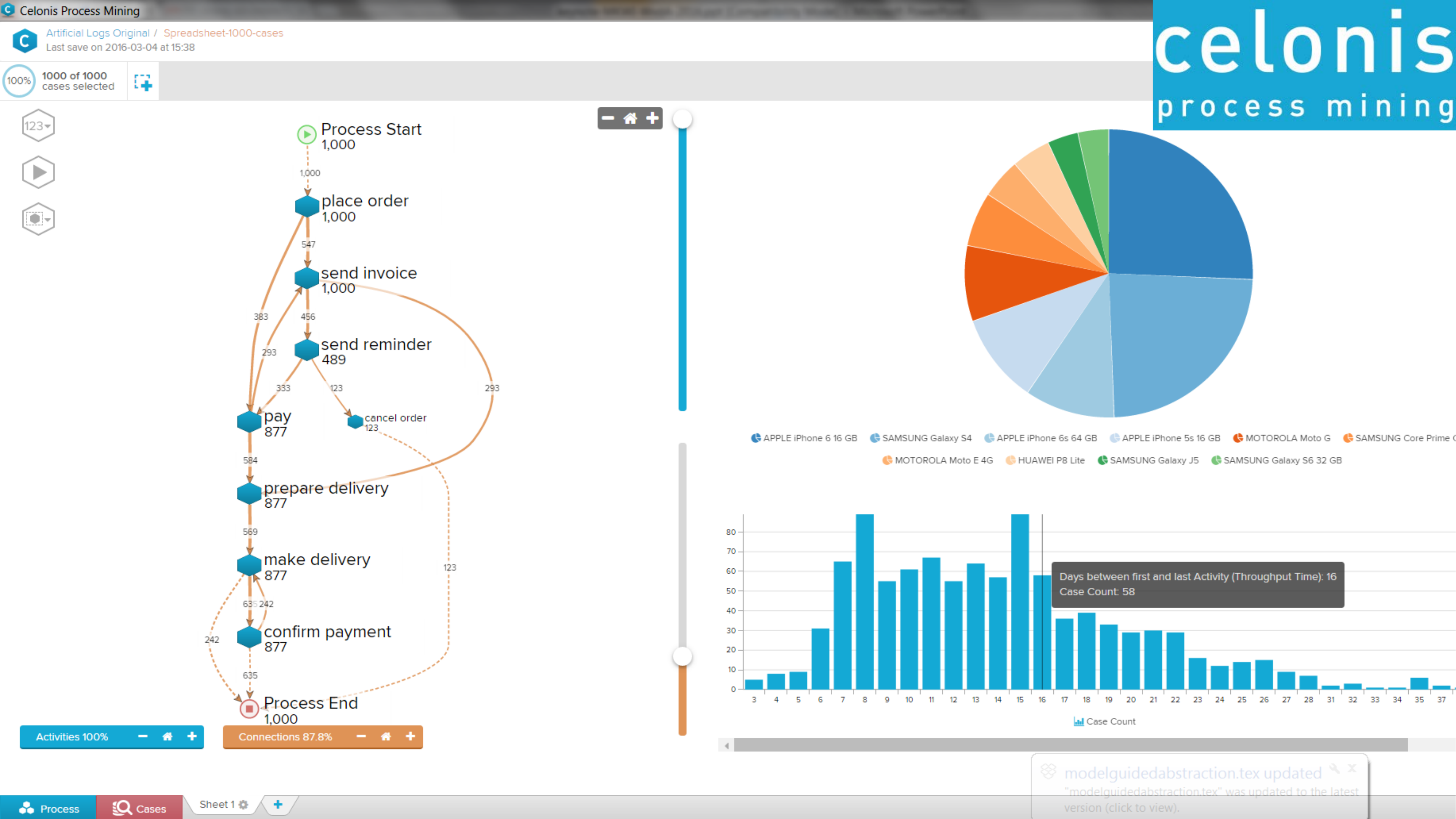
Events	63,763
Cases	10,000
Activities	8
Median case duration	13.9 d
Mean case duration	14.9 d
Start	05.01.2015 09:00:07
End	31.12.2019 14:46:02

Variant	Started	Finished	Duration
7	05.01.2015 09:00:07	26.01.2015 16:42:28	21 days, 7 hours
1	05.01.2015 10:18:21	15.01.2015 15:52:30	10 days, 5 hours
4	05.01.2015 11:54:49	09.01.2015 18:38:58	4 days, 6 hours
3	05.01.2015 14:07:45	22.01.2015 13:18:30	16 days, 23 hours
1	05.01.2015 15:33:38	12.01.2015 17:27:36	7 days, 1 hour
5	05.01.2015 17:25:23	02.02.2015 12:31:09	27 days, 19 hours
4	05.01.2015 19:08:53	15.01.2015 14:56:54	9 days, 19 hours
9	05.01.2015 21:54:00	13.01.2015 15:49:53	7 days, 17 hours
4	06.01.2015 07:25:13	15.01.2015 11:27:50	9 days, 4 hours
1	06.01.2015 10:09:51	15.01.2015 19:15:18	9 days, 9 hours
1	06.01.2015 11:37:49	14.01.2015 09:14:28	7 days, 21 hours
4	06.01.2015 13:33:45	14.01.2015 11:30:05	7 days, 21 hours
4	06.01.2015 15:25:38	13.01.2015 12:25:34	6 days, 20 hours
2	06.01.2015 17:09:23	22.01.2015 18:59:10	16 days, 1 hour
3	06.01.2015 18:36:53	22.01.2015 14:39:39	15 days, 20 hours
8	06.01.2015 21:26:54	26.01.2015 17:16:02	19 days, 19 hours
1	07.01.2015 04:42:36	16.01.2015 10:17:14	9 days, 5 hours
3	07.01.2015 10:10:58	21.01.2015 17:31:29	14 days, 7 hours
8	07.01.2015 11:40:04	28.01.2015 10:27:12	20 days, 22 hours
9	07.01.2015 13:38:15	13.01.2015 13:22:15	5 days, 23 hours
1	07.01.2015 15:34:37	19.01.2015 09:11:23	11 days, 17 hours
1	07.01.2015 17:27:21	16.01.2015 09:09:25	8 days, 15 hours
5	07.01.2015 19:12:50	03.02.2015 14:34:33	26 days, 19 hours
6	07.01.2015 22:01:54	19.01.2015 13:15:02	11 days, 15 hours
8	08.01.2015 07:12:36	28.01.2015 10:41:14	20 days, 3 hours
3	08.01.2015 09:55:59	26.01.2015 15:52:42	18 days, 5 hours
6	08.01.2015 12:10:05	15.01.2015 13:54:59	7 days, 1 hour
1	08.01.2015 13:38:17	14.01.2015 12:30:26	5 days, 22 hours
5	08.01.2015 15:34:42	02.02.2015 14:10:36	24 days, 22 hours
2	08.01.2015 17:27:31	29.01.2015 11:26:06	20 days, 17 hours
3	08.01.2015 19:13:09	26.01.2015 14:16:02	17 days, 19 hours
4	08.01.2015 22:02:32	20.01.2015 10:35:40	11 days, 12 hours


Disco - New project
event_log_10000_cases
+17
Map
Statistics
Cases
Academic
w.m.p.v.d.aalst@tue.nl
Disco
Zoom:
159%
search...
Detail
Disco
by Fluxicon.
place order
instant
5.4 d
send invoice
instant
59.3 hrs
61.7 hrs
4.4 d
40.3 hrs
send reminder
instant
7.8 d
3 d
51.3 hrs
33.3 hrs
pay
instant
4.1 d
cancel order
instant
prepare delivery
instant
13.7 hrs
19.6 hrs
confirm payment
instant
28.1 hrs
12.5 hrs
make delivery
instant
send invoice
send reminder
pay
place order
send invoice
send reminder
pay
cancel order
prepare delivery
confirm payment
make delivery
Activities
Paths
100%
100%
Frequency
Performance
Show:
Mean duration
instant
instant
instant
instant
6.3 d
4.7 d
3.1 d
37.6 hrs
Add secondary metrics
Filter
Animation
Copy
Delete
Export
Version 1.9.1

