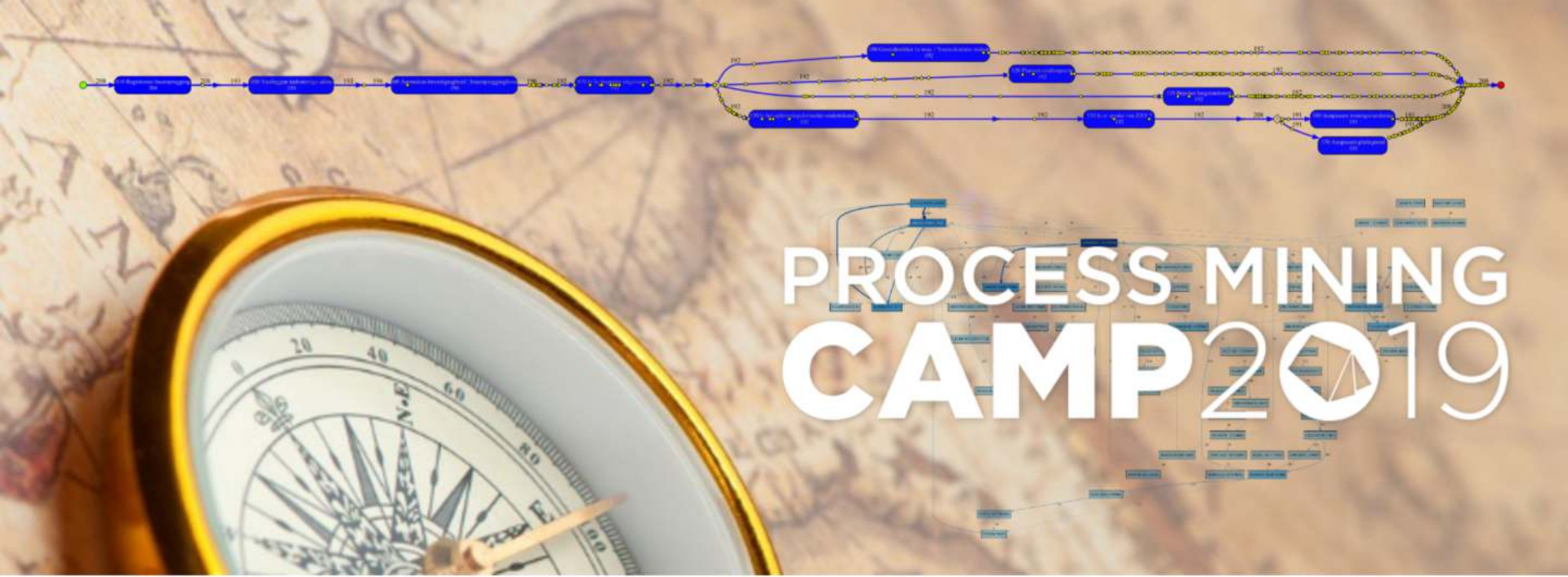


Responsible Data Science for Process Miners

prof.dr.ir. Wil van der Aalst
RWTH Aachen University
W: vdaalst.com T:@wvdaalst

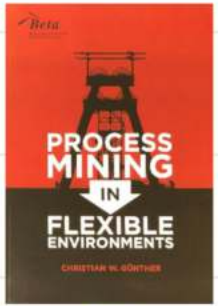
Process Mining Camp 2019, June 20th, Eindhoven

Fraunhofer FIT

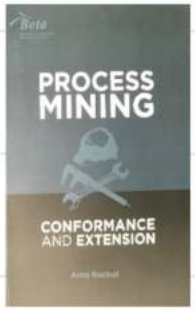
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Uptake of process mining (Scopus, June 2019)



2009



2010

start process mining at TU/e

Alpha miner and Heuristic miner

first version of ProM

first process mining company (Futura PI)

IEEE Task Force

Fluxicon

ProcessGold

Celonis

first process mining book

PM manifesto

many new process mining companies created

second process mining book

1999 2001 2003 2005 2007 2009 2011 2013 2015 2017

ICPM

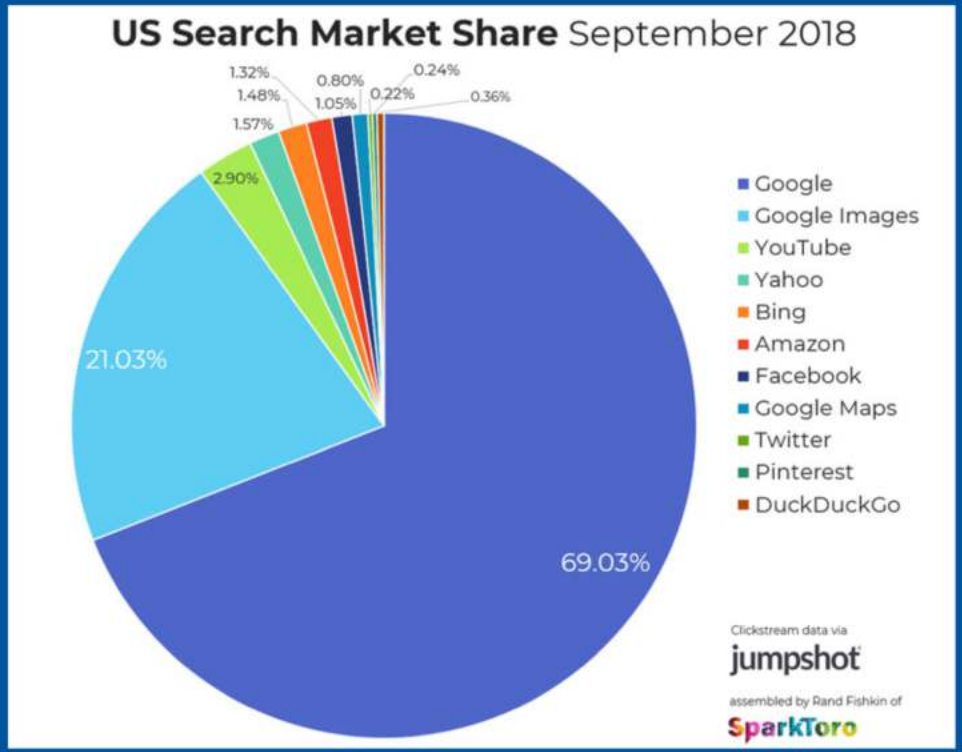
Chair of Process and Data Science

Using data in a responsible manner.

EU General Data Protection Regulation (GDPR)

(GDPR came into force on 25th May 2018.)

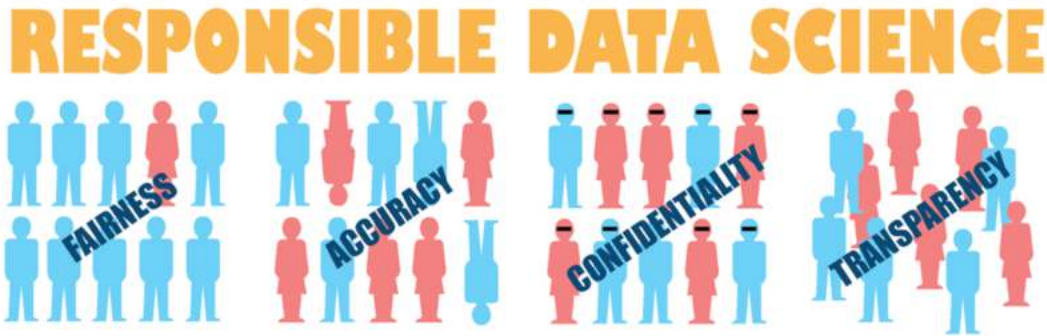
Does Google have a monopoly? The winner takes it all!



GDPR probably strengthened the monopolies of US tech companies.

The economic/ competitive perspective should have more emphasis.

Responsible Data Science initiative (2015-2018)



Focus on positive technological breakthroughs to prevent "pollution" by "bad data science".

https://redasci.org/

Chair of Process and Data Science

Fairness

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

Accuracy

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

Confidentiality

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

Chair of Process and Data Science

Transparency

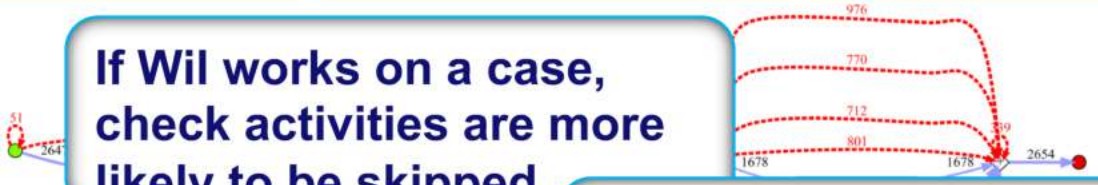
Data Science that provides transparency: How to clarify answers such that they become indisputable?


Let’s focus on fairness and
confidentiality in process mining


Fairness

Process mining can be used to identify compliance and performance problems

If Wil works on a case, check activities are more likely to be skipped.

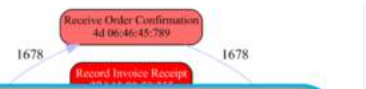


In this department, checks are performed after the legal deadline.

