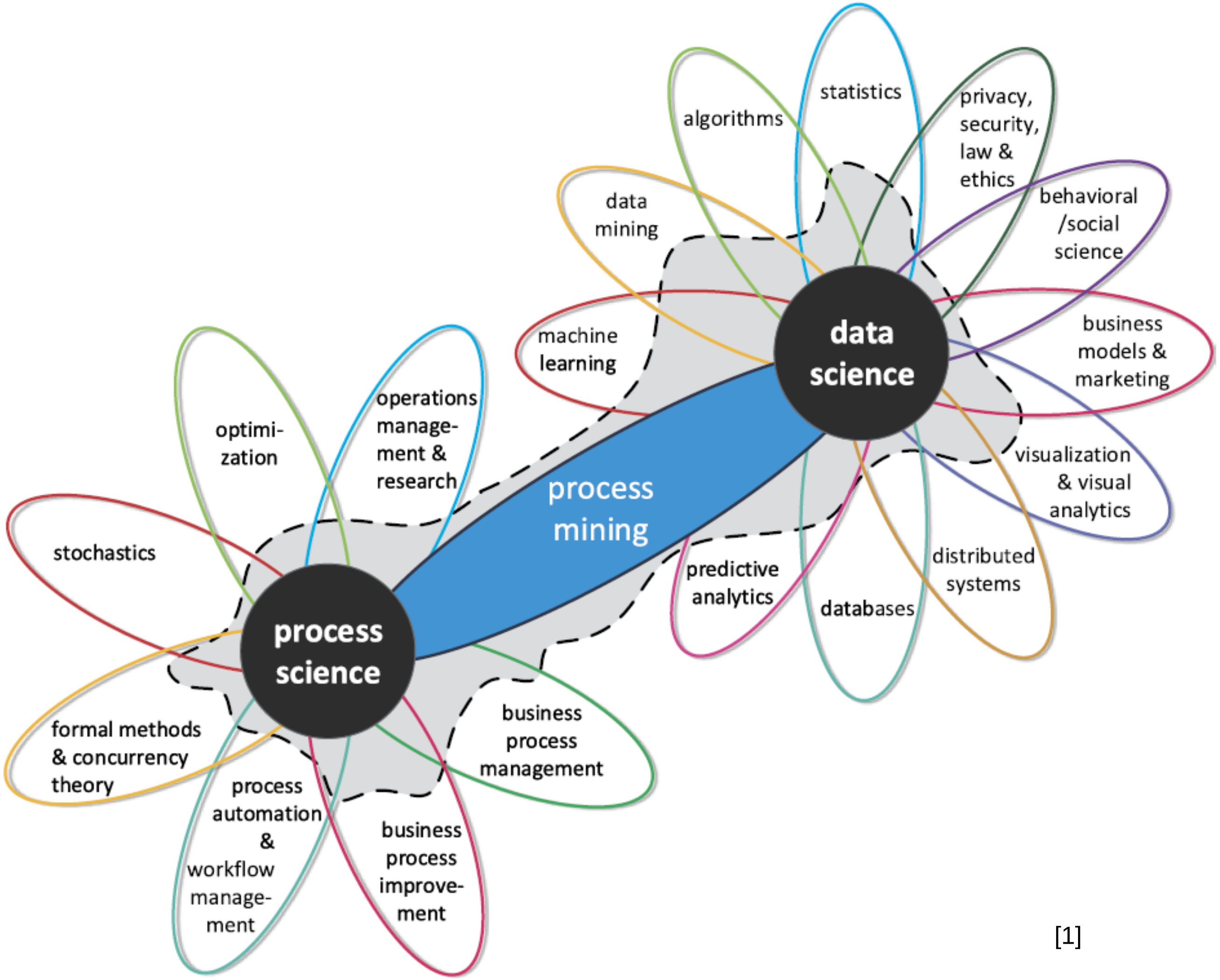
Text-Aware Predictive Monitoring of Business Processes