300+ universities use Disco





Celonis Process Mining

Artificial Logs Original / Spreadsheet-1000-cases

Last save on 2016-03-04 at 15:38

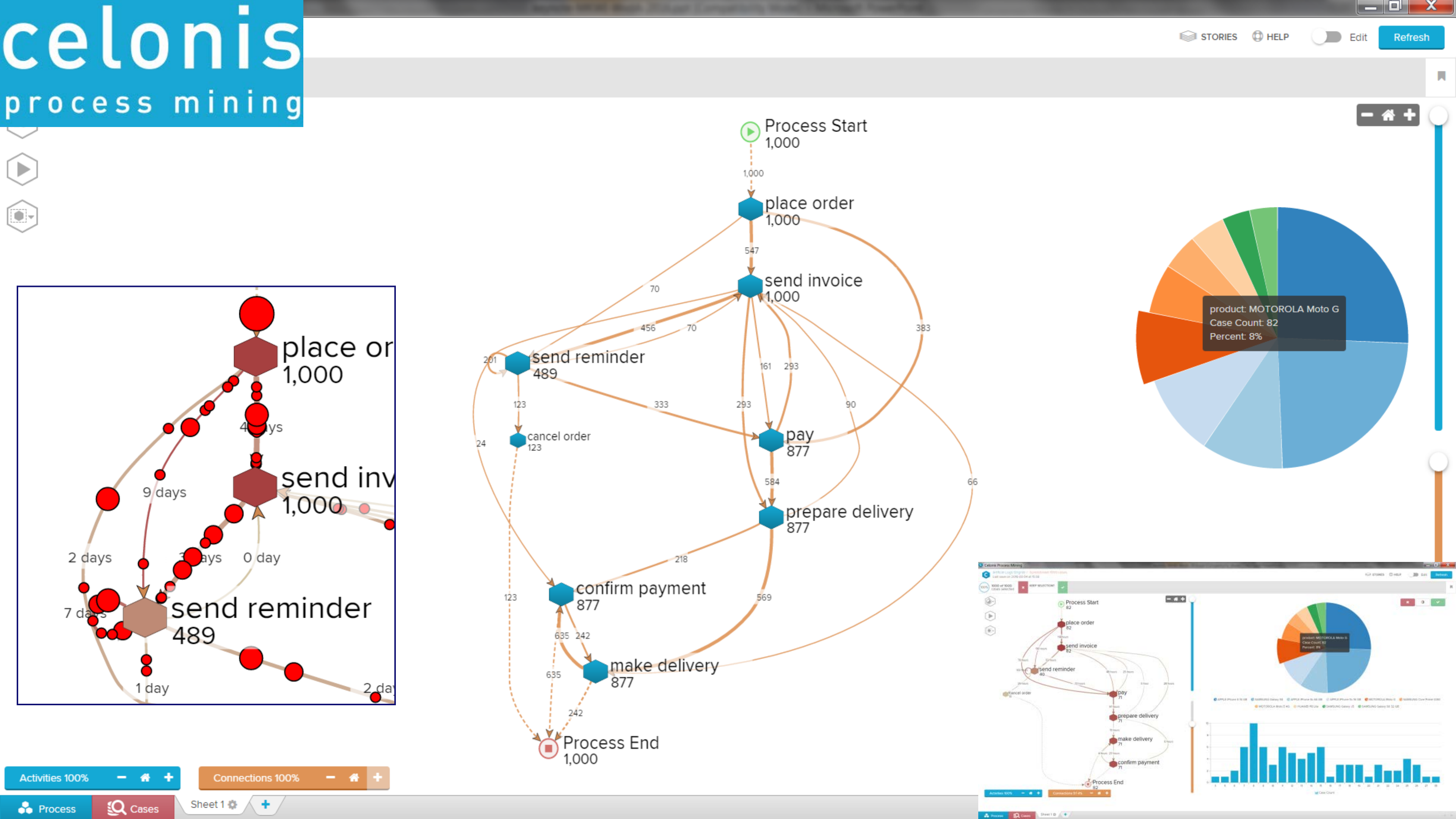
100% 1000 of 1000 cases selected

Days between first and last Activity (Throughput Time): 16
Case Count: 58

modelguidedabstraction.tex updated

"modelguidedabstraction.tex" was updated to the latest version (click to view).

celonis
process mining



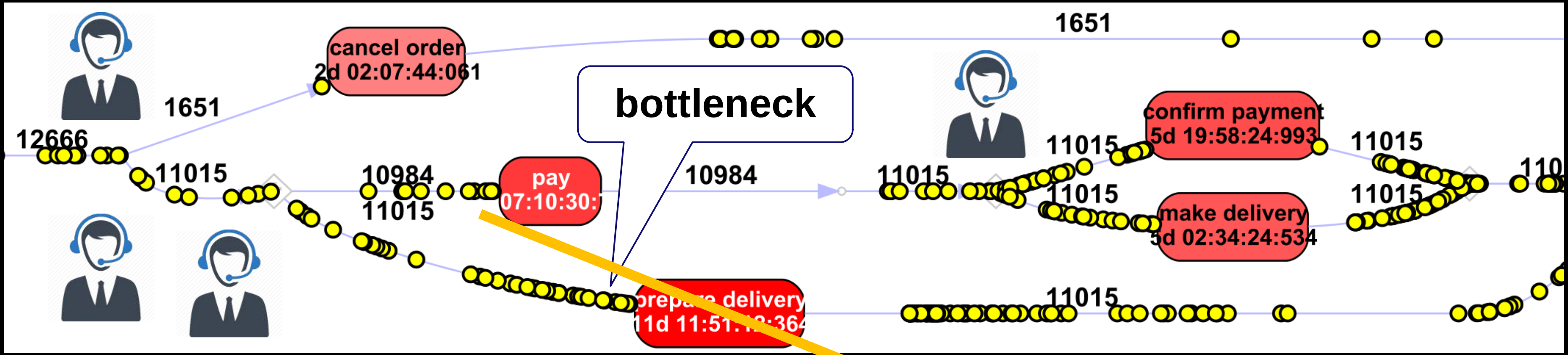
Part II

responsible data science: our next big challenge

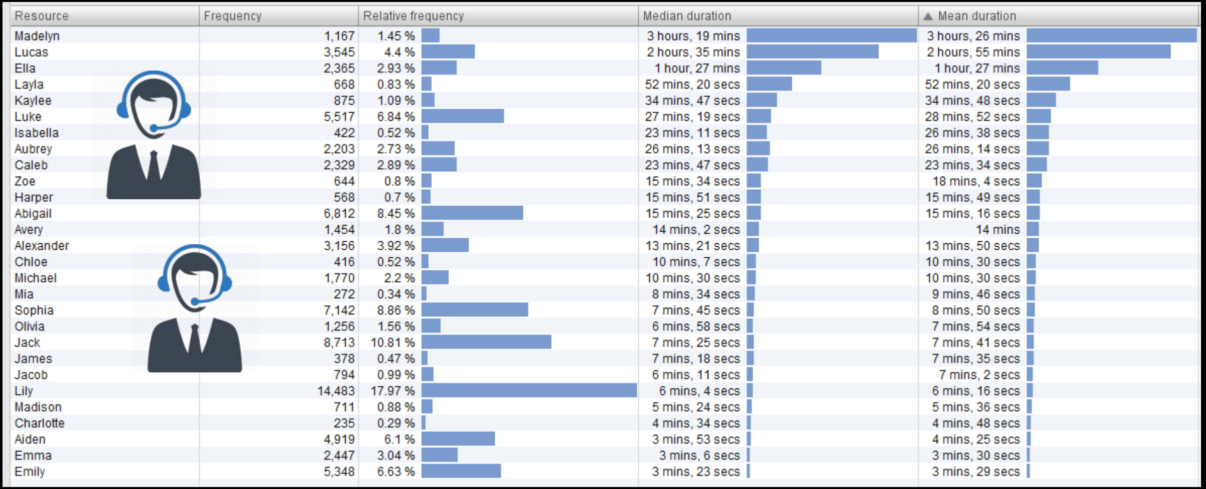
010110
10010
11011
10110
01111
10010
1101110
1001011
0101110
1001011
0100011
1000001
0101100
1001011
0101111
1001011
0101100

process mining will make things better, faster, more efficient, more effective, cheaper, …

With Great Power Comes Great Responsibility!!