- mandatory activity is skipped
- activity is performed too late
- wrong order
- unauthorized resource

Process mining can be used to show the root causes of such problems, but …

Cases for this supplier tend to have many price changes.

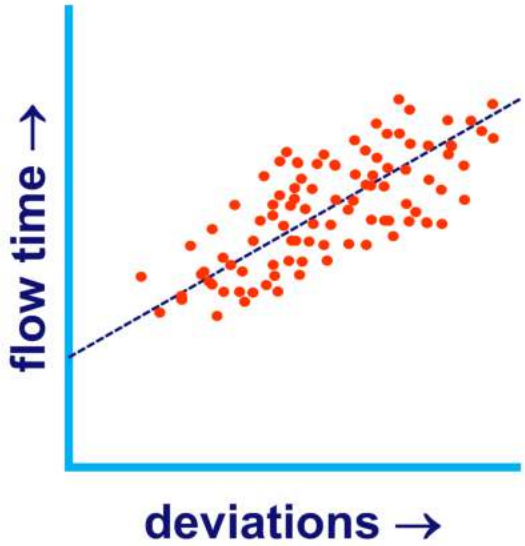


There are often delays in the back office on Fridays.

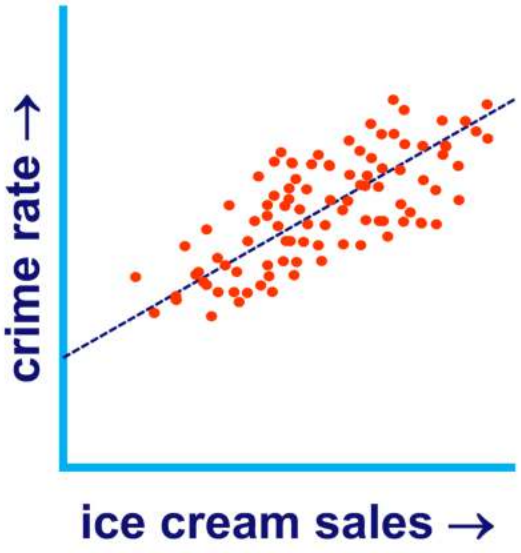
- bottlenecks and delays
- unnecessary rework
- waste
- overproduction

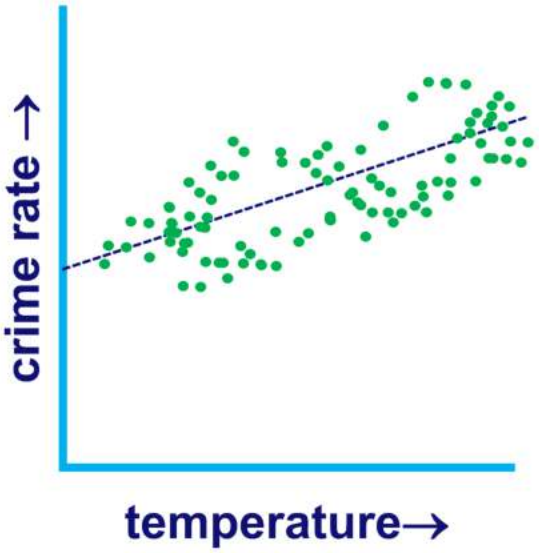
Chair of Process and Data Science

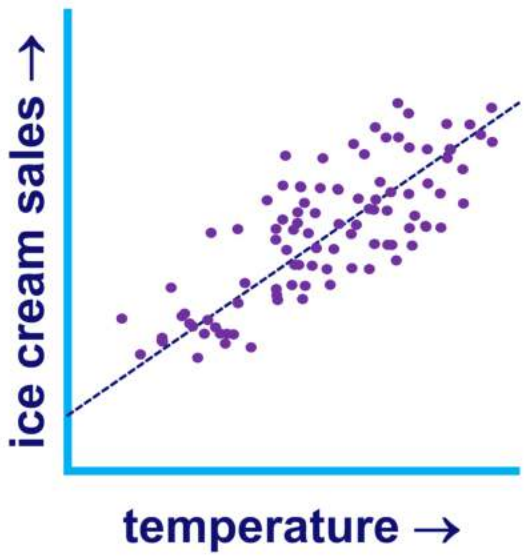
Root case analysis?



Correlation ≠ Causality







Ban ice-cream, umbrellas, etc. ?

Simpson's paradox

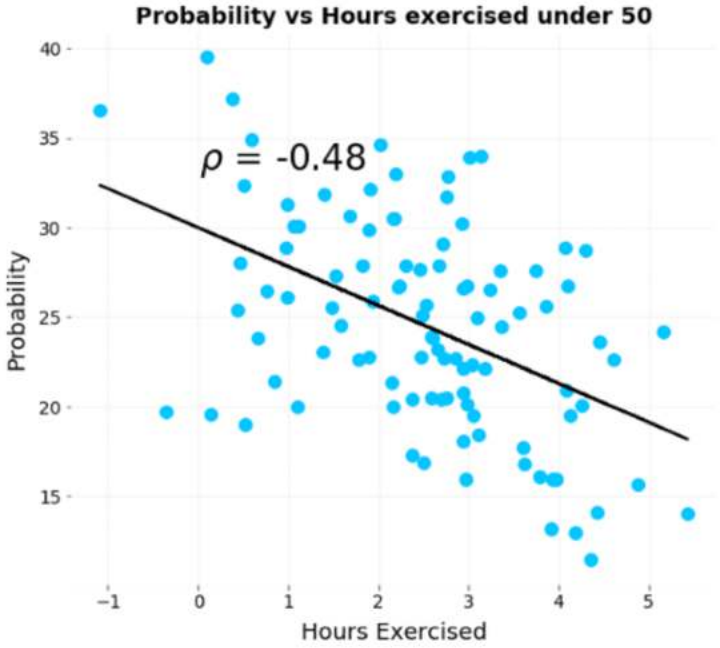
	Recommend Sophia's	Recommend Carlo's
Male	$\frac{50}{150} = 30\%$	$\frac{180}{360} = 50\%$
Female	$\frac{200}{250} = 80\%$	$\frac{36}{40} = 90\%$
Combined	$\frac{250}{400} = 62.5\%$	$\frac{216}{400} = 54\%$

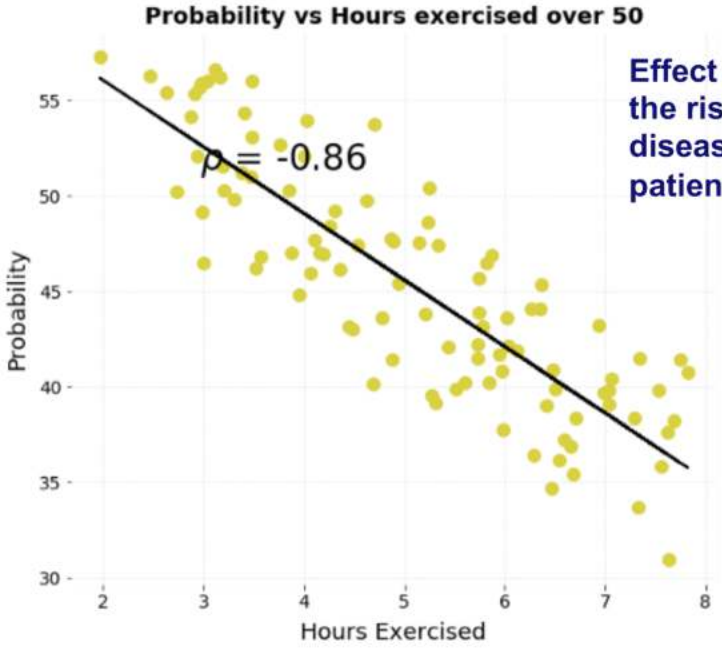
Each fraction shows the number of users who would recommend the restaurant out of the number asked. Carlo's has far more responses from men than from women while the reverse is true for Sophia's.

The data clearly show that Carlo's is preferred when the data are separated, but Sophia's is preferred when the data are combined!

Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com

Simpson's paradox (it is good to exercise)

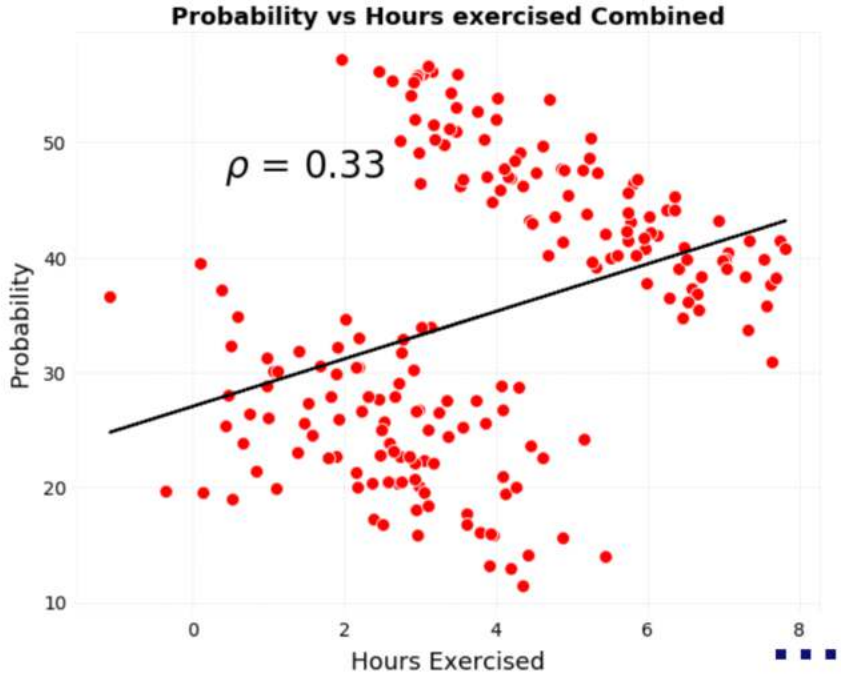




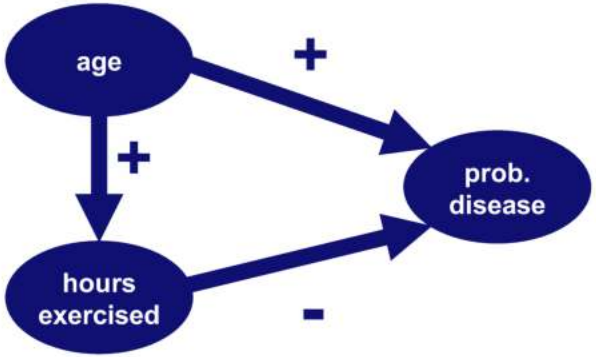
Effect of exercising on the risk of developing a disease for two sets of patients (below/over 50).

Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com

Simpson's paradox (exercising will kill you)



Effect of exercising on the risk of developing a disease for the whole group.



Simpson's Paradox: How to Prove Opposite Arguments with the Same Data by William Koehrsen towardsdatascience.com

Chair of Process and Data Science

Fairness in processes

- Don’t blame overloaded resources for causing bottlenecks.
- Don’t blame the most experienced resources taking the most difficult cases for deviating.
- Discrimination-aware data/process mining aims to avoid such errors.
- Trade-off between fairness and accuracy.

Root-case analysis in process mining

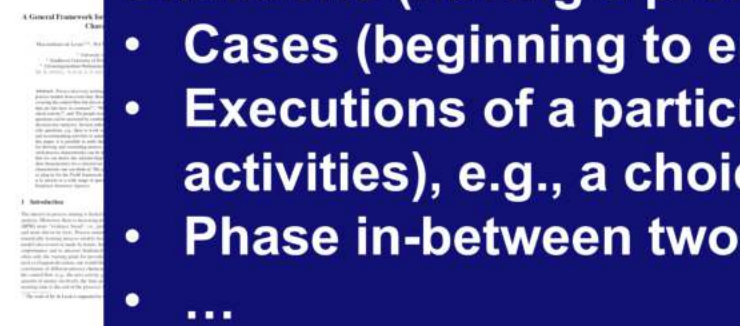
2006

20

Features:

- Target features (dependent variable)
- Descriptive features (independent variables)

2014



20

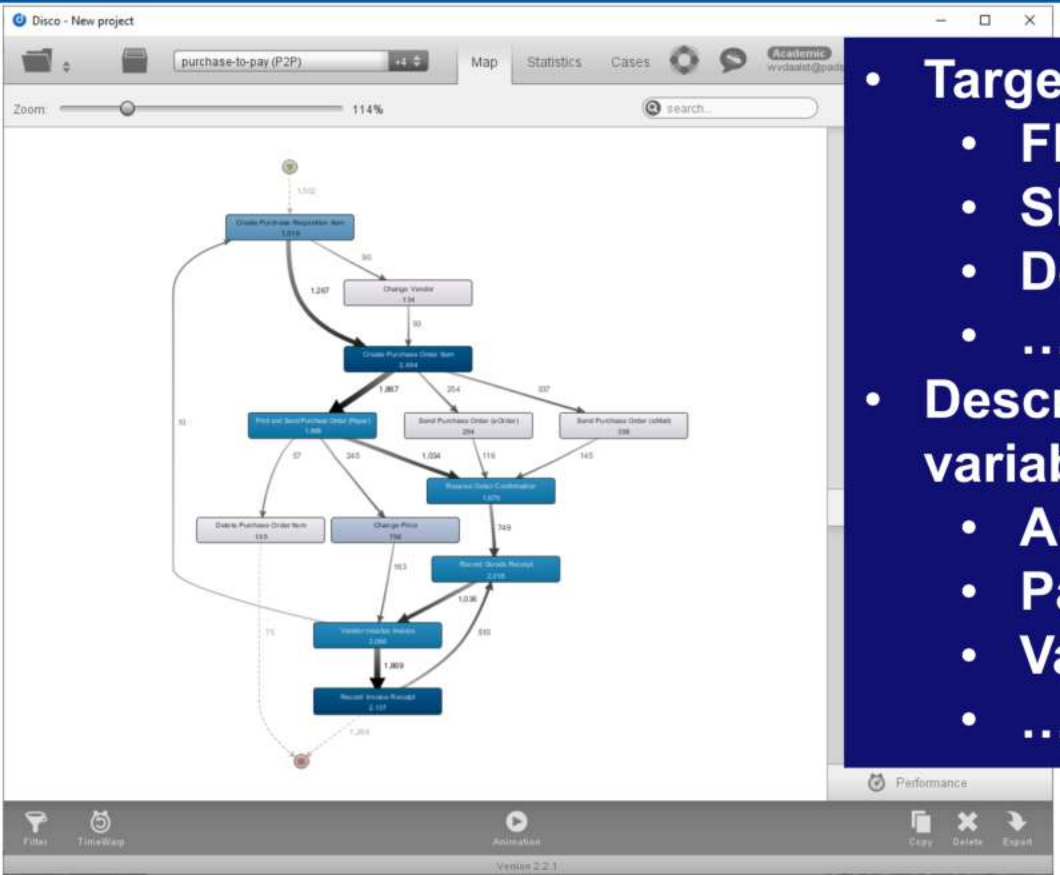
2014

2016

Situations (having a problem or not):

- Cases (beginning to end)
- Executions of a particular activity (or set of activities), e.g., a choice to skip or redo
- Phase in-between two milestone activities
- …

Situations and features



- Target features (dependent variable)
  - Flow time
  - SLA measurements
  - Deviation/non-deviation
  - …
- Descriptive features (independent variables)
  - Activity skipped
  - Particular resources
  - Value case attribute
  - …

Chair of Process and Data Science

Features corresponding to a situation

- $\mathcal{F} = \{f^0, f^1, f^2, \ldots, f^n\}$ is a set of features.
- f^0 is the target feature (the one we try to predict).
- $S = \{1, 2, \ldots, m\}$ is a set of situations.
- f_s^i is the value of feature i $(0 \leq i \leq n)$ for situation s $(1 \leq s \leq m)$.

s	f^0	f^1	f^2	…	f^n
1	f_1^0	f_1^1	f_1^2		f_1^n
2	f_2^0	f_2^1	f_2^2		f_2^n
m	f_m^0	f_m^1	f_m^2		f_m^n

Features corresponding to a situation

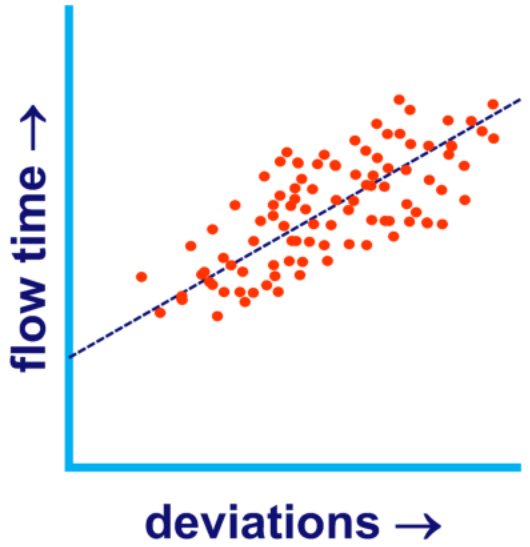
- $\mathcal{F} = \{f^0, f^1, f^2, \ldots, f^n\}$ is a set of features.
- f^0 is the target feature (the one we try to predict).
- $S = \{1, 2, \ldots, m\}$ is a set of situations.
- f_s^i is the value of feature i $(0 \leq i \leq n)$ for situation s $(1 \leq s \leq m)$.

