Queue Mining:
The journey from predictive to prescriptive analytics in congested systems

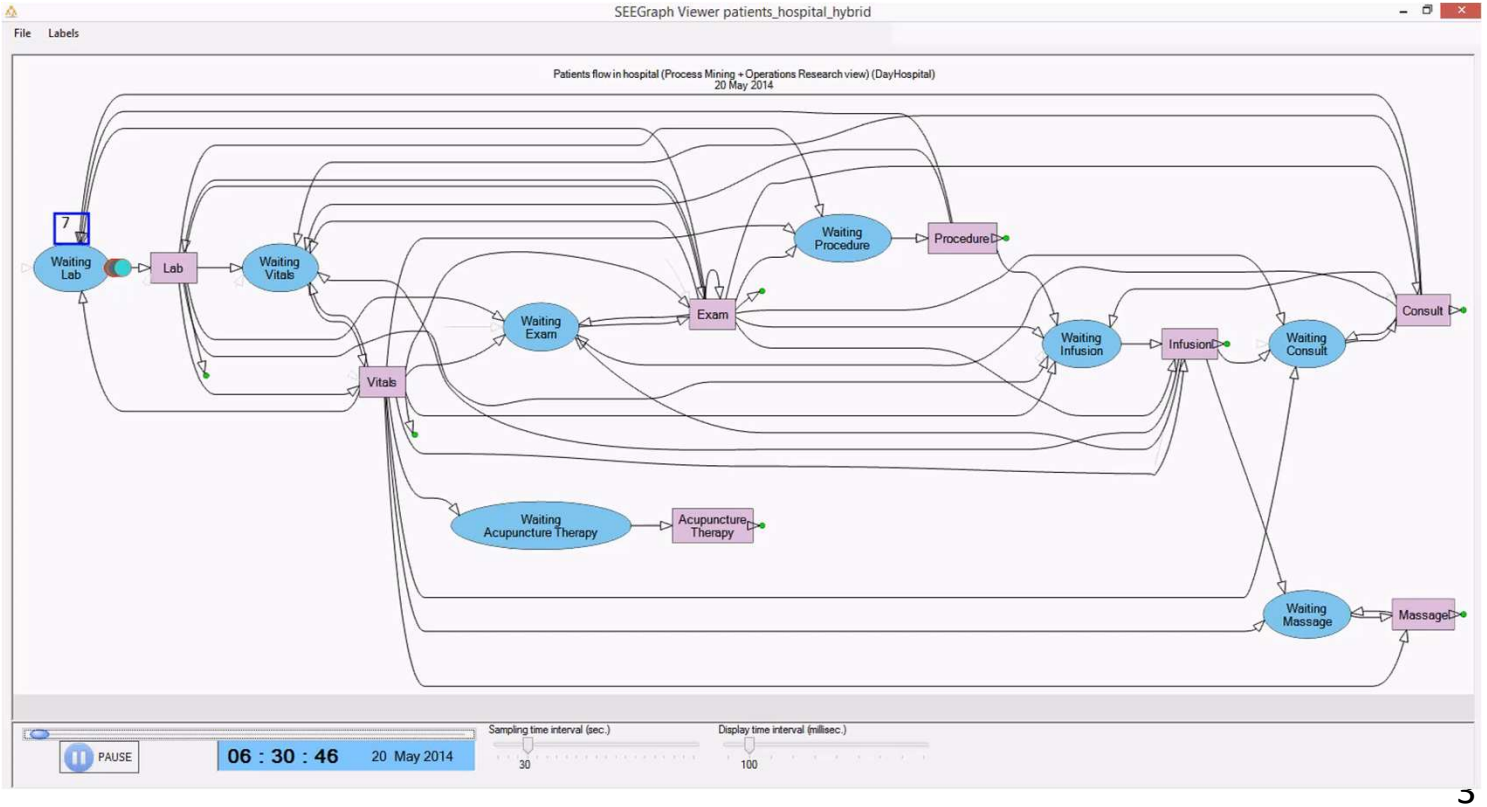
AAAI2023 Bridge on AI4BPM

Arik Senderovich
School of Information Technology, York University
Rotman School of Management, University of Toronto

2/8/2023

Motivation: Congested Systems

Descriptive Analytics

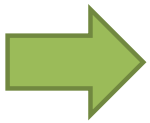


Business Analytics in Congested Systems

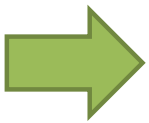
- **Descriptive analytics**: (Past)
  - Is the system performing as expected?
  - Where are the bottlenecks?
- **Predictive analytics**: (Future)
  - How long will the patient wait?
  - When will the patient be discharged?
- **Prescriptive analytics**: (Future + Optimization)
  - What is the optimal number of resources that we must add?
  - What is the optimal patient routing?