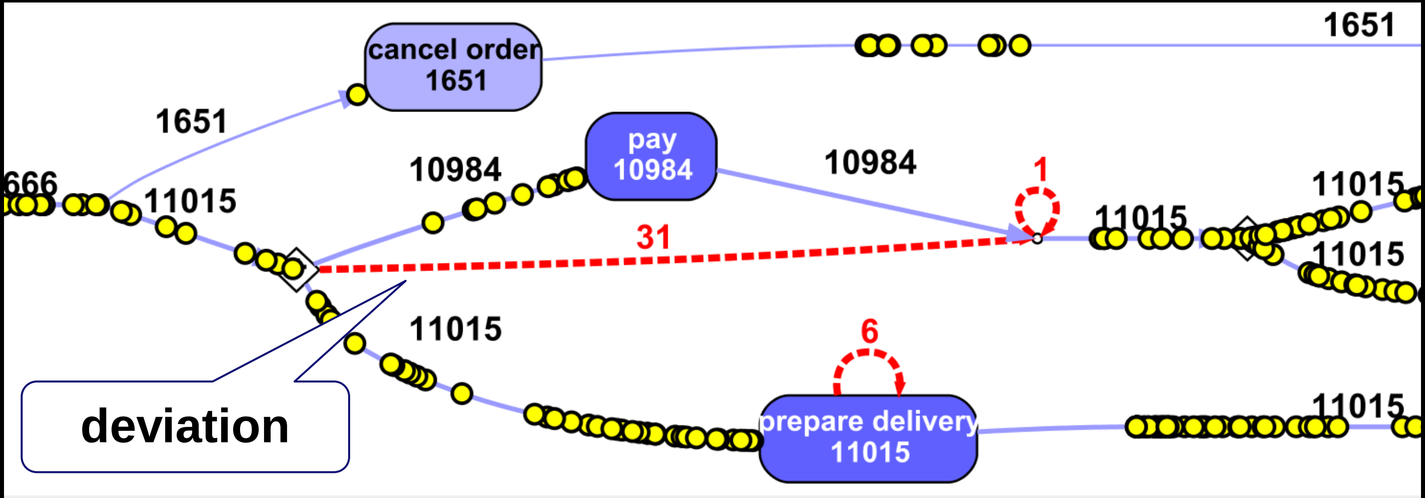


Who is causing these delays?



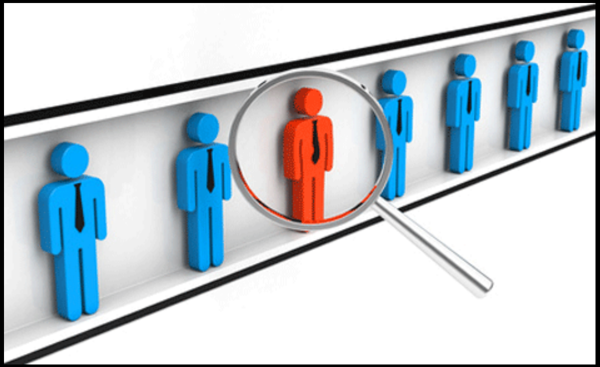
Resource	Frequency	Relative frequency	Median duration	▲ Mean duration
Madelyn	1,167	1.45 %	3 hours, 19 mins	3 hours, 26 mins
Lucas	3,545	4.4 %	2 hours, 35 mins	2 hours, 55 mins
Ella	2,365	2.93 %	1 hour, 27 mins	1 hour, 27 mins
Layla	668	0.83 %	52 mins, 20 secs	52 mins, 20 secs
Kaylee	875	1.09 %	34 mins, 47 secs	34 mins, 48 secs
Luke	5,517	6.84 %	27 mins, 19 secs	28 mins, 52 secs
Isabella	422	0.52 %	23 mins, 11 secs	26 mins, 38 secs
Aubrey	2,203	2.73 %	26 mins, 13 secs	26 mins, 14 secs
Caleb	2,329	2.89 %	23 mins, 47 secs	23 mins, 34 secs
Zoe	644	0.8 %	15 mins, 34 secs	18 mins, 4 secs
Harper	568	0.7 %	15 mins, 51 secs	15 mins, 49 secs
Abigail	6,812	8.45 %	15 mins, 25 secs	15 mins, 16 secs
Avery	1,454	1.8 %	14 mins, 2 secs	14 mins
Alexander	3,156	3.92 %	13 mins, 21 secs	13 mins, 50 secs
Chloe	416	0.52 %	10 mins, 7 secs	10 mins, 30 secs
Michael	1,770	2.2 %	10 mins, 30 secs	10 mins, 30 secs
Mia	272	0.34 %	8 mins, 34 secs	9 mins, 46 secs
Sophia	7,142	8.86 %	7 mins, 45 secs	8 mins, 50 secs
Olivia	1,256	1.56 %	6 mins, 58 secs	7 mins, 54 secs
Jack	8,713	10.81 %	7 mins, 25 secs	7 mins, 41 secs
James	378	0.47 %	7 mins, 18 secs	7 mins, 35 secs
Jacob	794	0.99 %	6 mins, 11 secs	7 mins, 2 secs
Lily	14,483	17.97 %	6 mins, 4 secs	6 mins, 16 secs
Madison	711	0.88 %	5 mins, 24 secs	5 mins, 36 secs
Charlotte	235	0.29 %	4 mins, 34 secs	4 mins, 48 secs
Aiden	4,919	6.1 %	3 mins, 53 secs	4 mins, 25 secs
Emma	2,447	3.04 %	3 mins, 6 secs	3 mins, 30 secs
Emily	5,348	6.63 %	3 mins, 23 secs	3 mins, 29 secs

Which customers pay late?



Which customers don’t pay at all?

Why is this employee deviating?

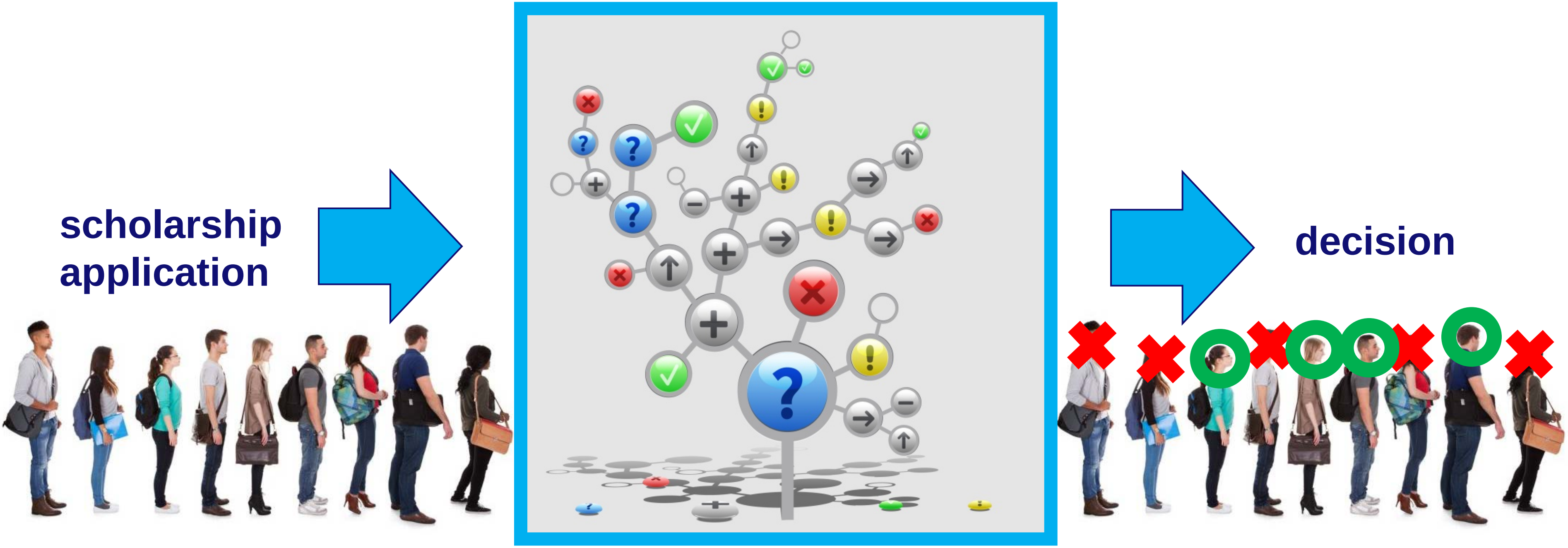


Responsible Data Science

Aim for positive solutions rather than avoiding the use of data

Fairness: Data Science without prejudice: How to avoid unfair conclusions even if they are true?