$s = order$	f^0= flow time	f^1= team	f^2= deviations	...	f^n=shipper
435351	2.5 days	Köln	3		DHL
565452	6.4 days	Köln	0		DHL
43455	4.8 days	Aachen	2		TNT

Correlation

- $corr(f^i, f^j) \in [-1, 1]$ is the sample correlation coefficient of features i and j.
  - $corr(f^i, f^j) \approx 0$: uncorrelated
  - $corr(f^i, f^j) \gg 0$: positively correlated
  - $corr(f^i, f^j) \ll 0$: negatively correlate

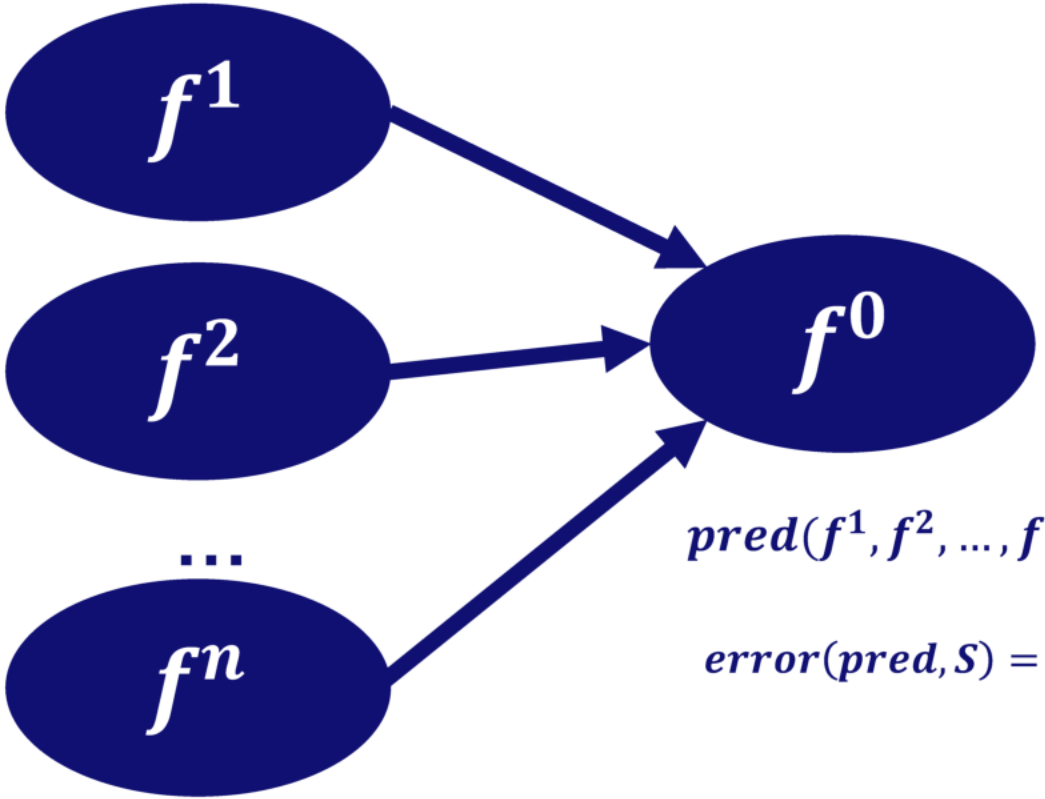
$corr(\text{deviations}, \text{flow time}) \gg 0$



Prediction technique

- $pred(f^1, f^2, \ldots, f^n)$ aims to predict the target variable f^0 in terms of the other variables.
- $error(pred, S) = \sum_{s \in S} |f_s^0 - pred(f_s^1, f_s^2, \ldots, f_s^n)| / |S|$ is the mean sample error.
- $pred$ aims to minimize the error. We can use decision trees, SVMs, (logistic) regression, neural networks, etc.
- Assume we have "the best predictor possible" (smallest mean error, also on unseen data).

Prediction error

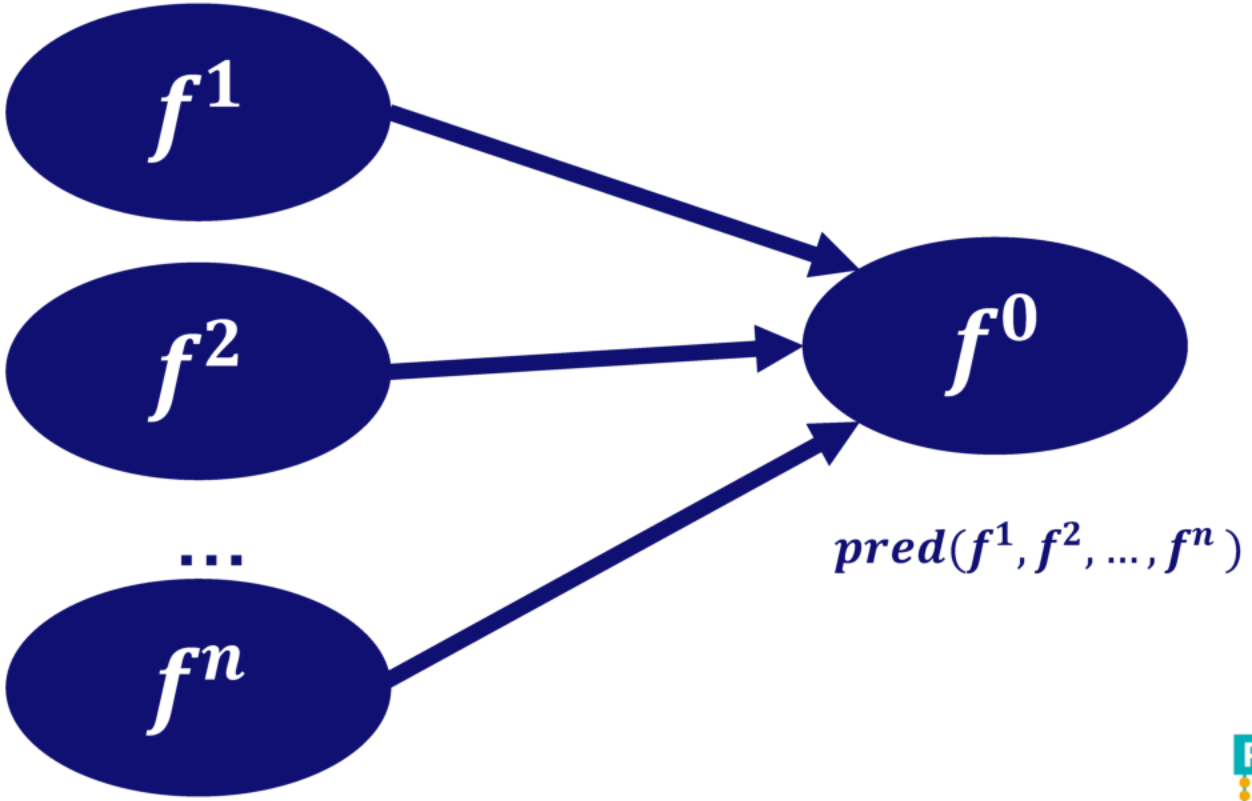


predicted value	real values	error
1.6 days	2.5 days	0.9 days
7.8 days	6.4 days	1.4 days
2.8 days	4.8 days	2.0 days

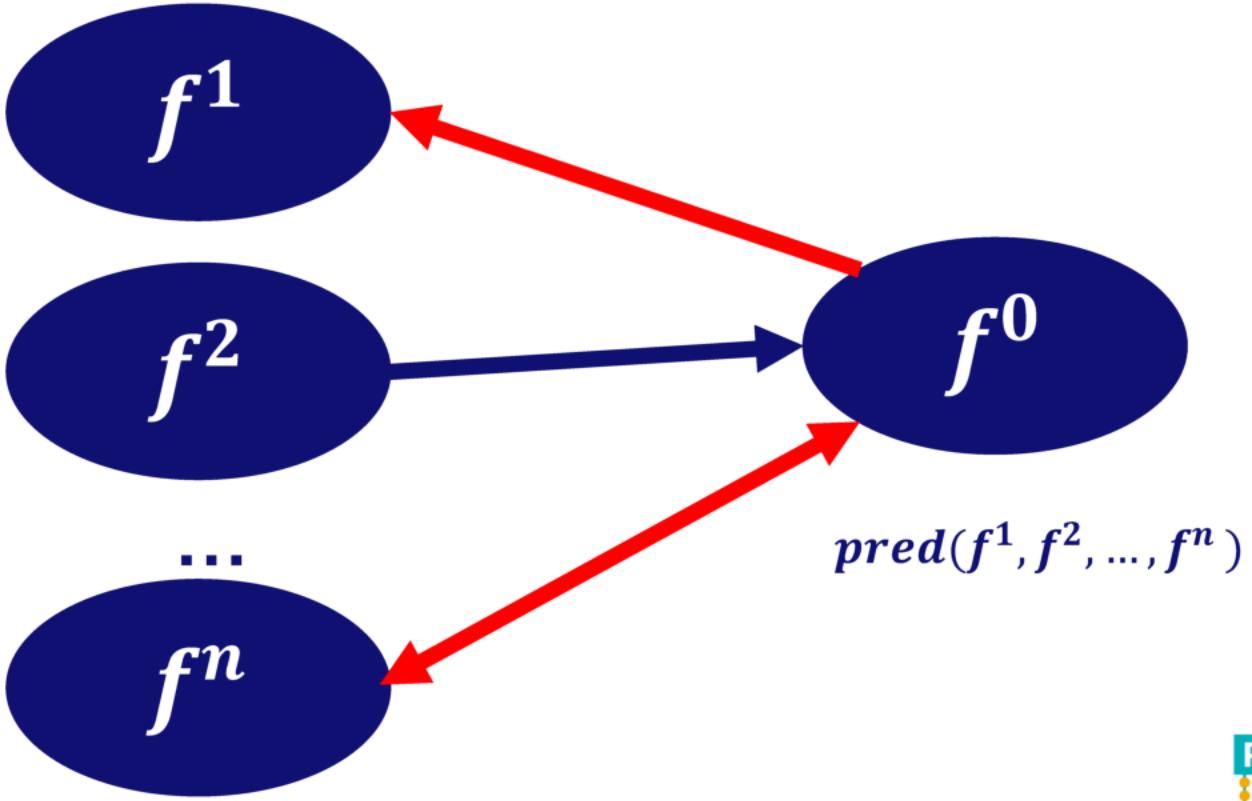
$$pred(f^1, f^2, \ldots, f^n)$$

$$error(pred, S) = \sum_{s \in S} |f_s^0 - pred(f_s^1, f_s^2, \ldots, f_s^n)| / |S|$$

