Stochastically Known Logs

Avigdor Gal
Faculty of Data & Decision Sciences
Technion – Israel Institute of Technology

The Machine Take on Things

Probability Matrix

Activity:	Probability:
cut tomato	0.985865890979767
place tomato into bowl	0.00171934533864
cut cheese	0.00106682535260
place cheese into bowl	0.00030539551516
cut lettuce	0.00149893620982
place lettuce into bowl	0.00126593722961
add salt	0.00064346694853
add vinegar	0.00042187844519
add oil	0.00051470356993
add pepper	0.00027871355996
mix dressing	0.00070176448207
peel cucumber	0.00168040674179
cut cucumber	0.00144467677455
place cucumber into bowl	0.00084491760935
add dressing	0.00009947862417
mix ingredients	0.00029063280089
serve salad onto plate	0.00058529624948
action start	0.00040305560105
action end	0.00042604035115

The Machine Take on Things

Probability Matrix

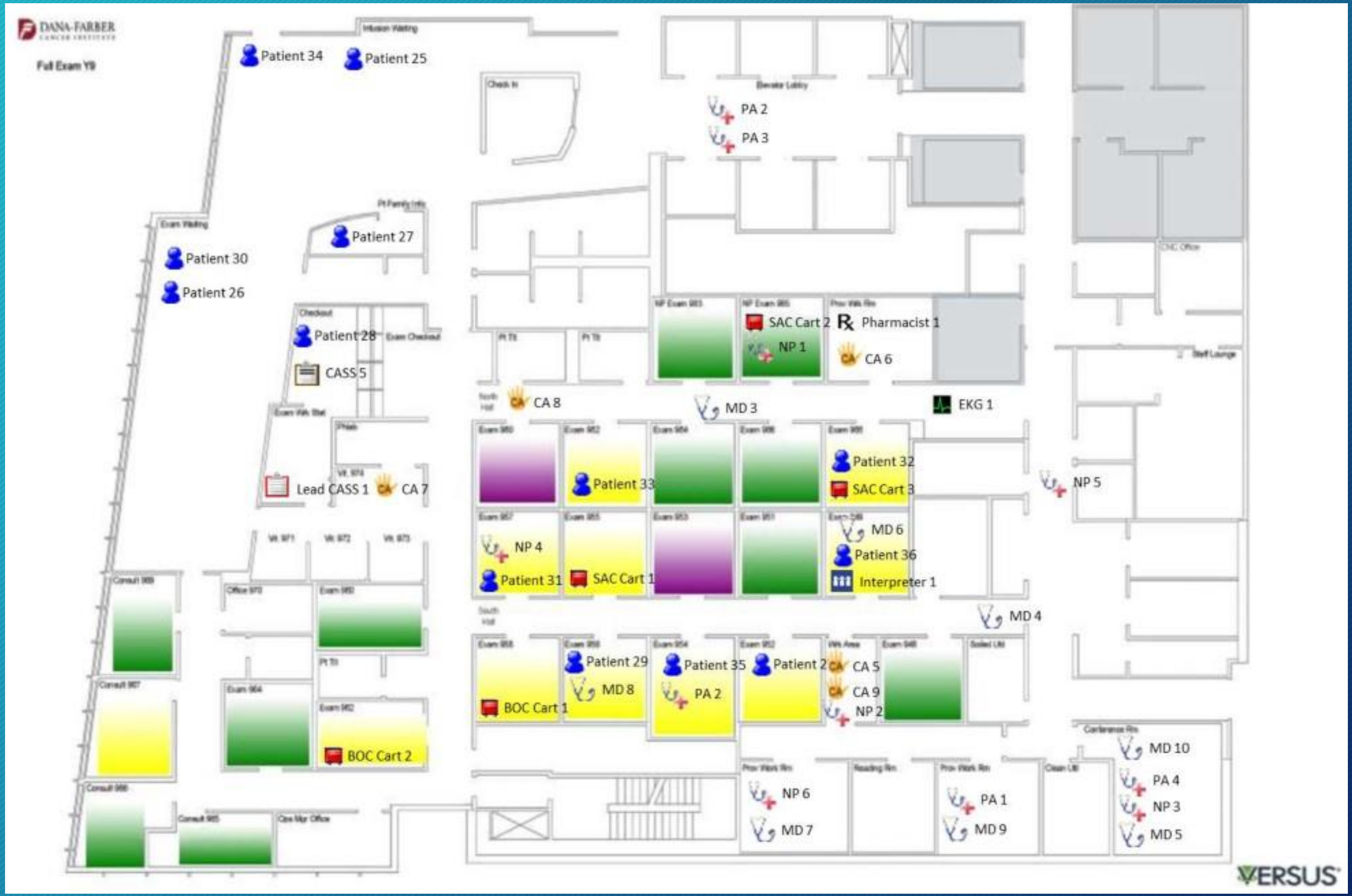
Activity:	Probability:
cut tomato	0.00074690836481
place tomato into bowl	0.978379130363464
cut cheese	0.01348170731216
place cheese into bowl	0.00250011822208
cut lettuce	0.00042915481026
place lettuce into bowl	0.00084187183529
add salt	0.00017669517546
add vinegar	0.00005881859760
add oil	0.00022545302635
add pepper	0.00006381660932
mix dressing	0.00013917377509
peel cucumber	0.00031478240271
cut cucumber	0.00006369184848
place cucumber into bowl	0.00154672493226
add dressing	0.00009407318430
mix ingredients	0.00063657807186
serve salad onto plate	0.00004826133954
action start	0.00019006876391
action end	0.00006290576857