Standard classification problem

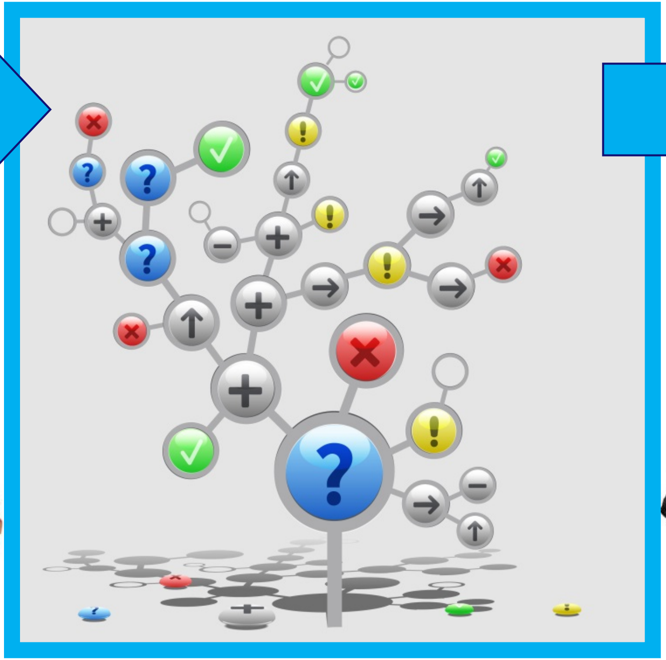


scholarship application

decision

Learn classifier using training data

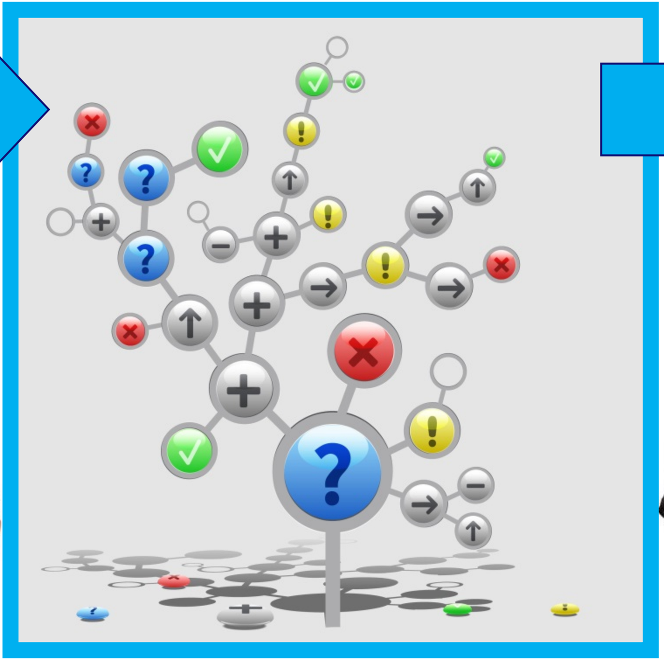
Name: Peter
Age: 28
Gender: Male
Country: German
Hobbies: Soccer
Fav. food: Sauerkraut
…



Graduated: Yes
Duration: 8 years
Average grade: 6.4
…

Tend to reject older male German students

Name: Peter
~~Age: 28~~
~~Gender: Male~~
~~Country: German~~
Hobbies: Soccer
Fav. food: Sauerkraut
…



Graduated: Yes
Duration: 8 years
Average grade: 6.4
…

Tend to reject “sauerkraut eating soccer fans”

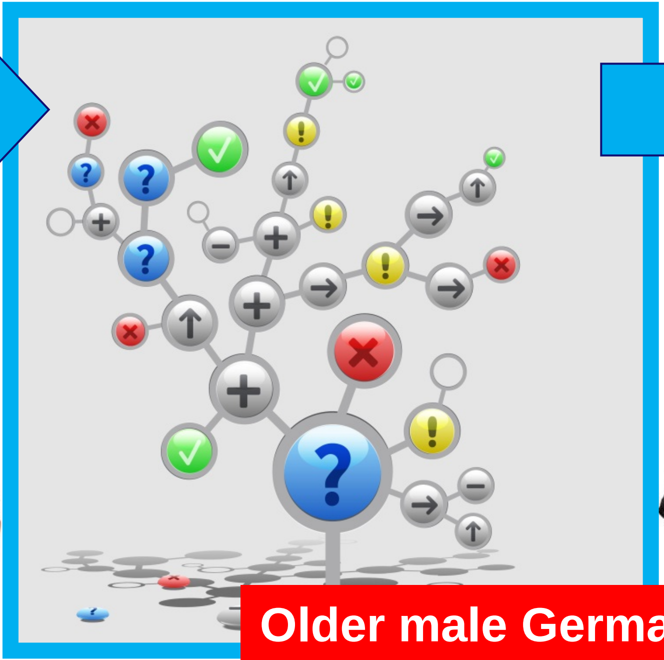
Name: Peter

confidential

Hobbies: Soccer

Fav. food: Sauerkraut

…



Graduated: Yes

Duration: 8 years

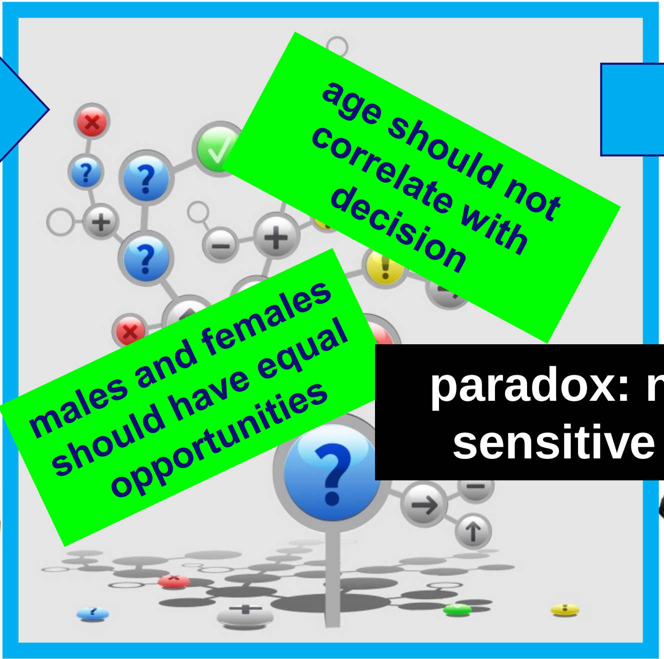
Average grade: 6.4

…

Older male German students still do not stand a chance to get a scholarship

Discrimination-aware classification

Name: Peter
Age: 28
Gender: Male
Country: German
Hobbies: Soccer
Fav. food: Sauerkraut
...



age should not correlate with decision

males and females should have equal opportunities

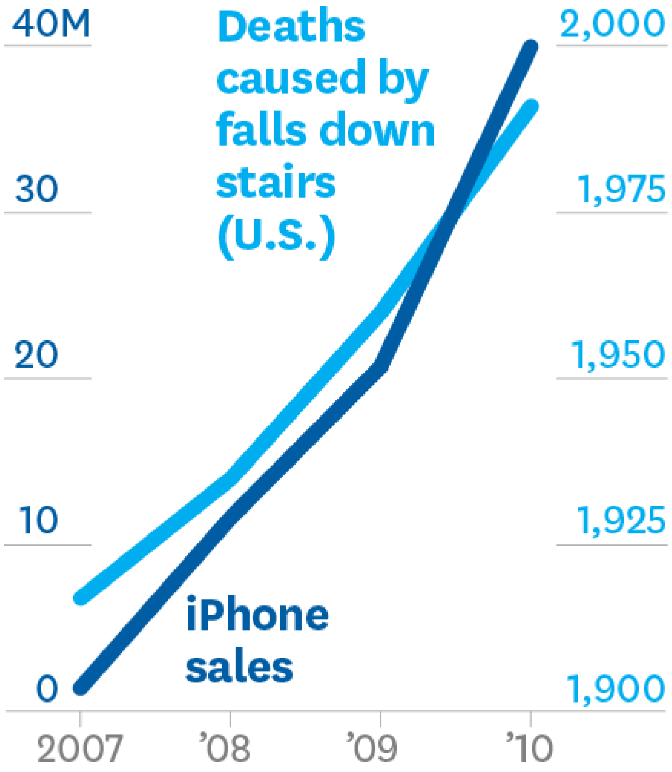
paradox: need to use sensitive attributes