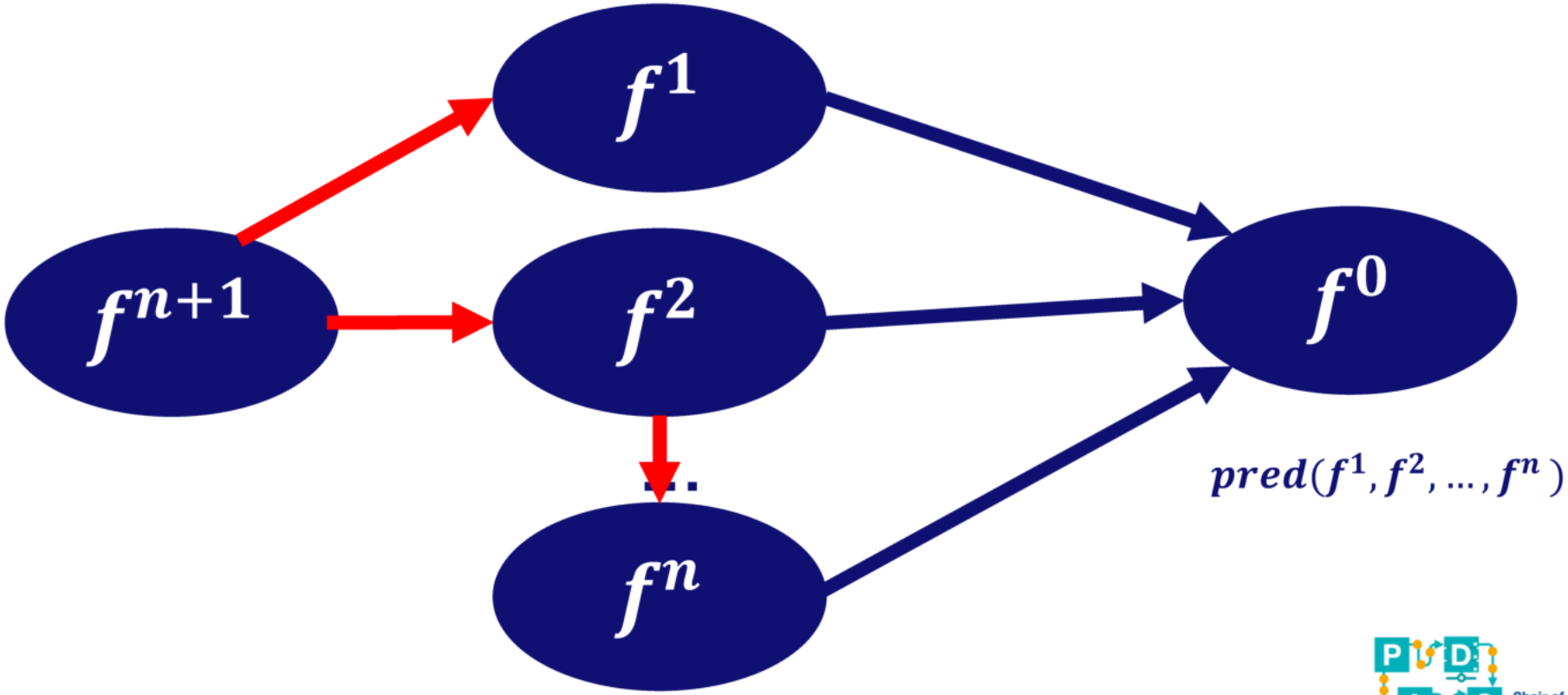
Correlation does not imply causality



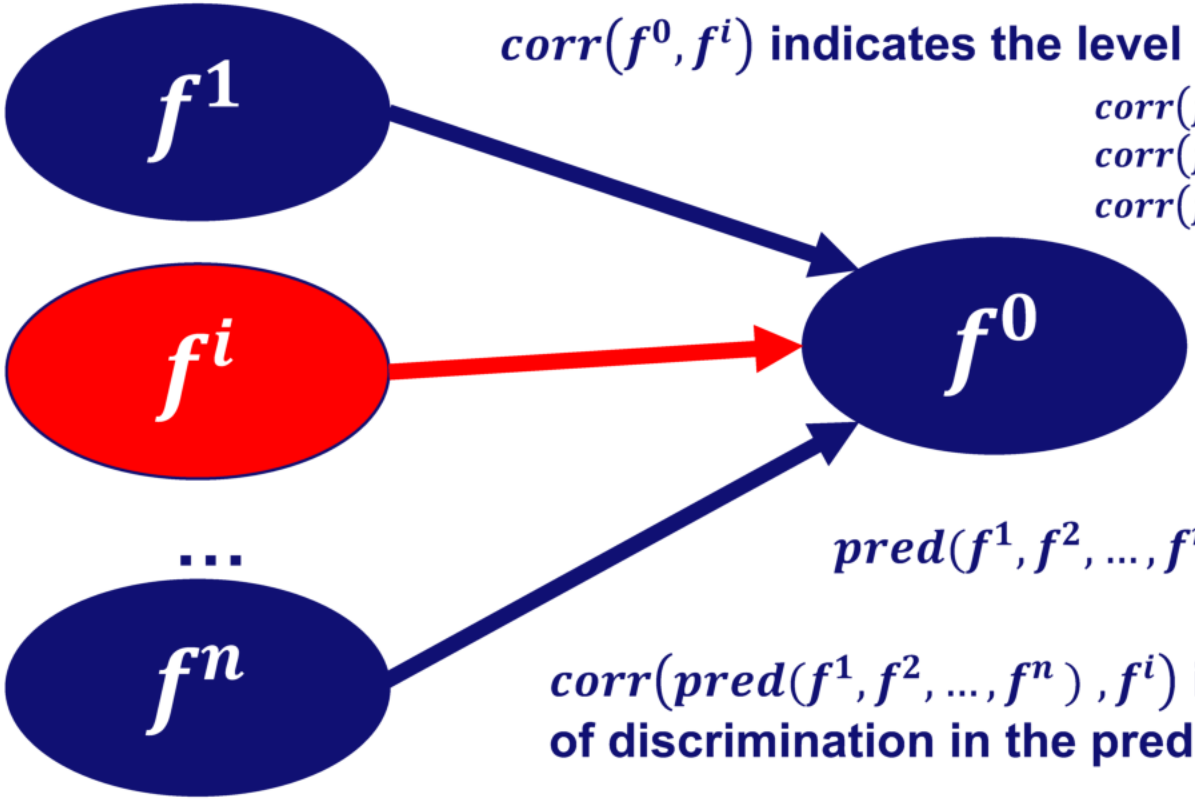
Correlation does not imply causality



Correlation does not imply causality



Fairness: f^i is the sensitive feature



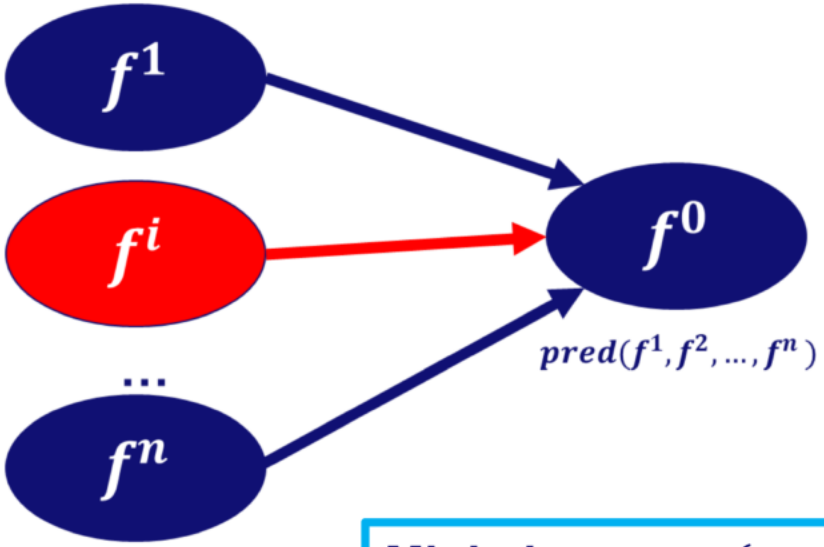
$corr(f^0, f^i)$ indicates the level of discrimination in the data