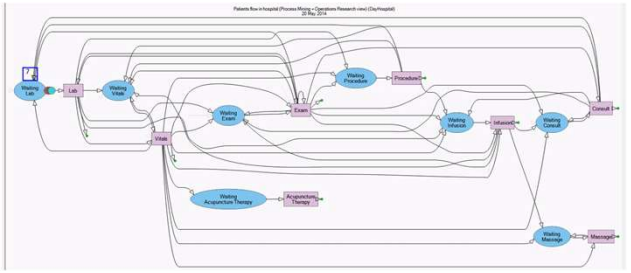
Queue Mining=
Process mining in congested systems

Patient ID	Event	Timestamp	Remaining Time
111	Reception_End	7:30:04	03:45:16
111	Triage_End	7:47:12	03:28:08
111	Treatment_End	9:10:10	02:05:10
222	Triage_End	7:35:52	00:15:32
222	Discharge	7:51:24	00:00:00



Model Learning



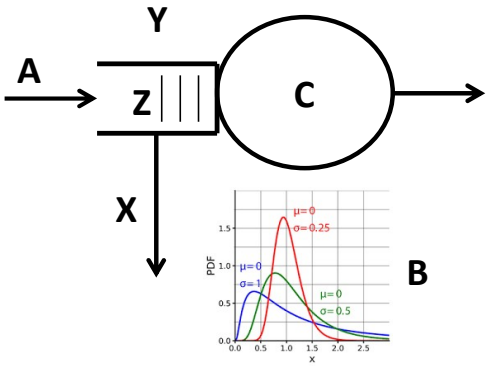


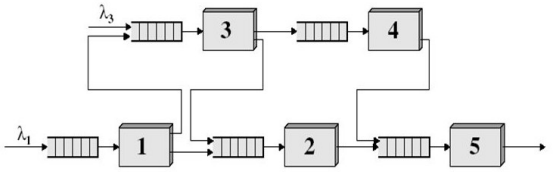
Senderovich, Weidlich, Gal, Mandelbaum. *Queue Mining:...* Information Systems, 2014

What do we need to learn?

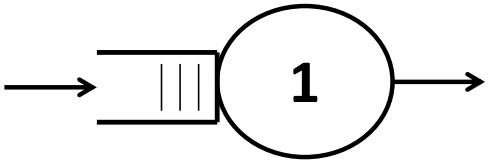
Queueing model via Kendall's notation: **A/B/C/Y/Z +X**

- **A** – arrivals (Poisson, appointment based,…)
- **B** – service times (Exponential, lognormal, empirical,…)
- **C** – server capacity (static/dynamic)
- **Y** – queue capacity (=K, infinite)
- **Z** – service policy (FCFS, LCFS, Processor Sharing…)
- **X** – (Im)patience (Exponential, general)
- Add **R** - routing scheme in networks (e.g., Markovian)
- Typically: independence between building blocks





Common Example: M/M/1



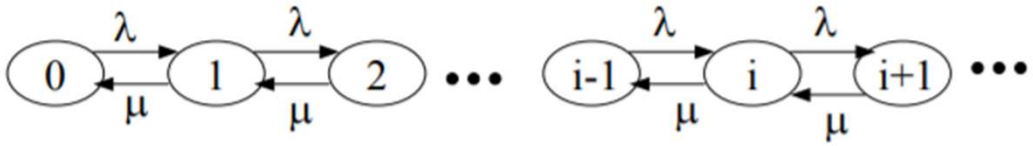
Assumptions (on A/B/C/Y/Z+X):

- Dropped notation Y, Z, X (defaults are taken): infinite queue size, FCFS policy, no abandonments, single server
- M - Poisson arrivals (completely random, one at a time, constant rate)
- M - Exponentially distributed service times
- Easy to analyze when parameters are known (or fitted from data)

Closed-Form Analysis of M/M/1

- The model is “Markovian”
- One can represent M/M/1 using a continuous-time Markov chain (CTMC) that counts the number of customers in the system:

  - Poisson arrivals, rate λ;
  - Single exponential server, rate μ; $\mathrm{E}[S] = 1/\mu$.