The Machine Take on Things

Probability Matrix

Activity:	Probability:
cut tomato	0.00108002044726
place tomato into bowl	0.00017580932762
cut cheese	0.00141588645055
place cheese into bowl	0.00107793929055
cut lettuce	0.00106934632640
place lettuce into bowl	0.00040337929385
add salt	0.00043896827264
add vinegar	0.00016904433141
add oil	0.00007885327795
add pepper	0.00039472218486
mix dressing	0.00026627627084
peel cucumber	0.00145996094215
cut cucumber	0.989480376243591
place cucumber into bowl	0.00093776948051
add dressing	0.00068892521085
mix ingredients	0.00004418128446
serve salad onto plate	0.00025483779609
action start	0.00034179221256
action end	0.00022189987066

One More Example for the Road



A. Senderovich, A. Rogge-Solti, A. Gal, J. Mendling, and A. Mandelbaum, “The road from sensor data to process instances via interaction mining,” in Advanced Information Systems Engineering, S. Nurcan, P. Soffer, M. Bajec, and J. Eder, Eds. Cham: Springer International Publishing, 2016, pp. 257–273.

What's Stochastic in Stochastic Processes?

Model (Data set) →	Single process		Multiple processes	
↓ Observation (Log)	DK	SK	DK	SK
Deterministically Known (DK)	1	2	3	4
Stochastically Known (SK)	5	6	7	8

I. Cohen and A. Gal, "Uncertain process data with probabilistic knowledge: Problem characterization and challenges," Proceedings of the International Workshop Problems21, collocated with the 19th International Conference on Business Process Management BPM 2021, Italy, published in CEUR Workshop Proceedings, vol. 2938, pp. 51–56, 2021.

Stochastic Logs: what does it Mean?

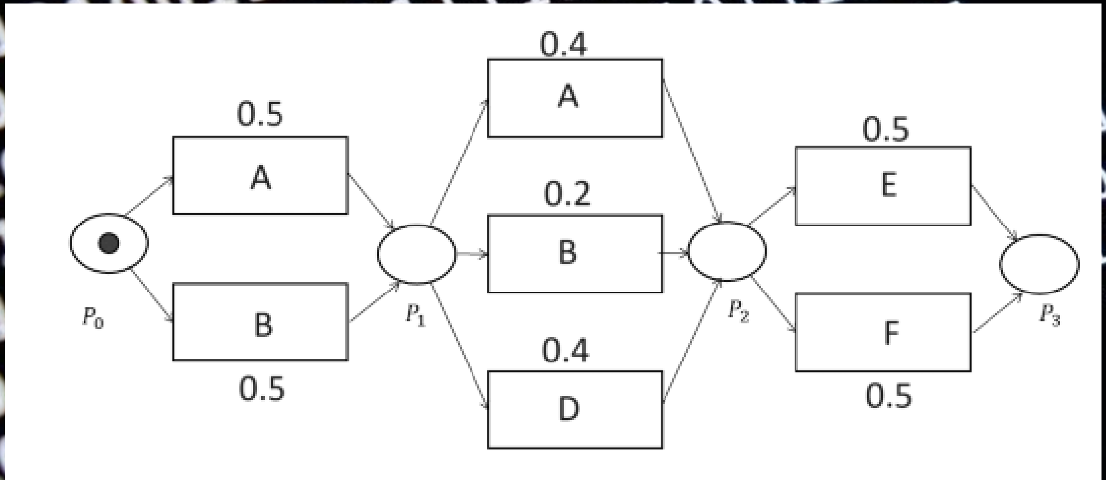
Case ID	Event ID	Activity	Timestamp	Event Probability	Process
1	e_1	{a:0.5, b:0.5}	11-06-2020T00:00	0.8	{P1:0.8, P2:0.2}
1	e_2	{a:0.3, b:0.7}	[12-06-2020T13:52 12-06-2020T14:14]	0.5	{P1:0.8, P2:0.2}
1	e_3	{a:0.2, b:0.3, c:0.5}	13-06-2020T15:39	0.3	{P1:0.8, P2:0.2}
2	e_4	{c:1.0}	15-06-2020T11:23	0.7	{P3:1.0}
2	e_5	{d:1.0}	15-06-2020T11:25	0.2	{P3:1.0}