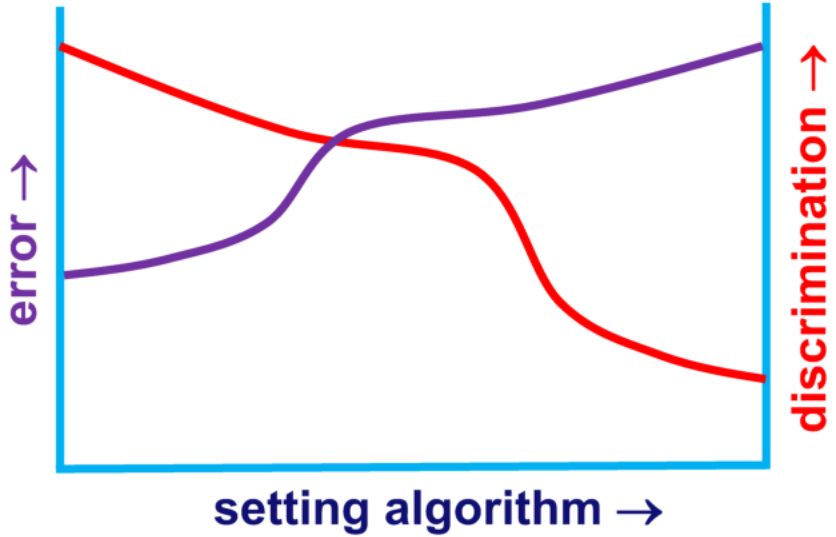
$corr(f^0, f^i) \approx 0$: no discrimination
$corr(f^0, f^i) \gg 0$: positive discrimination
$corr(f^0, f^i) \ll 0$: negative discrimination

$pred(f^1, f^2, \ldots, f^n)$

$corr(pred(f^1, f^2, \ldots, f^n), f^i)$ indicates the level of discrimination in the predictor function

Goal:
Minimize discrimination and minimize error



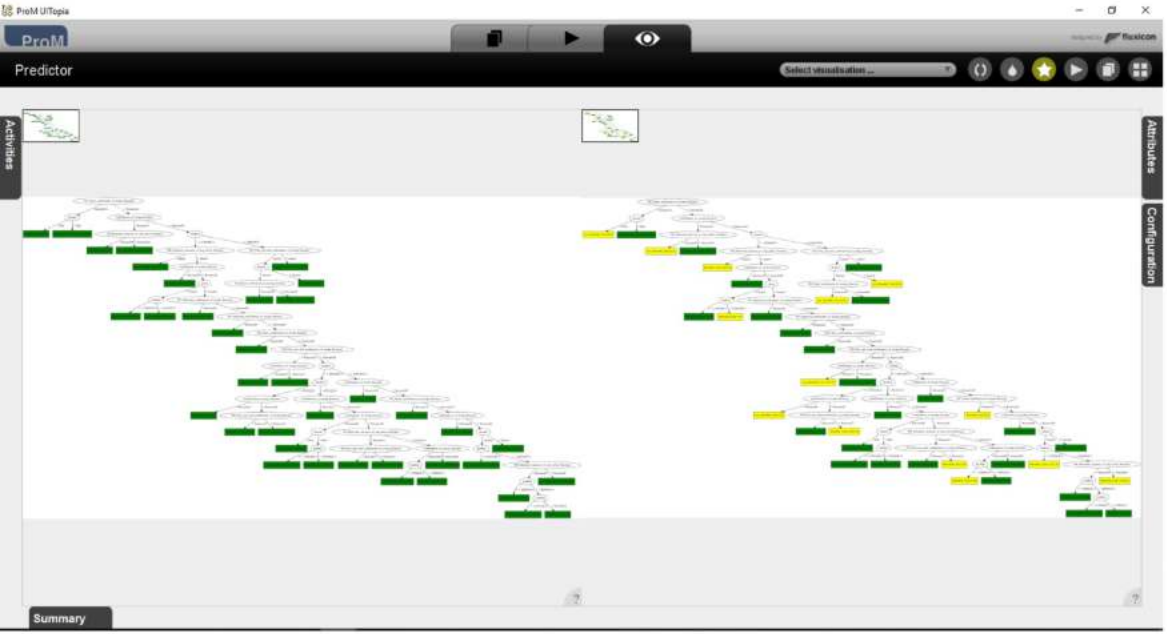


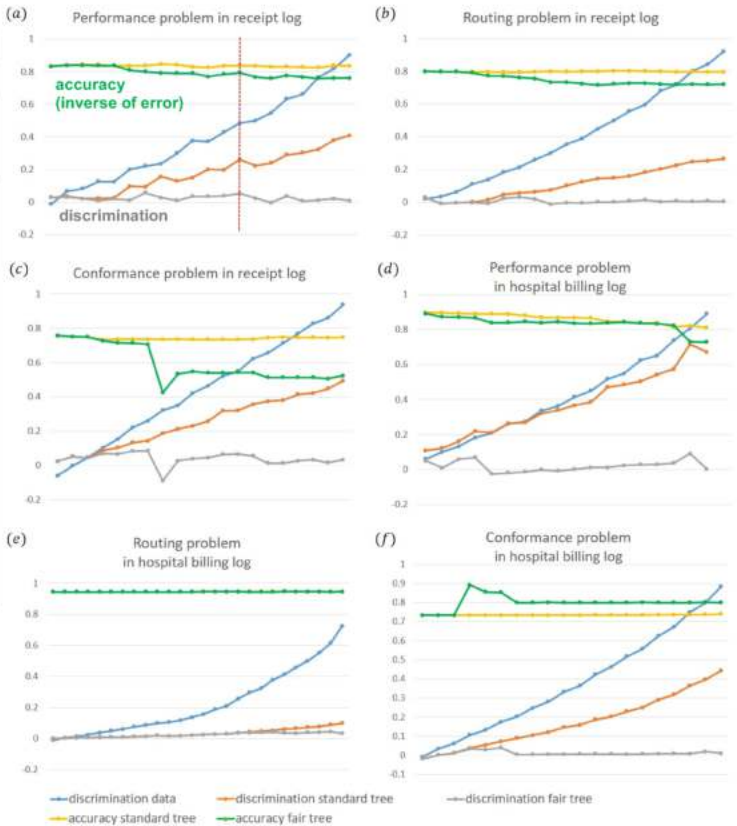
Minimize $error(pred, S)$ under the constraint

$$\left|corr\left(pred\left(f^1, f^2, \ldots, f^n\right), f^i\right)\right| \leq threshold$$

Example: Work Mahnaz Sadat Qafari

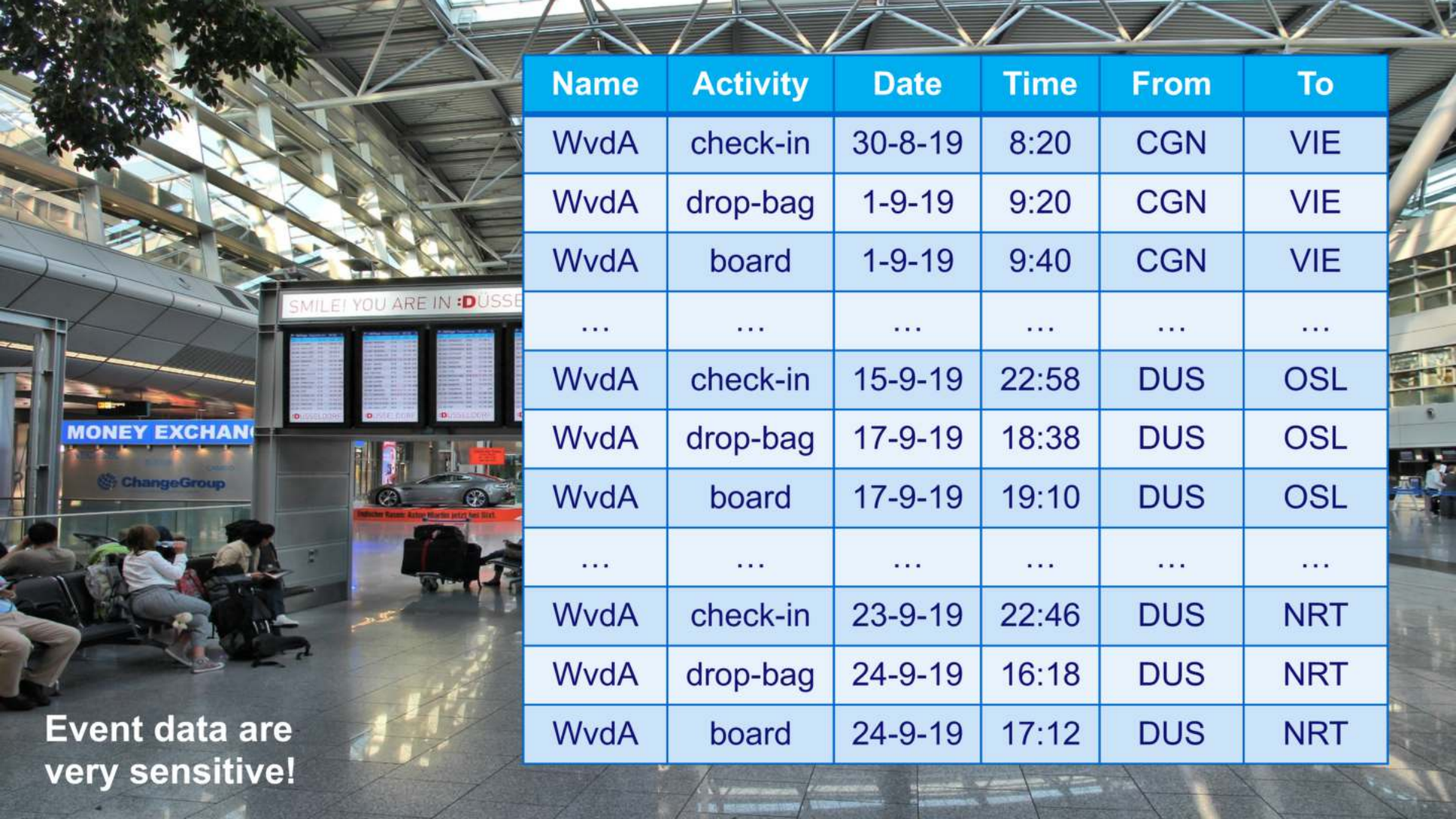
(Technique used: relabeling leaves in decision tree.)