Marco Pegoraro, Merih Seran Uysal, David Benedikt Georgi, Wil M.P. van der Aalst

24th International Conference on Business Information Systems
June 16th 2021

Process mining

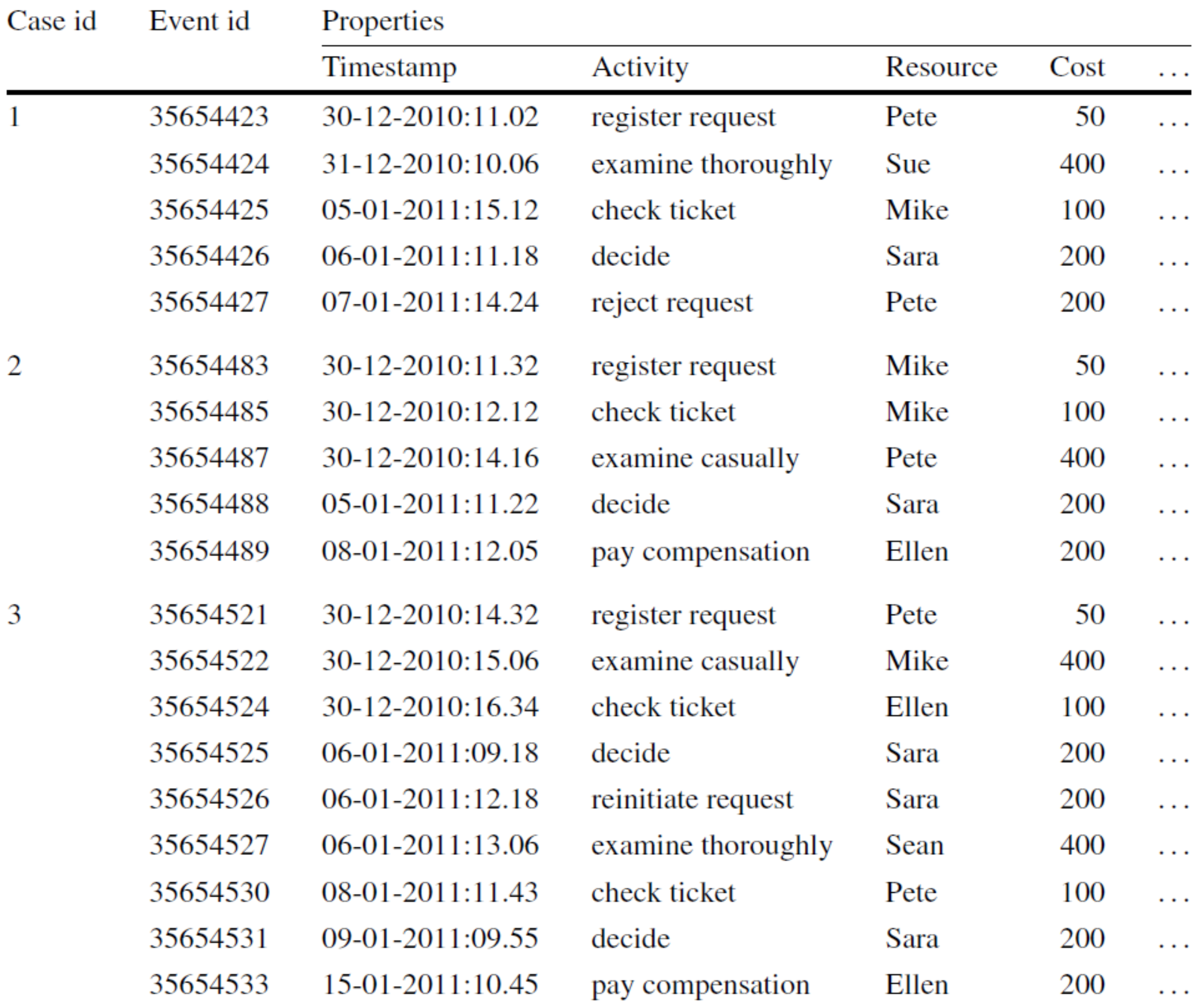

algorithms
statistics
privacy, security, law & ethics
data mining
behavioral /social science
machine learning
data science
business models & marketing
optimi-zation
operations manage-ment & research
process mining
visualization & visual analytics
stochastics
distributed systems
predictive analytics
databases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

[1]

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Event log

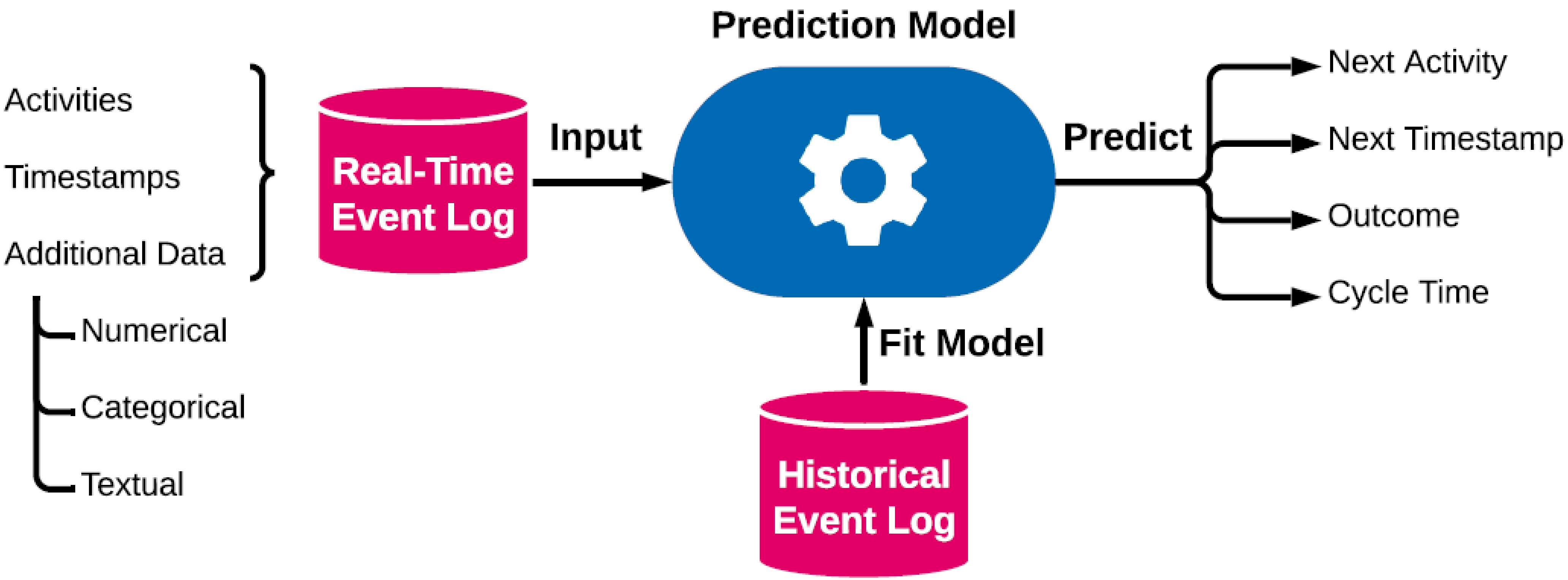


Case id	Event id	Properties				
		Timestamp	Activity	Resource	Cost	...
1	35654423	30-12-2010:11.02	register request	Pete	50	...
	35654424	31-12-2010:10.06	examine thoroughly	Sue	400	...
	35654425	05-01-2011:15.12	check ticket	Mike	100	...
	35654426	06-01-2011:11.18	decide	Sara	200	...
	35654427	07-01-2011:14.24	reject request	Pete	200	...
2	35654483	30-12-2010:11.32	register request	Mike	50	...
	35654485	30-12-2010:12.12	check ticket	Mike	100	...
	35654487	30-12-2010:14.16	examine casually	Pete	400	...
	35654488	05-01-2011:11.22	decide	Sara	200	...
	35654489	08-01-2011:12.05	pay compensation	Ellen	200	...
3	35654521	30-12-2010:14.32	register request	Pete	50	...
	35654522	30-12-2010:15.06	examine casually	Mike	400	...
	35654524	30-12-2010:16.34	check ticket	Ellen	100	...
	35654525	06-01-2011:09.18	decide	Sara	200	...
	35654526	06-01-2011:12.18	reinitiate request	Sara	200	...
	35654527	06-01-2011:13.06	examine thoroughly	Sean	400	...
	35654530	08-01-2011:11.43	check ticket	Pete	100	...
	35654531	09-01-2011:09.55	decide	Sara	200	...
	35654533	15-01-2011:10.45	pay compensation	Ellen	200	...