Probabilistic Databases

- Tuple-level semantics
- Attribute-level semantics:
  - Independence selection of values from a distribution.
- Possible world semantics:
  - The likelihood of each possible world is given by the product of marginal probabilities of tuples' selections.
- More complex models exist

M. Weidlich, "Database systems and process management – a call for a closer look," in Business Process Management - 21st International Conference, BPM 2023, Utrecht, The Netherlands, September 11-15, 2023, Proceedings, ser. Lecture Notes in Computer Science, C. D. Francescomarino, A. Burattin, C. Janiesch, and S. Sadiq, Eds., vol. 14159. Springer, 2023. [Online]. Available: https://doi.org/10.1007/978-3-031-41620-0

A quick illustration



Case ID	Activity	Timestamp
1	$\{(A, 0.5), (B, 0.5)\}$	3-6-22T12:00
1	$\{(A, 0.4), (B, 0.2), (D, 0.4)\}$	3-6-22T14:55
1	$\{(E, 0.5), (F, 0.5)\}$	4-6-22T17:39

Values	Probability
$S(e_1) = A, S(e_2) = B, S(e_3) = E$	$0.5 \cdot 0.2 \cdot 0.5 = 0.05$
$S(e_1) = A, S(e_2) = B, S(e_3) = F$	$0.5 \cdot 0.2 \cdot 0.5 = 0.05$
$S(e_1) = A, S(e_2) = A, S(e_3) = E$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = A, S(e_2) = A, S(e_3) = F$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = A, S(e_2) = D, S(e_3) = E$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = A, S(e_2) = D, S(e_3) = F$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = B, S(e_2) = B, S(e_3) = E$	$0.5 \cdot 0.2 \cdot 0.5 = 0.05$
$S(e_1) = B, S(e_2) = B, S(e_3) = F$	$0.5 \cdot 0.2 \cdot 0.5 = 0.05$
$S(e_1) = B, S(e_2) = A, S(e_3) = E$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = B, S(e_2) = A, S(e_3) = F$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = B, S(e_2) = D, S(e_3) = E$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$
$S(e_1) = B, S(e_2) = D, S(e_3) = F$	$0.5 \cdot 0.4 \cdot 0.5 = 0.1$