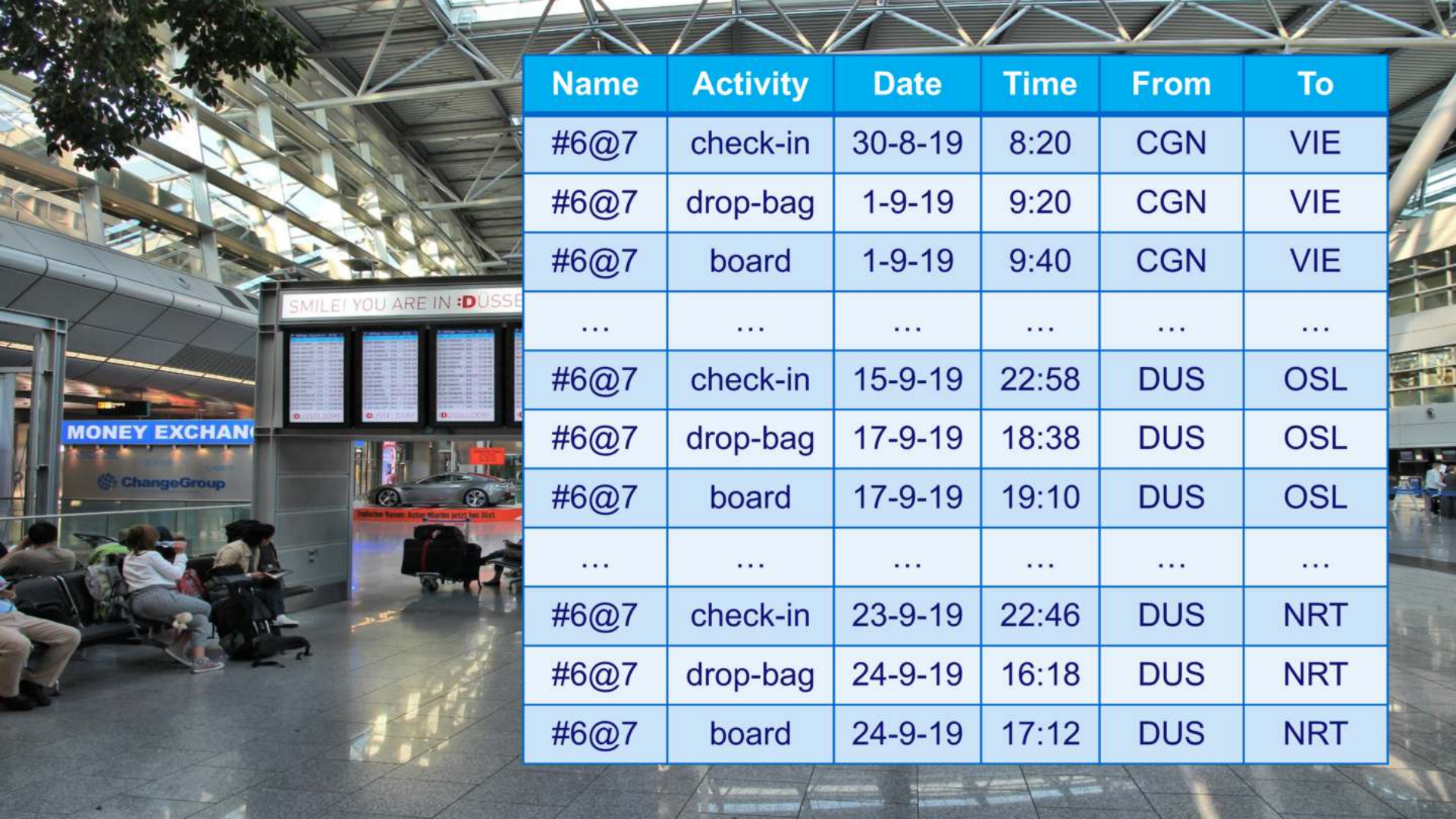
Bottom-line: We can reduce discrimination without sacrificing accuracy too much.


Confidentiality

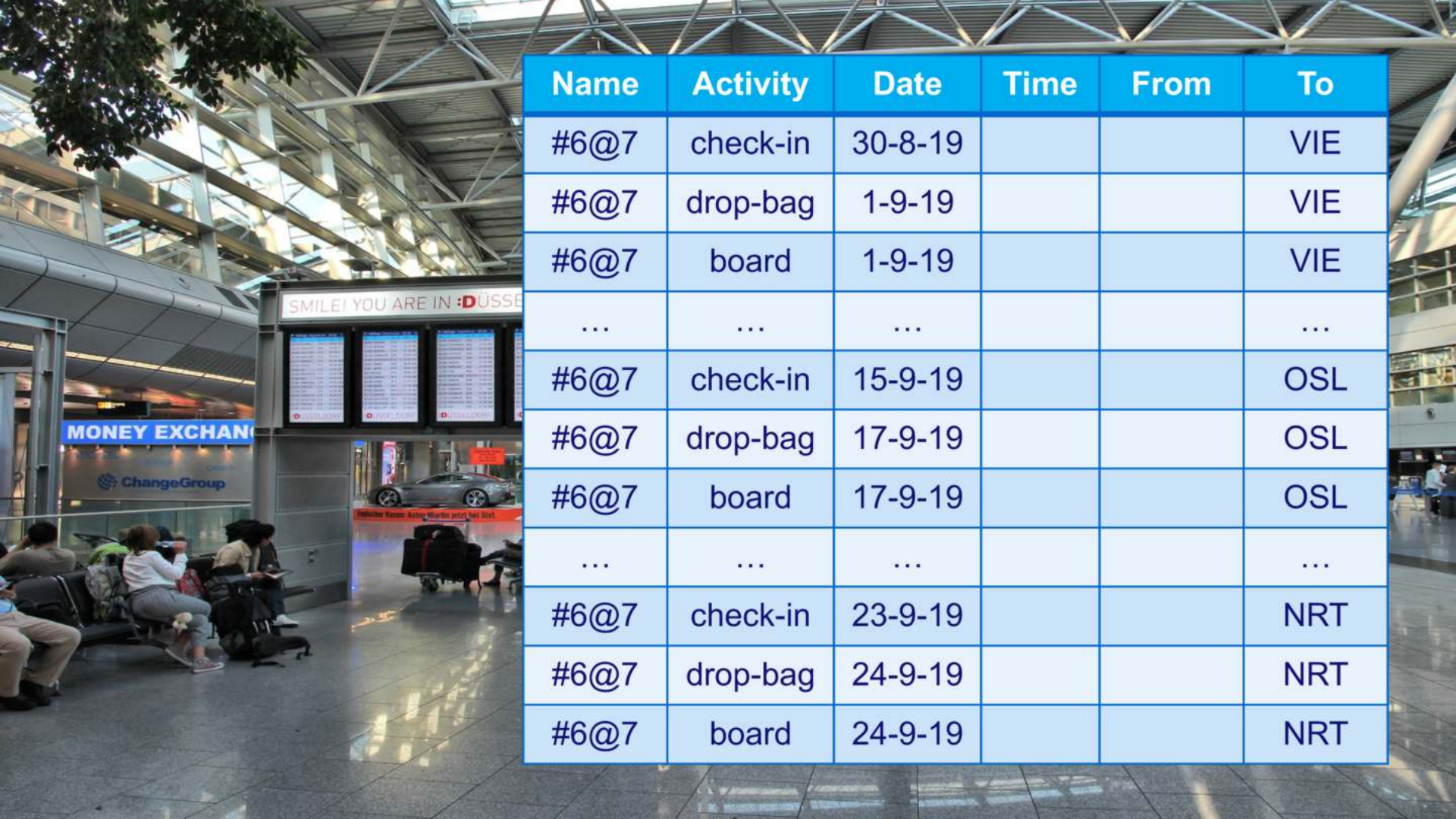


Name	Activity	Date	Time	From	To
WvdA	check-in	30-8-19	8:20	CGN	VIE
WvdA	drop-bag	1-9-19	9:20	CGN	VIE
WvdA	board	1-9-19	9:40	CGN	VIE
…	…	…	…	…	…
WvdA	check-in	15-9-19	22:58	DUS	OSL
WvdA	drop-bag	17-9-19	18:38	DUS	OSL
WvdA	board	17-9-19	19:10	DUS	OSL
…	…	…	…	…	…
WvdA	check-in	23-9-19	22:46	DUS	NRT
WvdA	drop-bag	24-9-19	16:18	DUS	NRT
WvdA	board	24-9-19	17:12	DUS	NRT