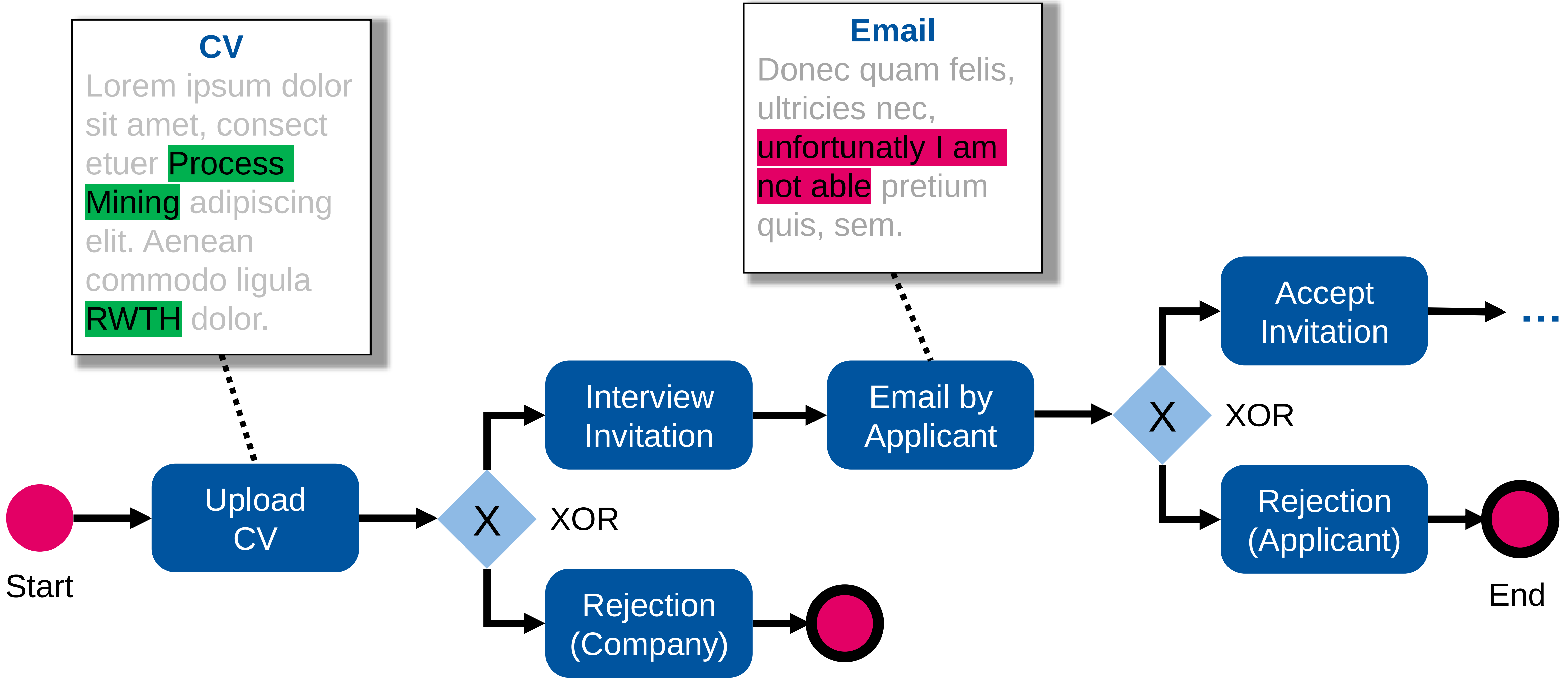
Traces

[1]

Processes: feature prediction



Text matter!