add fairness constraint(s) to problem

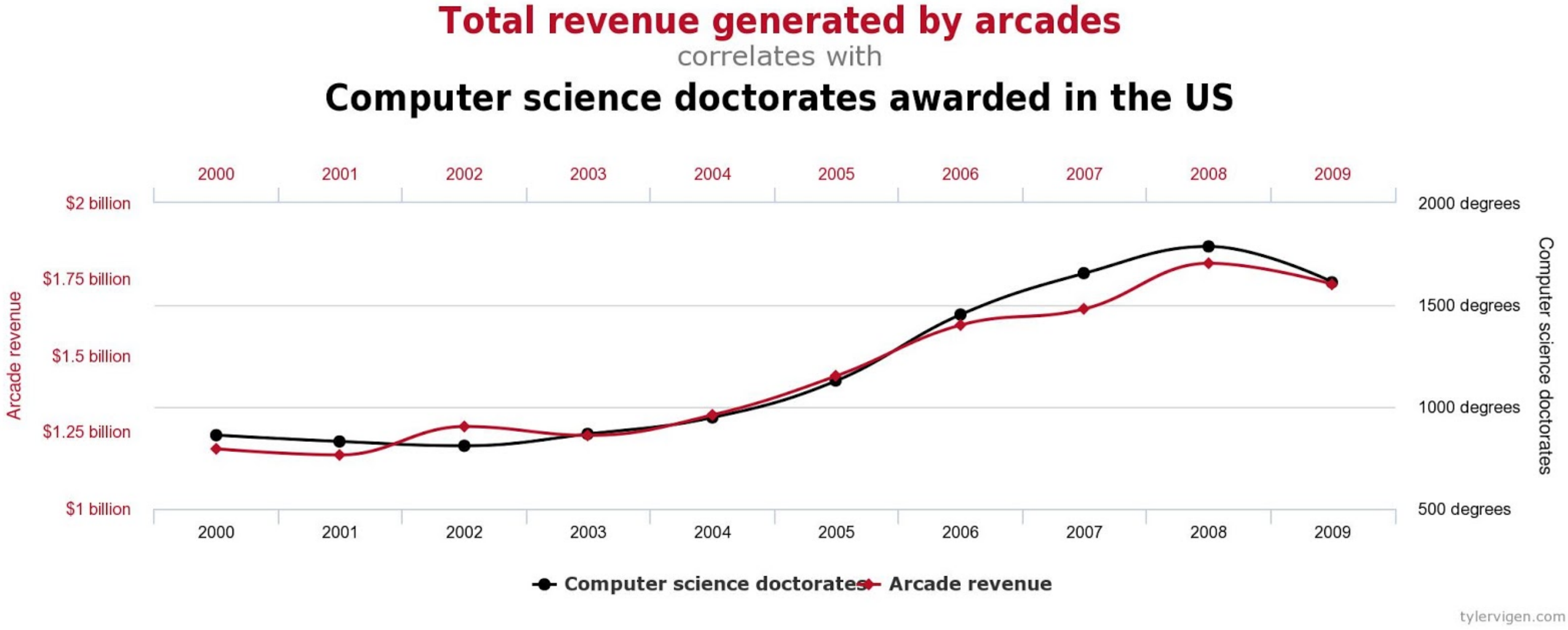
Graduated: Yes
Duration: 8 years
Average grade: 6.4
...

Accuracy: Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?

Spurious Correlations



Spurious Correlations



Curse of dimensionality

Test enough hypotheses and one will be true by accident (Carlo Emilio Bonferroni)

find the terrorists

Assumptions:

- 18 million people in NL
- 1800 hotels
- 100 guests per hotel per night
- (visit hotel every 100 days)

Algemene Inlichtingen-
en Veiligheidsdienst

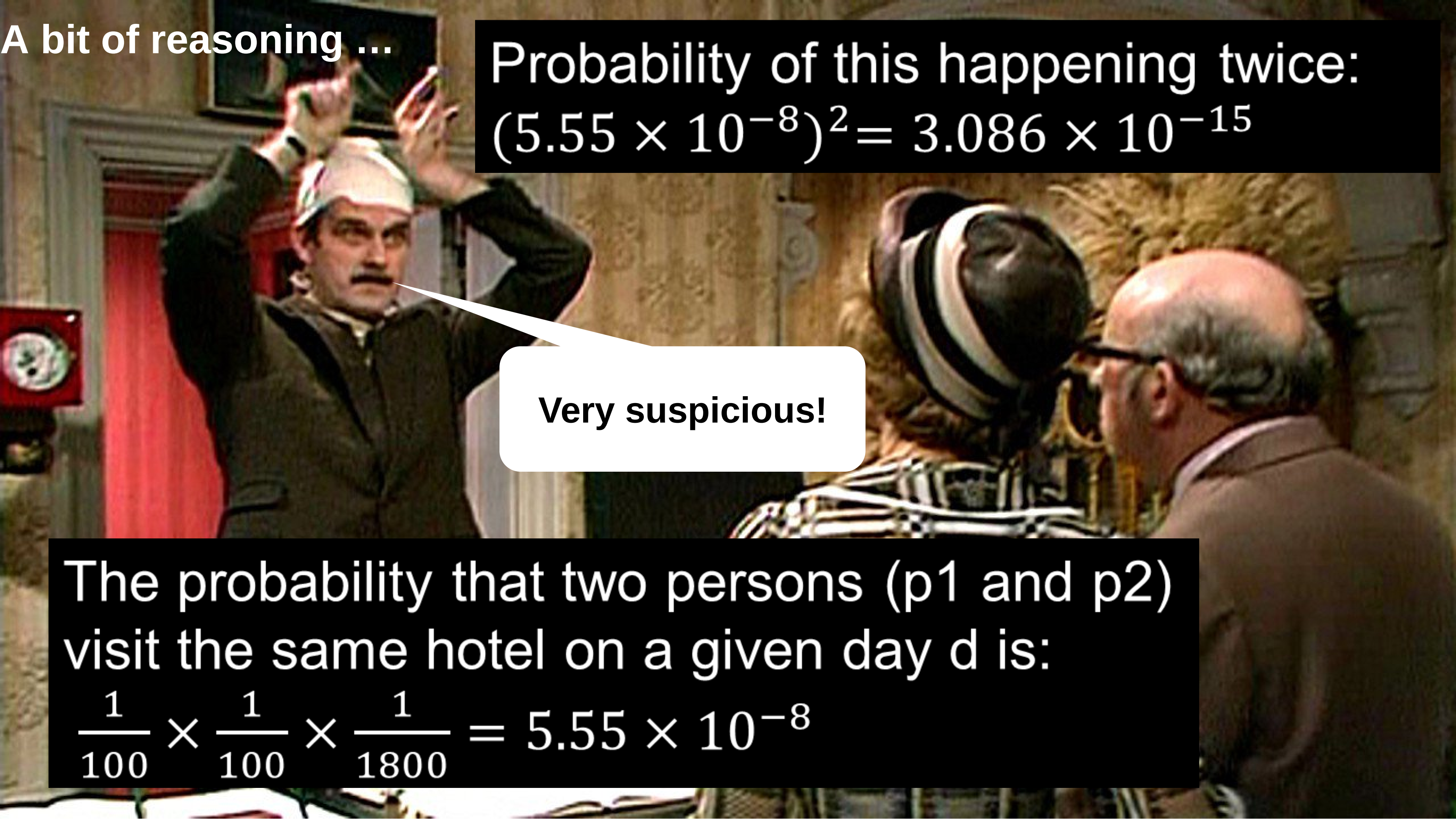
Suspicious event: two persons stay in the same hotel on two different dates

How many suspicious events in a 1000 day period?

A bit of reasoning …

Probability of this happening twice:

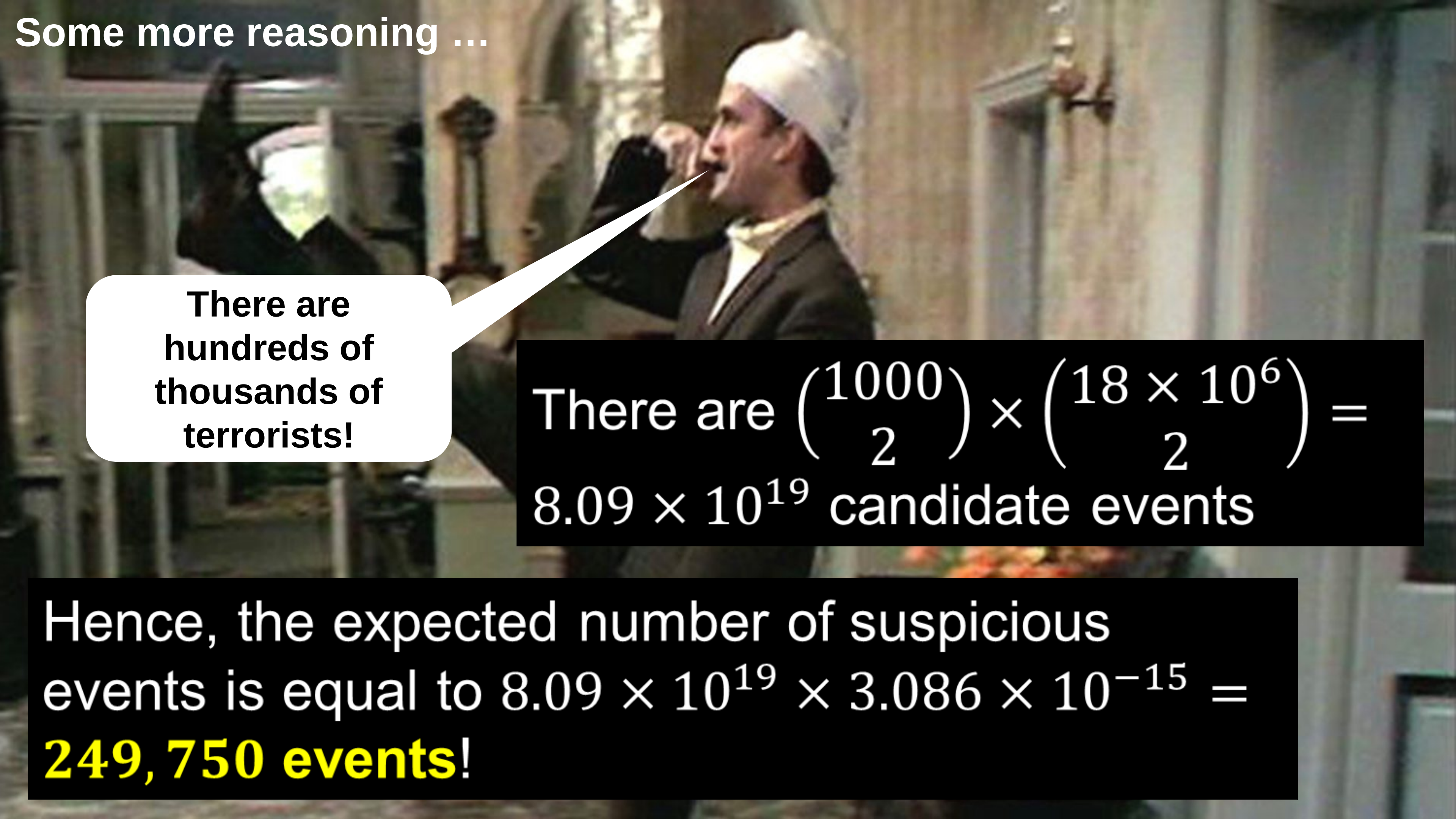
$$(5.55 \times 10^{-8})^2 = 3.086 \times 10^{-15}$$



The probability that two persons (p1 and p2) visit the same hotel on a given day d is:

$$\frac{1}{100} \times \frac{1}{100} \times \frac{1}{1800} = 5.55 \times 10^{-8}$$

Some more reasoning …



There are $\binom{1000}{2} \times \binom{18 \times 10^6}{2} =$ 8.09×10^{19} candidate events

Hence, the expected number of suspicious events is equal to $8.09 \times 10^{19} \times 3.086 \times 10^{-15} =$ **$249,750$ events**!

Confidentiality: Data Science that ensures confidentiality: How to answer questions without revealing secrets?

How to compute results with a predefined “privacy budget”?

How to distribute analysis such that nobody has the data?

How to share data in a safe manner?

Jason Ford

Transparency: Data Science that provides transparency: How to clarify answers such that they become indisputable?

How to present results such that people understand?

How to make the “data science pipeline” transparent?

Do analysis results indeed influence people as intended?

How to reveal analysis choices and risks related to the input data?

Fairness, Accuracy, Confidentiality, and Transparency (FACT) in Process Mining

Green Data Science
Using Big Data in an “Environmentally Friendly” Manner

Wil M. P. van der Aalst
Eindhoven University of Technology, Department of Mathematics and Computer Science,
PO Box 513, NL-5600 MB Eindhoven, The Netherlands
w.m.p.v.d.aalst@tue.nl

Keywords: Data Science, Big Data, Fairness, Confidentiality, Accuracy, Transparency, Process Mining.

Abstract: The widespread use of “Big Data” is heavily impacting organizations and individuals for which these data are collected. Sophisticated data science techniques aim to extract as much value from data as possible. Powerful mixtures of Big Data and analytics are rapidly changing the way we do business, socialize, conduct research, and govern society. Big Data is considered as the “new oil” and data science aims to transform this into new forms of “energy”: insights, diagnostics, predictions, and automated decisions. However, the process of transforming “new oil” (data) into “new energy” (analytics) may negatively impact citizens, patients, customers, and employees. Systematic discrimination based on data, invasions of privacy, non-transparent life-changing decisions, and inaccurate conclusions illustrate that data science techniques may lead to new forms of “pollution”. We use the term “Green Data Science” for technological solutions that enable individuals, organizations and society to reap the benefits from the widespread availability of data while ensuring fairness, confidentiality, accuracy, and transparency. To illustrate the scientific challenges related to “Green Data Science”, we focus on process mining as a concrete example. Recent breakthroughs in process mining resulted in powerful techniques to discover the real processes, to detect deviations from normative process models, and to analyze bottlenecks and waste. Therefore, this paper poses the question: How to benefit from process mining while avoiding “pollutions” related to unfairness, undesired disclosures, inaccuracies, and non-transparency?

1 INTRODUCTION

In recent years, data science emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. We use the following definition: *“Data science is an interdisciplinary field aiming to turn data into real value. Data may be structured or unstructured, big or small, static or streaming. Value may be provided in the form of predictions, models learned from data, or any type*
of data visualizati
includes data ex
ploration, data tr
computing infras
and learning, pr
dictions, and the
account ethical, s
(Aalst, 2016).

Related to dat
Data” that is use
of data collected
ing in Big Data

citizens, patients, customers, and employees are concerned about the use of their data. We live in an era characterized by unprecedented opportunities to sense, store, and analyze data related to human activities in great detail and resolution. This introduces new risks and intended or unintended abuse enabled by powerful analysis techniques. Data may be sensitive and personal, and should not be revealed or used for proposes different from what was agreed upon. Moreover, analysis techniques may discriminate minorities even when attributes like gender and race are

Aalst, W.
Green Data Science - Using Big Data in an “Environmentally Friendly” Manner.
In *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016) - Volume 1*, pages 9-21
ISBN: 978-989-758-187-8

Green Data Science - Using Big Data in an “Environmentally Friendly” Manner.
In *Proceedings of the 18th International Conference on Enterprise Information Systems (ICEIS 2016) - Volume 1*, pages 9-21
ISBN: 978-989-758-187-8