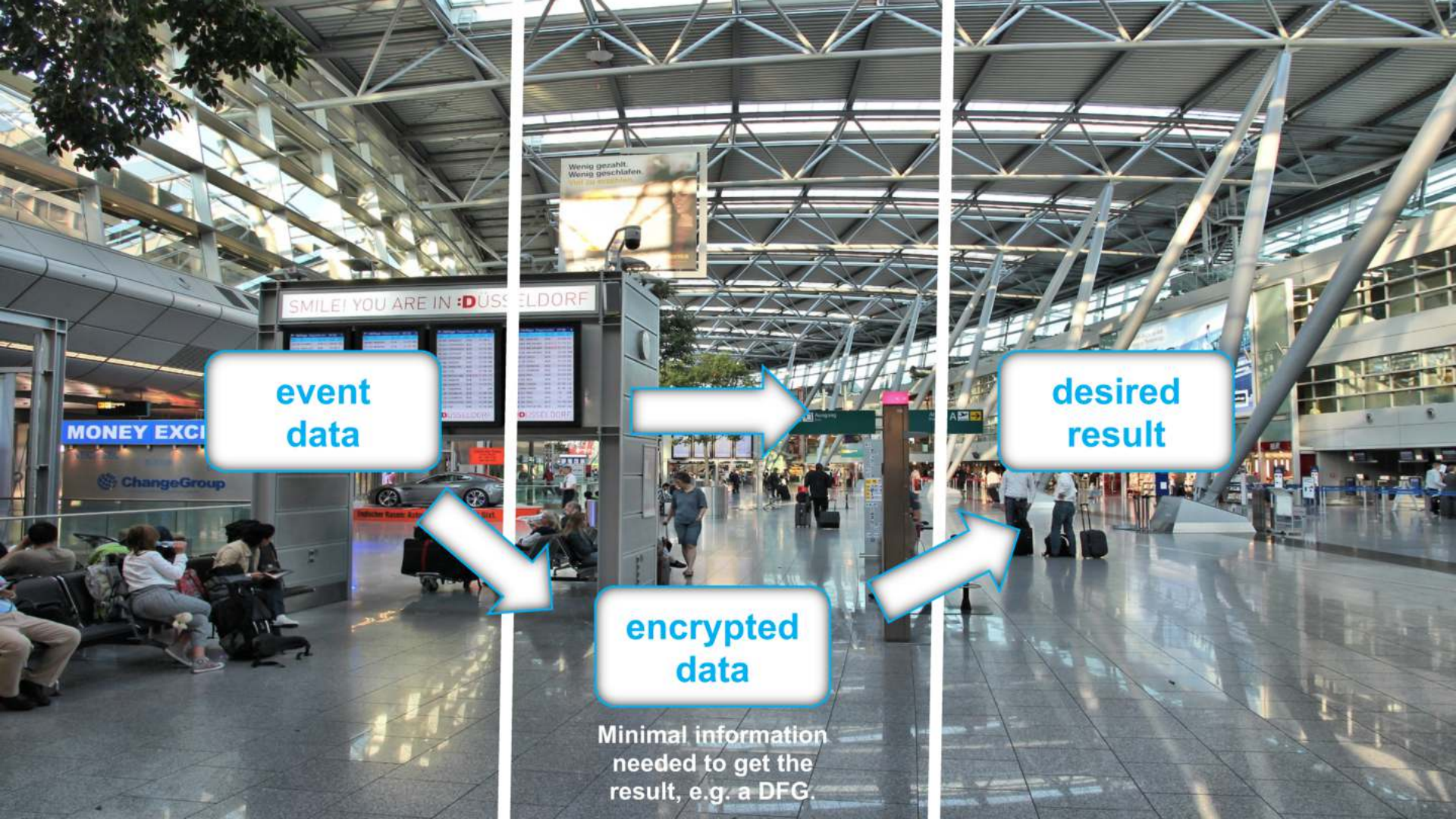
Event data are very sensitive!



Name	Activity	Date	Time	From	To
#6@7	check-in	30-8-19	8:20	CGN	VIE
#6@7	drop-bag	1-9-19	9:20	CGN	VIE
#6@7	board	1-9-19	9:40	CGN	VIE
…	…	…	…	…	…
#6@7	check-in	15-9-19	22:58	DUS	OSL
#6@7	drop-bag	17-9-19	18:38	DUS	OSL
#6@7	board	17-9-19	19:10	DUS	OSL
…	…	…	…	…	…
#6@7	check-in	23-9-19	22:46	DUS	NRT
#6@7	drop-bag	24-9-19	16:18	DUS	NRT
#6@7	board	24-9-19	17:12	DUS	NRT



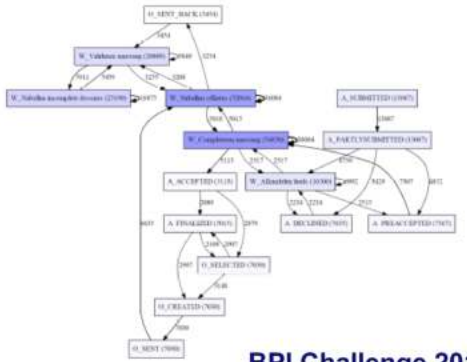
Name	Activity	Date	Time	From	To
#6@7	check-in	30-8-19			VIE
#6@7	drop-bag	1-9-19			VIE
#6@7	board	1-9-19			VIE
…	…	…			…
#6@7	check-in	15-9-19			OSL
#6@7	drop-bag	17-9-19			OSL
#6@7	board	17-9-19			OSL
…	…	…			…
#6@7	check-in	23-9-19			NRT
#6@7	drop-bag	24-9-19			NRT
#6@7	board	24-9-19			NRT


event data
encrypted data
desired result
Minimal information needed to get the result, e.g. a DFG.

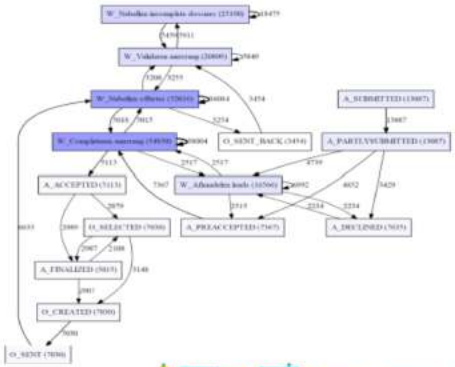
Example: Work of Majid Rafiei

Case ID	Timestamp	Activity	Resource	Cost
1	01-01-2018:08.00	Register (R)	Frank (F)	1000
2	01-01-2018:10.00	Register (R)	Frank (F)	1000
3	01-01-2018:12.10	Register (R)	Joey (J)	1000
3	01-01-2018:13.00	Verify-Documents (V)	Monica (M)	50
1	01-01-2018:13.55	Verify-Documents (V)	Paolo (P)	50
1	01-01-2018:14.57	Check-Vacancies (C)	Frank (F)	100
2	01-01-2018:15.20	Check-Vacancies (C)	Paolo (P)	100
4	01-01-2018:15.22	Register (R)	Joey (J)	1000
2	01-01-2018:16.00	Verify-Documents (V)	Frank (F)	50
2	01-01-2018:16.10	Decision (D)	Alex (A)	500
5	01-01-2018:16.30	Register (R)	Joey (J)	1000
4	01-01-2018:16.55	Check-Vacancies (C)	Monica (M)	100
1	01-01-2018:17.57	Decision (D)	Alex (A)	500
3	01-01-2018:18.20	Check-Vacancies (C)	Joey (J)	50
3	01-01-2018:19.00	Decision (D)	Alex (A)	500
4	01-01-2018:19.20	Verify-Documents (V)	Joey (J)	50
5	01-01-2018:20.00	Special-Case (S)	Katy (K)	800
5	01-01-2018:20.10	Decision (D)	Katy (K)	500
4	01-01-2018:20.55	Decision (D)	Alex (A)	500

Timestamp	Activity	Prev. Activity	Resource	Prev. Resource	Cost	Connector
08.00	R	START	Frank (F)	START	0820315	1¡@sadd21?
01.02	C	V	Frank (F)	Paolo (P)	0650900	!s*f*+dsf3
10.00	R	START	Frank (F)	START	0820315	ça/ds23"w'
15.22	R	START	Joey (J)	START	0820315	.,m;lo,mh
00.50	V	R	Monica (M)	Joey (J)	0650210	;l4;l,'kjh
00.40	V	C	Frank (F)	Paolo (P)	0650210	*';k!kjm."
12.10	R	START	Joey (J)	START	0820315	l:mj/.m @p
05.20	C	R	Paolo (P)	Frank (F)	0650900	;k;lm.lå@,
05.55	V	R	Paolo (P)	Frank (F)	0650210	=ó@k;d/f.m
00.10	D	V	Alex (A)	Frank (F)	0710155	';,lk.;hj!



≈



BPI Challenge 2012

Chair of Process and Data Science


Conclusion


process mining that adds business value
EU General Data Protection Regulation (GDPR)
responsible process mining ensuring fairness and confidentiality

International Conference on Process Mining
Aachen, June 24-26, 2019

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Supported by "all that matter in process mining".