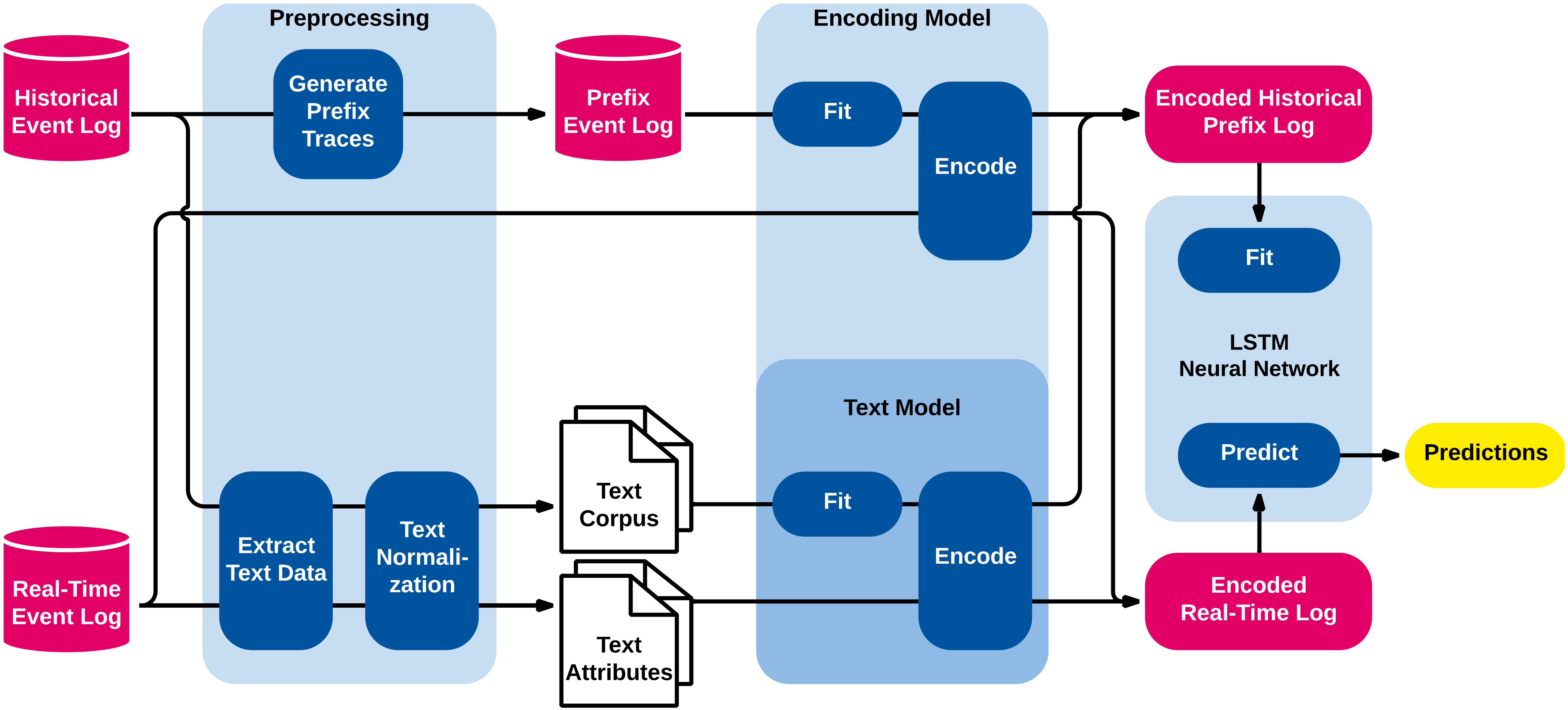
CV

Lorem ipsum dolor sit amet, consectetuer Process Mining adipiscing elit. Aenean commodo ligula RWTH dolor.

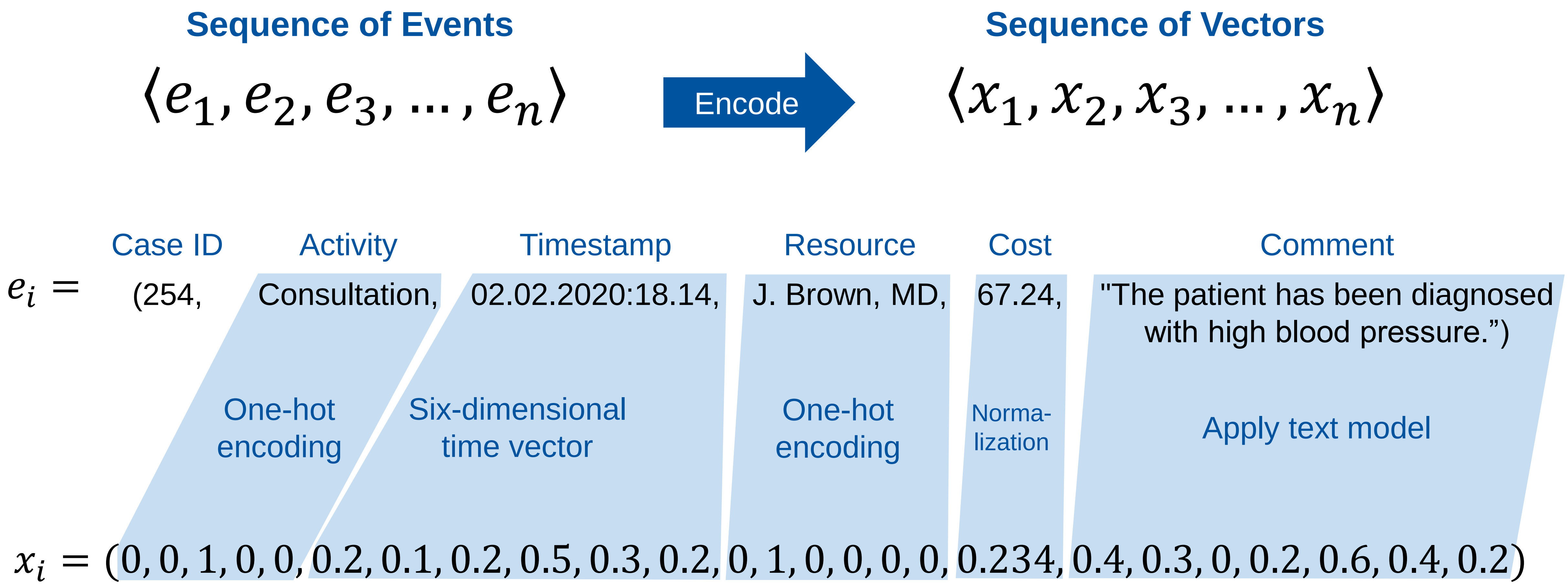
Email

Donec quam felis, ultricies nec, unfortunatly I am not able pretium quis, sem.