Probabilistic Databases: a Model for Stochastic Logs

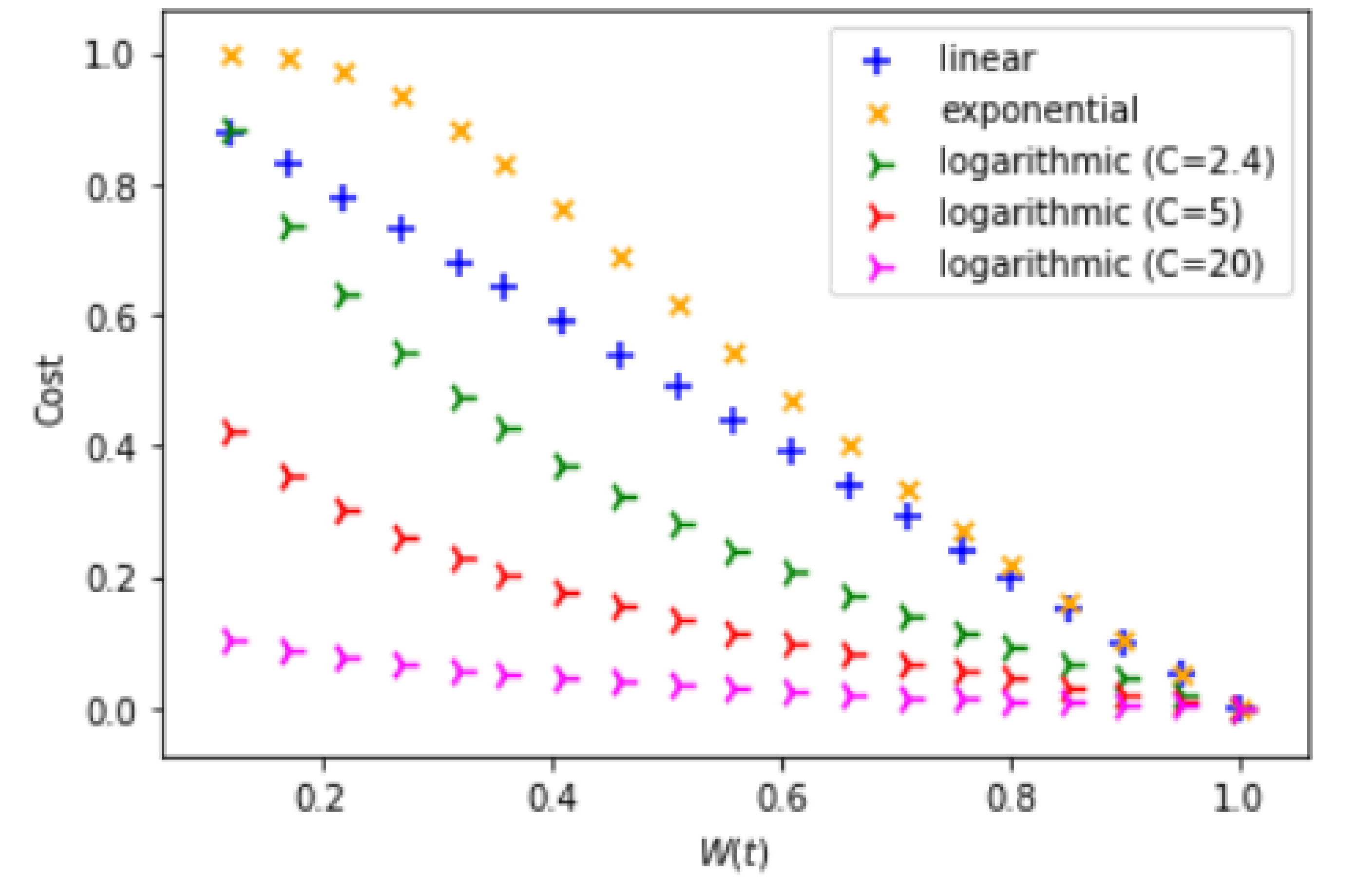
- Log as a probabilistic relation
- Possible world semantics assists in creating deterministic logs from stochastic ones:
  - What’s the motivation?
  - Pros and Cons
- Independence assumptions:
  - Activity independence?
  - Trace independence?

What are my options?

Encapsulate

Expose

Stochastic Log Encapsulation



Low

High

quality

- Stochastic on the inside, deterministic on the outside
- Trace Recovery from Stochastically Known Logs
- Attend the "Fresh View on Process Data" session at 16:00

Stochastic Log Exposure: the Selection Game

- Recall possible worlds
- The right log for the job
- Top-K logs:
  - a ranked list of the K logs with the highest probabilities.
  - Recursive definition:
    - the top ranked log is a log with the highest probability
    - i-th highest ranked log is the top ranked log that is different from the $i - 1$ logs that preceded it in at least one element
  - What's the point in having top-K logs?
  - Computation

Encapsulate or Expose: Pros and Cons

	Encapsulate	Expose
Informativeness	↓	↑
Back compatibility	↑	↓
Computation	=	=
Complexity	↑	↓
End-user orientation	↑	↓

Challenges

- Are all uncertatinties born equal? Should we treat sensor limited accuracy similarly to machine learning prediction indecisiveness?
- Probabilistic information acquisition
- End user stochastic orientation
- Data dependencies?
- Other representations: fuzzy set theory
- Performance: # possible worlds explode

Final notes

- Growing interest