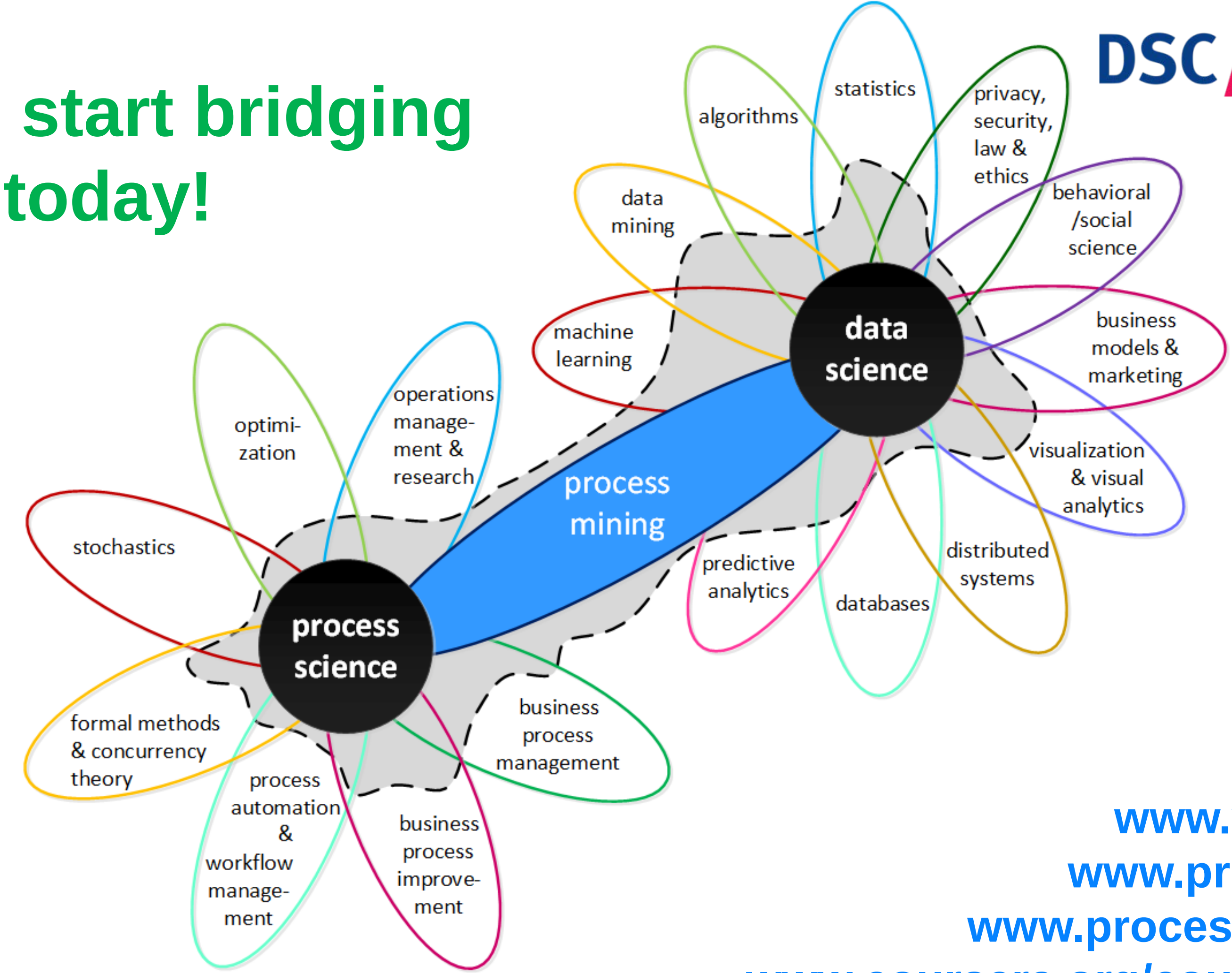
	creating and managing event data	process discovery	conformance checking	performance analysis	operational support
fairness **Data Science without prejudice: How to avoid unfair conclusions even if they are true?**	The input data may be biased, incomplete or incorrect such that the analysis reconfirms prejudices. By resampling or relabeling the data, undesirable forms of discrimination can be avoided. Note that both cases and resources (used to execute activities) may refer to individuals having sensitive attributes such as race, gender, age, etc.	The discovered model may abstract from paths followed by certain under-represented groups of cases. Discrimination-aware process-discovery algorithms can be used to avoid this. For example, if cases are handled differently based on gender, we may want to ensure that both are equally represented in the model.	Conformance checking can be used to “blame” individuals, groups, or organizations for deviating from some normative model. Discrimination-aware conformance checking (e.g., alignments) needs to separate (1) likelihood, (2) severity and (3) blame. Deviations may need to be interpreted differently for different groups of cases and resources.	Straightforward performance measurements may be unfair for certain classes of cases and resources (e.g., not taking into account the context). Discrimination-aware performance analysis detects unfairness and supports process improvements taking into account trade-offs between internal fairness (worker’s perspective) and external fairness (citizen/patient/customer’s perspective).	Process-related predictions, recommendations and decisions may discriminate (un)intentionally. This problem can be tackled using techniques from discrimination-aware data mining.
confidentiality **Data Science that ensures confidentiality: How to answer questions without revealing secrets?**	Event data (e.g., XES files) may reveal sensitive information. Anonymization and de-identification can be used to avoid disclosure. Note that timestamps and paths may be unique and a source for re-identification (e.g., all paths are unique).	The discovered model may reveal sensitive information, especially with respect to infrequent paths or small event logs. Drilling-down from the model may need to be blocked when numbers get too small (cf. k-anonymity).	Conformance checking shows diagnostics for deviating cases and resources. Access-control is important and diagnostics need to be aggregated to avoid revealing compliance problems at the level of individuals.	Performance analysis shows bottlenecks and other problems. Linking these problems to cases and resources may disclose sensitive information.	Process-related predictions, recommendations and decisions may disclose sensitive information, e.g., based on a rejection other properties can be derived.
accuracy **Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?**	Event data (e.g., XES files) may have all kinds of quality problems. Attributes may be incorrect, imprecise, or uncertain. For example, timestamps may be too coarse (just the date) or reflect the time of recording rather than the time of the event’s occurrence.	Process discovery depends on many parameters and characteristics of the event log. Process models should better show the confidence level of the different parts. Moreover, additional information needs to be used better (domain knowledge, uncertainty in event data, etc.).	Often multiple explanations are possible to interpret non-conformance. Just providing one alignment based on a particular cost function may be misleading. How robust are the findings?	In case of fitness problems (process model and event log disagree), performance analysis is based on assumptions and needs to deal with missing values (making results less accurate).	Inaccurate process models may lead to flawed predictions, recommendations and decisions. Moreover, not communicating the (un)certainty of predictions, recommendations and decisions, may negatively impact processes.
transparency **Data Science that provides transparency: How to clarify answers such that they become indisputable?**	Provenance of event data is key. Ideally, process mining insights can be related to the event data they are based on. However, this may conflict with confidentiality concerns.	Discovered process models depend on the event data used as input and the parameter settings and choice of discovery algorithm. How to ensure that the process model is interpreted correctly? End-users need to understand the relation between data and model to trust analysis.	When modeled and observed behavior disagree there may be multiple explanations. How to ensure that conformance diagnostics are interpreted correctly?	When detecting performance problems, it should be clear how these were detected and what the possible causes are. Animating event logs on models helps to make problems more transparent.	Predictions, recommendations and decisions are based on process models. If possible, these models should be transparent. Moreover, explanations should be added to predictions, recommendations and decisions (“We predict that this case be late, because ...”).

Conclusion

2 responsible data science: our next big challenge

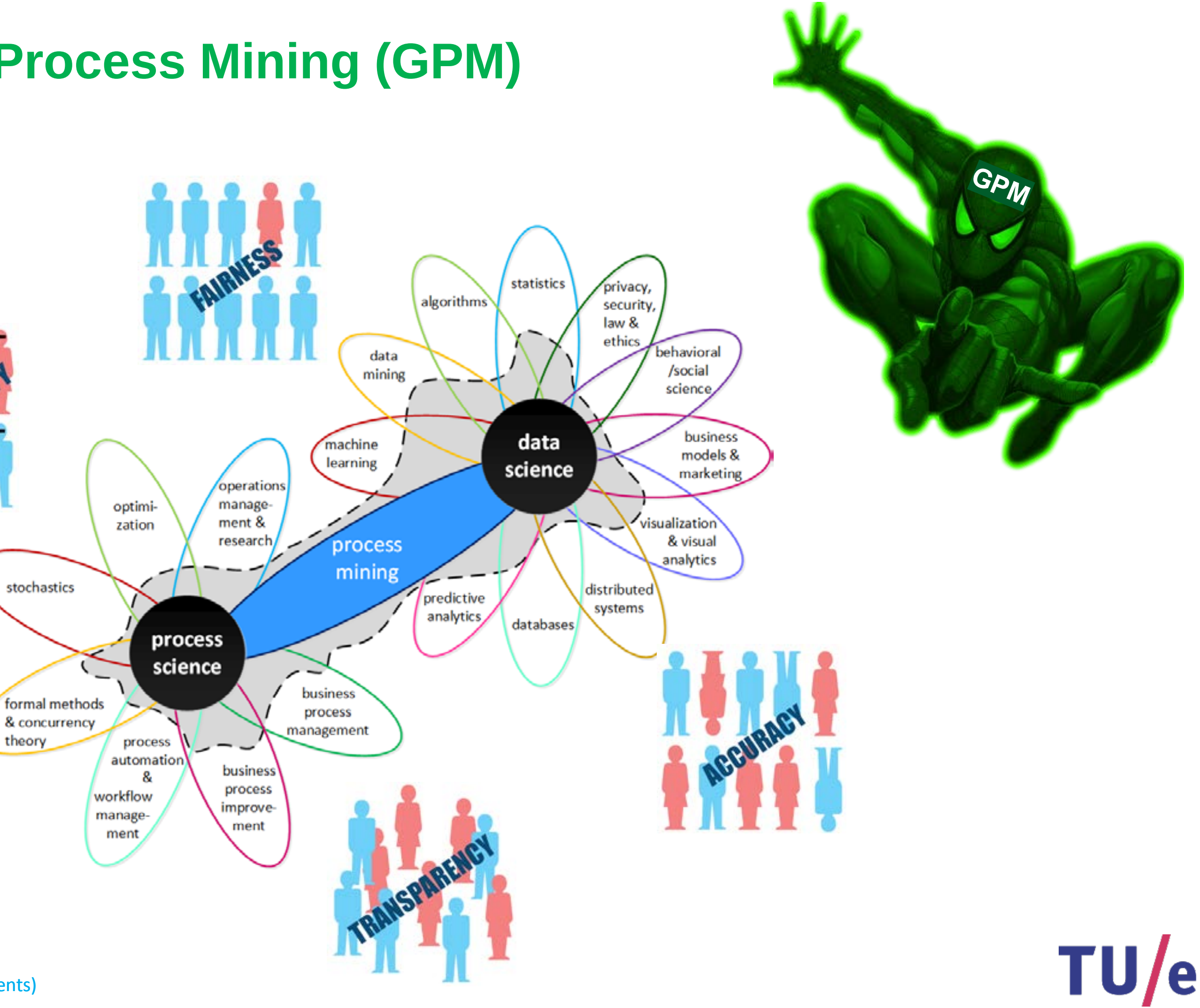
1 process mining: creating value from data