Exploiting free text in predictive monitoring



Exploiting free text in predictive monitoring



Sequence of Events

$$\langle e_1, e_2, e_3, \dots, e_n \rangle$$

Encode

Sequence of Vectors

$$\langle x_1, x_2, x_3, \dots, x_n \rangle$$

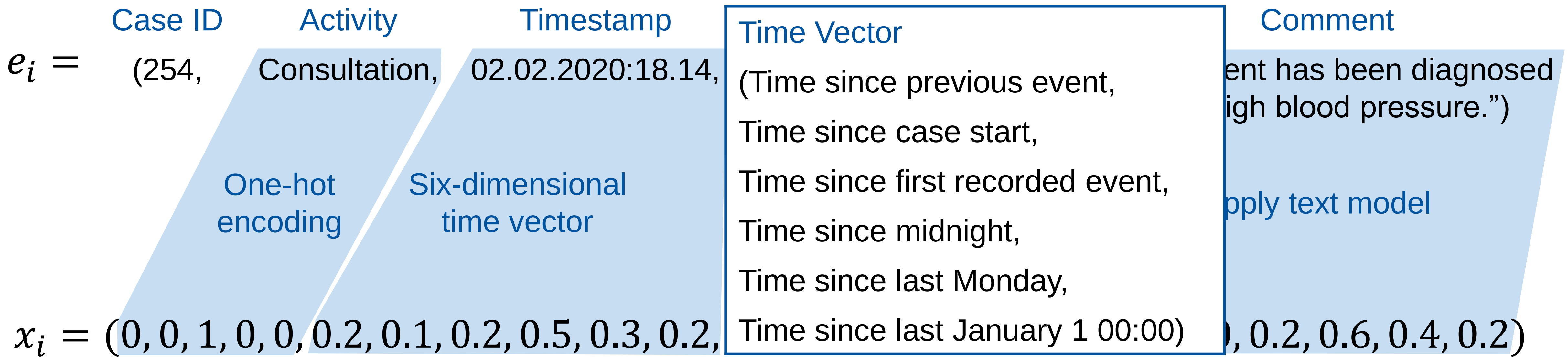
	Case ID	Activity	Timestamp	Resource	Cost	Comment
$e_i =$	(254,	Consultation,	02.02.2020:18.14,	J. Brown, MD,	67.24,	"The patient has been diagnosed with high blood pressure.")
		One-hot encoding	Six-dimensional time vector	One-hot encoding	Norma-lization	Apply text model

$$x_i = (0, 0, 1, 0, 0, 0.2, 0.1, 0.2, 0.5, 0.3, 0.2, 0, 1, 0, 0, 0, 0, 0.234, 0.4, 0.3, 0, 0.2, 0.6, 0.4, 0.2)$$

Exploiting free text in predictive monitoring

Sequence of Events → Encode → **Sequence of Vectors**

$$\langle e_1, e_2, e_3, \dots, e_n \rangle \xrightarrow{\text{Encode}} \langle x_1, x_2, x_3, \dots, x_n \rangle$$



Case ID | Activity | Timestamp | Comment

$e_i =$ (254, Consultation, 02.02.2020:18.14, ent has been diagnosed igh blood pressure.")

One-hot encoding

Six-dimensional time vector

pply text model

$x_i = (0, 0, 1, 0, 0, 0.2, 0.1, 0.2, 0.5, 0.3, 0.2,$ $), 0.2, 0.6, 0.4, 0.2)$

Time Vector

(Time since previous event,

Time since case start,

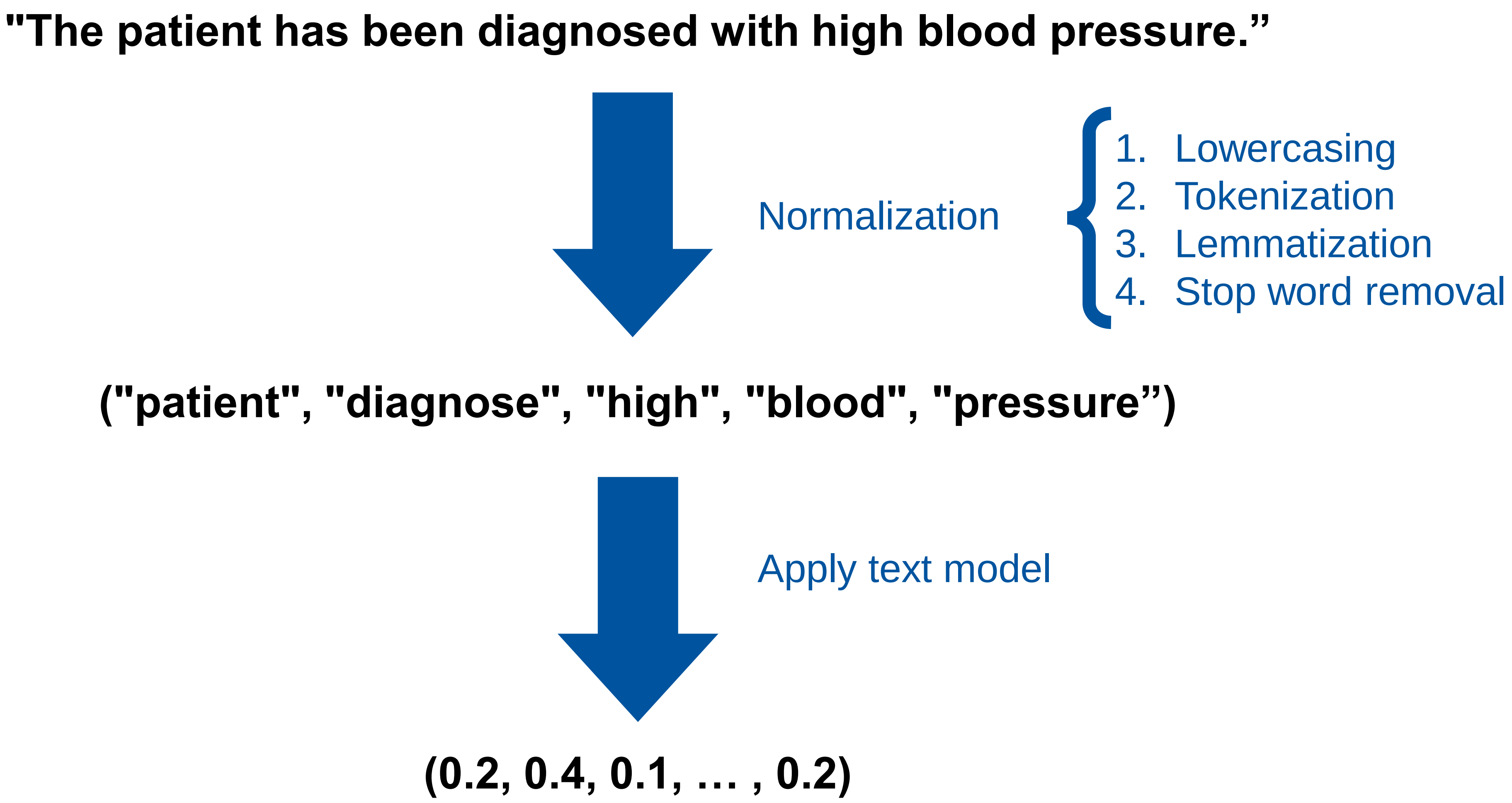
Time since first recorded event,

Time since midnight,

Time since last Monday,

Time since last January 1 00:00)

Text vectorization



"The patient has been diagnosed with high blood pressure.”

Normalization

1. Lowercasing
2. Tokenization
3. Lemmatization
4. Stop word removal

("patient", "diagnose", "high", "blood", "pressure”)

Apply text model

(0.2, 0.4, 0.1, … , 0.2)

Text vectorization models

Bag of Words

based on

word frequencies (tf-idf)

Zellig S. Harris [2]

Bag of N-Gram

based on

n-gram frequencies (tf-idf)

Peter F. Brown et al. [3]

Paragraph Vector

based on

neural network

Quoc V. Le and Tomas Mikolov [4]

Latent Dirichlet Allocation

based on

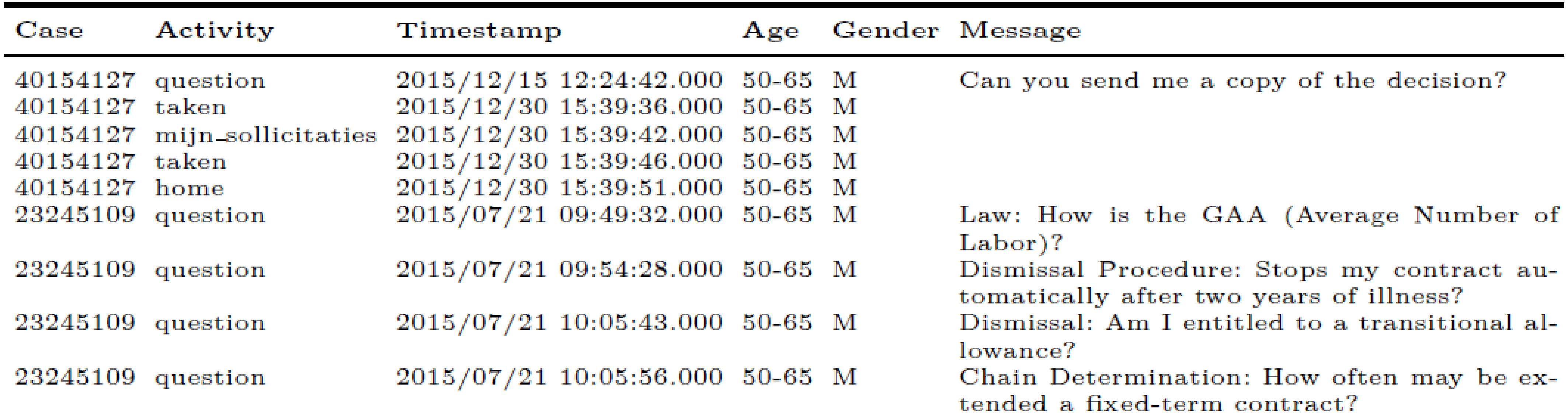
topic modeling

David M. Blei et al. [5]

Evaluation

We will validate our approach on 2 real-life logs:

- **BPI Challenge 2016: customer journey log**

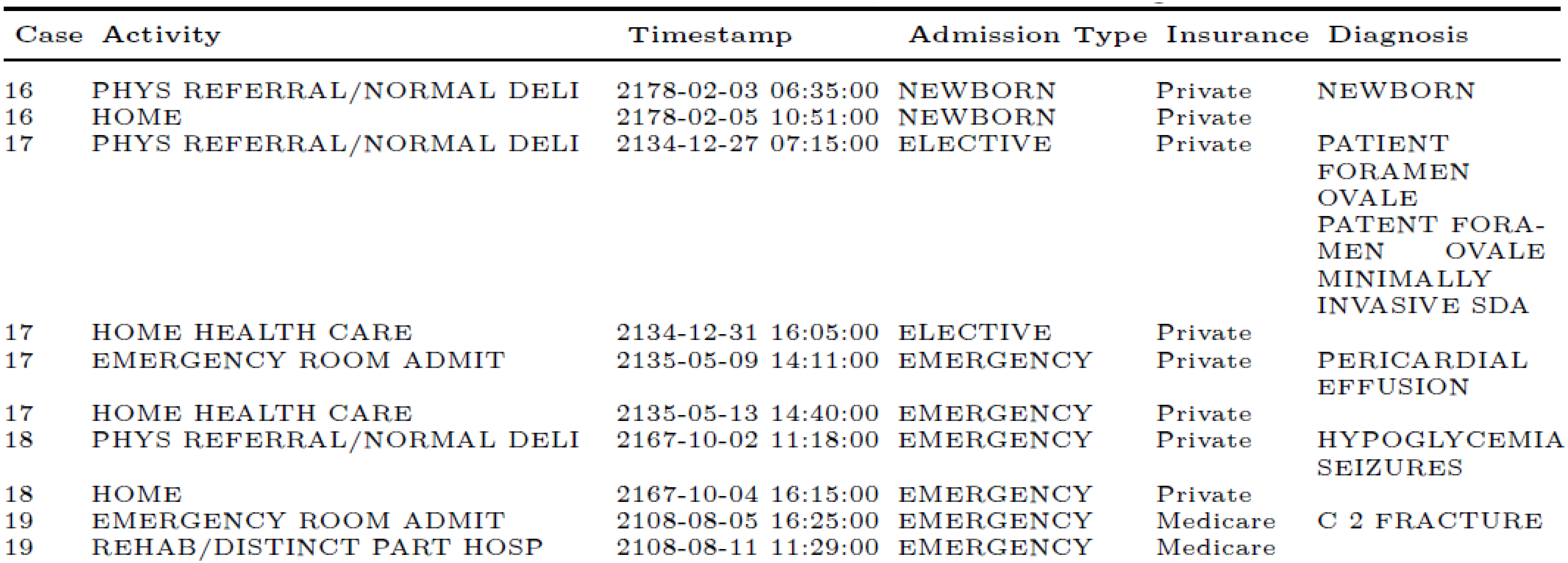


Case	Activity	Timestamp	Age	Gender	Message
40154127	question	2015/12/15 12:24:42.000	50-65	M	Can you send me a copy of the decision?
40154127	taken	2015/12/30 15:39:36.000	50-65	M	
40154127	mijn_sollicitaties	2015/12/30 15:39:42.000	50-65	M	
40154127	taken	2015/12/30 15:39:46.000	50-65	M	
40154127	home	2015/12/30 15:39:51.000	50-65	M	
23245109	question	2015/07/21 09:49:32.000	50-65	M	Law: How is the GAA (Average Number of Labor)?
23245109	question	2015/07/21 09:54:28.000	50-65	M	Dismissal Procedure: Stops my contract automatically after two years of illness?
23245109	question	2015/07/21 10:05:43.000	50-65	M	Dismissal: Am I entitled to a transitional allowance?
23245109	question	2015/07/21 10:05:56.000	50-65	M	Chain Determination: How often may be extended a fixed-term contract?

Evaluation

We will validate our approach on 2 real-life logs:

- **MIMIC-III: hospital admission log**



Case	Activity	Timestamp	Admission Type	Insurance	Diagnosis
16	PHYS REFERRAL/NORMAL DELI	2178-02-03 06:35:00	NEWBORN	Private	NEWBORN
16	HOME	2178-02-05 10:51:00	NEWBORN	Private	
17	PHYS REFERRAL/NORMAL DELI	2134-12-27 07:15:00	ELECTIVE	Private	PATIENT FORAMEN OVALE PATENT FORAMEN OVALE MINIMALLY INVASIVE SDA
17	HOME HEALTH CARE	2134-12-31 16:05:00	ELECTIVE	Private	
17	EMERGENCY ROOM ADMIT	2135-05-09 14:11:00	EMERGENCY	Private	PERICARDIAL EFFUSION
17	HOME HEALTH CARE	2135-05-13 14:40:00	EMERGENCY	Private	
18	PHYS REFERRAL/NORMAL DELI	2167-10-02 11:18:00	EMERGENCY	Private	HYPOGLYCEMIA SEIZURES
18	HOME	2167-10-04 16:15:00	EMERGENCY	Private	
19	EMERGENCY ROOM ADMIT	2108-08-05 16:25:00	EMERGENCY	Medicare	C 2 FRACTURE
19	REHAB/DISTINCT PART HOSP	2108-08-11 11:29:00	EMERGENCY	Medicare	

Evaluation

Text-Aware Process Prediction

LSTM + Text Model

Variants:
Bag of Words
Bag of N-Gram
Paragraph Vector
Latent Dirichlet Allocation

Baseline 1

LSTM Baseline

LSTM

Based on

Niek Tax et al. [6]
Nicolò Navarin et al. [7]

Baseline 2

Process Model Baseline

Annotated Transition System

Variants:
Sequence Abstraction
Bag Abstraction
Set Abstraction

Based on

Wil M. P. van der Aalst et al. [8]
Niek Tax et al. [9]