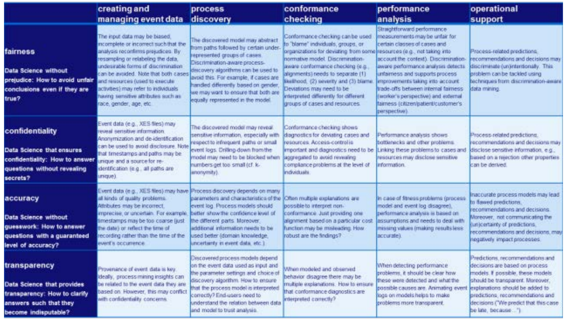
You can start bridging the gap today!


statistics
algorithms
privacy, security, law & ethics
data mining
behavioral /social science
machine learning
data science
business models & marketing
visualization & visual analytics
process mining
predictive analytics
databases
distributed systems
optimi- zation
operations manage- ment & research
stochastics
process science
formal methods & concurrency theory
business process management
process automation & workflow manage- ment
business process improve- ment

@wvdaalst
www.vdaalst.com
www.promtools.org
www.processmining.org
www.coursera.org/course/procmin

Next challenge: Green Process Mining (GPM)


GPM
FAIRNESS
statistics
algorithms
privacy, security, law & ethics
data mining
behavioral /social science
machine learning
data science
business models & marketing
operations manage-ment & research
optimi-zation
process mining
visualization & visual analytics
stochastics
predictive analytics
distributed systems
databases
process science
business process management
formal methods & concurrency theory
process automation & workflow manage-ment
business process improve-ment
ACCURACY
TRANSPARENCY



	creating and managing event data	process discovery	conformance checking	performance analysis	operational support
fairness **Data Science without prejudice: How to avoid unfair conclusions even if they are true?**	The input data may be biased, incomplete or incorrect such that the analysis reconfirms prejudices. By resampling or relabeling the data, undesirable forms of discrimination can be avoided. Note that both cases and resources (used to execute activities) may refer to individuals having sensitive attributes such as race, gender, age, etc.	The discovered model may abstract from paths followed by certain under-represented groups of cases. Discrimination-aware process-discovery algorithms can be used to avoid this. For example, if cases are handled differently based on gender, we may want to ensure that both are equally represented in the model.	Conformance checking can be used to "blame" individuals, groups, or organizations for deviating from some normative model. Discrimination-aware conformance checking (e.g., alignments) needs to separate (1) likelihood, (2) severity and (3) blame. Deviations may need to be interpreted differently for different groups of cases and resources.	Straightforward performance measurements may be unfair for certain classes of cases and resources (e.g., not taking into account the context). Discrimination-aware performance analysis detects unfairness and supports process improvements taking into account trade-offs between internal fairness (worker's perspective) and external fairness (citizen/patient/customer's perspective).	Process-related predictions, recommendations and decisions may discriminate (un)intentionally. This problem can be tackled using techniques from discrimination-aware data mining.
confidentiality **Data Science that ensures confidentiality: How to answer questions without revealing secrets?**	Event data (e.g., XES files) may reveal sensitive information. Anonymization and de-identification can be used to avoid disclosure. Note that timestamps and paths may be unique and a source for re-identification (e.g., all paths are unique).	The discovered model may reveal sensitive information, especially with respect to infrequent paths or small event logs. Drilling-down from the model may need to be blocked when numbers get too small (cf. k-anonymity).	Conformance checking shows diagnostics for deviating cases and resources. Access-control is important and diagnostics need to be aggregated to avoid revealing compliance problems at the level of individuals.	Performance analysis shows bottlenecks and other problems. Linking these problems to cases and resources may disclose sensitive information.	Process-related predictions, recommendations and decisions may disclose sensitive information, e.g., based on a rejection other properties can be derived.
accuracy **Data Science without guesswork: How to answer questions with a guaranteed level of accuracy?**	Event data (e.g., XES files) may have all kinds of quality problems. Attributes may be incorrect, imprecise, or uncertain. For example, timestamps may be too coarse (just the date) or reflect the time of recording rather than the time of the event's occurrence.	Process discovery depends on many parameters and characteristics of the event log. Process models should better show the confidence level of the different parts. Moreover, additional information needs to be used better (domain knowledge, uncertainty in event data, etc.).	Often multiple explanations are possible to interpret non-conformance. Just providing one alignment based on a particular cost function may be misleading. How robust are the findings?	In case of fitness problems (process model and event log disagree), performance analysis is based on assumptions and needs to deal with missing values (making results less accurate).	Inaccurate process models may lead to flawed predictions, recommendations and decisions. Moreover, not communicating the (un)certainty of predictions, recommendations and decisions, may negatively impact processes.
transparency **Data Science that provides transparency: How to clarify answers such that they become indisputable?**	Provenance of event data is key. Ideally, process mining insights can be related to the event data they are based on. However, this may conflict with confidentiality concerns.	Discovered process models depend on the event data used as input and the parameter settings and choice of discovery algorithm. How to ensure that the process model is interpreted correctly? End-users need to understand the relation between data and model to trust analysis.	When modeled and observed behavior disagree there may be multiple explanations. How to ensure that conformance diagnostics are interpreted correctly?	When detecting performance problems, it should be clear how these were detected and what the possible causes are. Animating event logs on models helps to make problems more transparent.	Predictions, recommendations and decisions are based on process models. If possible, these models should be transparent. Moreover, explanations should be added to predictions, recommendations and decisions ("We predict that this case be late, because ...").

Warning

advertisements ahead

85.000 people joined our MOOC on process mining

Process Mining
Data Science in Action

Runs every four weeks: also used for flipped classroom!

Info MOOC: www.coursera.org/course/procmin

Wil van der Aalst

Process Mining

Data Science in Action, *Second Edition*

This is the second edition of Wil van der Aalst's seminal book on process mining, which now discusses the field also in the broader context of data science and big data approaches. It includes several additions and updates, e.g. on inductive mining techniques, the notion of alignments, a considerably expanded section on software tools and a completely new chapter on process mining in the large. It is self-contained, while at the same time covering the entire process-mining spectrum from process discovery to predictive analytics.

After a general introduction to data science and process mining in Part I, Part II provides the basics of business process modeling and data mining necessary to understand the remainder of the book. Next, Part III focuses on process discovery as the most important process mining task, while Part IV moves beyond discovering the control flow of processes, highlighting conformance checking, and organizational and time perspectives. Part V offers a guide to successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM and several commercial products. Lastly, Part VI takes a step back, reflecting on the material presented and the key open challenges.

Overall, this book provides a comprehensive overview of the state of the art in process mining. It is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers.

Features and Benefits:

- First book on process mining, bridging the gap between business process modeling and business intelligence and positioning process mining within the rapidly growing data science discipline
- This second edition includes over 150 pages of new material, e.g. on data quality, the relation to data science, inductive mining techniques and the notion of alignments
- Written by one of the most influential and most-cited computer scientists and the best-known BPM researcher
- Self-contained and comprehensive overview for a broad audience in academia and industry, including up-to-date information on tools and the exploitation of modern IT infrastructures

Computer Science

ISBN 978-3-662-49850-7

9 783662 498507

▸ springer.com

van der Aalst

Process Mining

2nd Ed.

Wil van der Aalst

Process Mining

Data Science in Action

Second Edition

Springer