Evaluation: metrics

Classification

Weighted-average F$_1$ Score

# Classes

Class *k* vs. rest F$_1$ score

$$\mathrm{F_1\ Score} = \frac{1}{n}\sum_{k=1}^{C} \mathrm{F}_1^k \cdot n_k$$

# Instances

Instances with class *k*

Regression

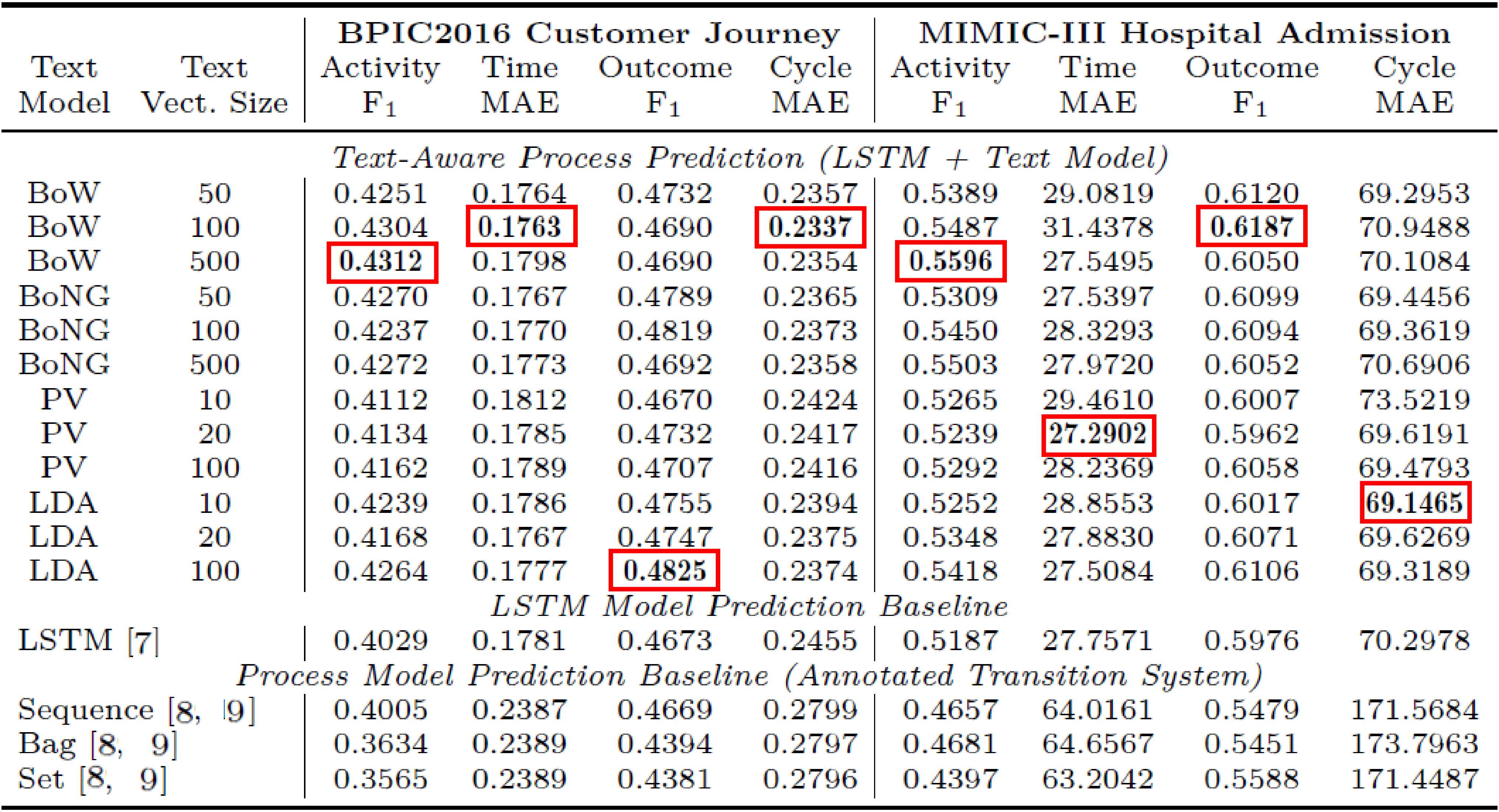
Mean Absolute Error

True value of instance *i*

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\widehat{y}_i - y_i|$$

Predicted value of instance *i*

Evaluation: results



Text Model	Text Vect. Size	BPIC2016 Customer Journey Activity F$_1$	 Time MAE	 Outcome F$_1$	 Cycle MAE	MIMIC-III Hospital Admission Activity F$_1$	 Time MAE	 Outcome F$_1$	 Cycle MAE
Text-Aware Process Prediction (LSTM + Text Model)									
BoW	50	0.4251	0.1764	0.4732	0.2357	0.5389	29.0819	0.6120	69.2953
BoW	100	0.4304	**0.1763**	0.4690	**0.2337**	0.5487	31.4378	**0.6187**	70.9488
BoW	500	**0.4312**	0.1798	0.4690	0.2354	**0.5596**	27.5495	0.6050	70.1084
BoNG	50	0.4270	0.1767	0.4789	0.2365	0.5309	27.5397	0.6099	69.4456
BoNG	100	0.4237	0.1770	0.4819	0.2373	0.5450	28.3293	0.6094	69.3619
BoNG	500	0.4272	0.1773	0.4692	0.2358	0.5503	27.9720	0.6052	70.6906
PV	10	0.4112	0.1812	0.4670	0.2424	0.5265	29.4610	0.6007	73.5219
PV	20	0.4134	0.1785	0.4732	0.2417	0.5239	**27.2902**	0.5962	69.6191
PV	100	0.4162	0.1789	0.4707	0.2416	0.5292	28.2369	0.6058	69.4793
LDA	10	0.4239	0.1786	0.4755	0.2394	0.5252	28.8553	0.6017	**69.1465**
LDA	20	0.4168	0.1767	0.4747	0.2375	0.5348	27.8830	0.6071	69.6269
LDA	100	0.4264	0.1777	**0.4825**	0.2374	0.5418	27.5084	0.6106	69.3189
LSTM Model Prediction Baseline									
LSTM [7]		0.4029	0.1781	0.4673	0.2455	0.5187	27.7571	0.5976	70.2978
Process Model Prediction Baseline (Annotated Transition System)									
Sequence [8, 9]		0.4005	0.2387	0.4669	0.2799	0.4657	64.0161	0.5479	171.5684
Bag [8, 9]		0.3634	0.2389	0.4394	0.2797	0.4681	64.6567	0.5451	173.7963
Set [8, 9]		0.3565	0.2389	0.4381	0.2796	0.4397	63.2042	0.5588	171.4487

Conclusion

- Text carries information that have a positive effect in predictive monitoring

- On short text, simple models such as BoW perform well (order have a smaller impact)

- To be addressed:
  - –Black box: decisions are not transparent and interpretable
  - –Confidentiality issues: anonymizing text is challenging

Marco Pegoraro
pegoraro@pads.rwth-aachen.de
www.mpegoraro.net

[1] Wil M.P. van der Aalst. “Process Mining: Data science in action." Springer.

[2] Zellig S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

[3] Peter F. Brown, Vincent J. Della Pietra, Peter V. Desouza, Jennifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4):467–480, 1992.

[4] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196. JMLR.org, 2014.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[6] Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. Predictive business process monitoring with LSTM neural networks. In Eric Dubois and Klaus Pohl, editors, *Advanced Information Systems Engineering - 29th International Conference, CAiSE 2017, Essen, Germany, June 12-16, 2017, Proceedings*, volume 10253 of *Lecture Notes in Computer Science*, pages 477–492. Springer, 2017.

[7] Nicolò Navarin, Beatrice Vincenzi, Mirko Polato, and Alessandro Sperduti. LSTM networks for data-aware remaining time prediction of business process instances. In *2017 IEEE Symposium Series on Computational Intelligence, SSCI 2017, Honolulu, HI, USA, November 27 - Dec. 1, 2017*, pages 1–7. IEEE, 2017.

[8] Wil M. P. van der Aalst, M. H. Schonenberg, and Minseok Song. Time prediction based on process mining. *Inf. Syst.*, 36(2):450–475, 2011.

[9] Niek Tax, Irene Teinemaa, and Sebastiaan J. van Zelst. An interdisciplinary comparison of sequence modeling methods for next-element prediction. *Softw. Syst. Model.*, 19(6):1345–1365, 2020.