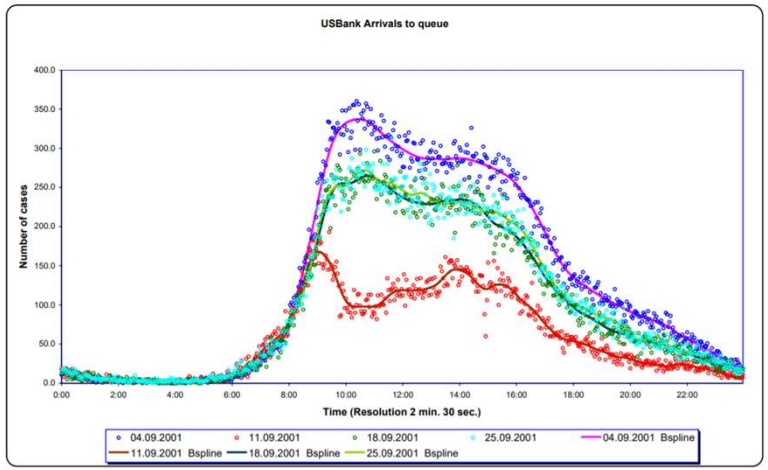


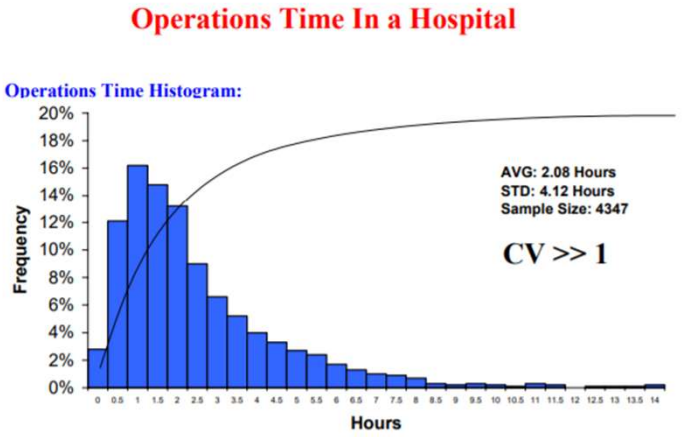
Important: small additional complexity may make the model “intractable”;

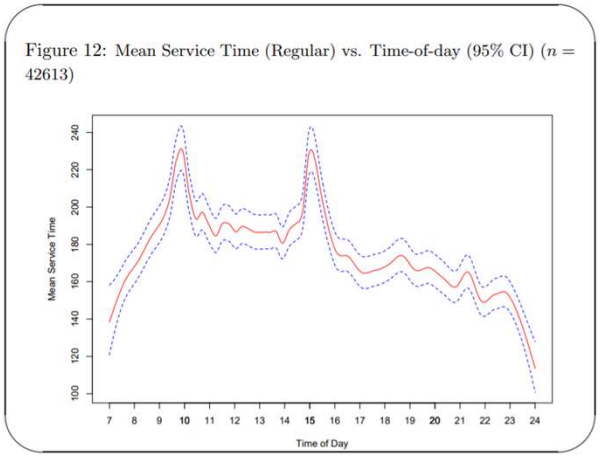
However:

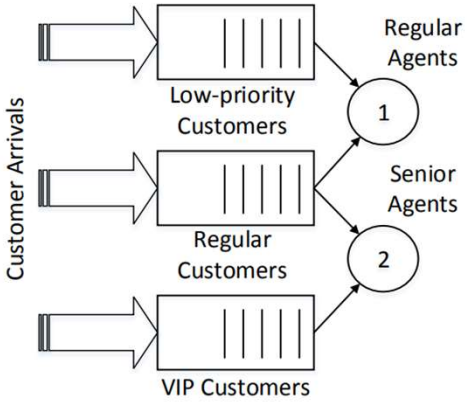
- Simulation
- ML

Gallery of Empirical Assumption Violations



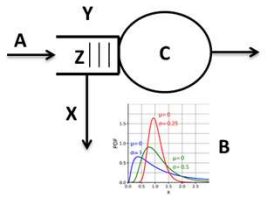






Queue Mining: Learning and Analyzing

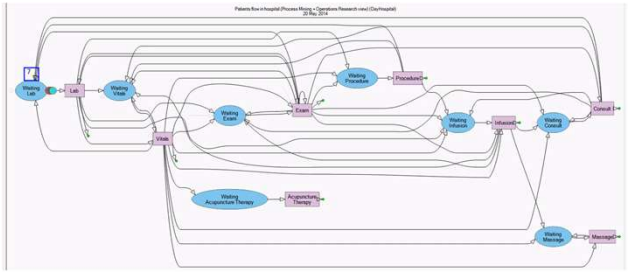
Patient ID	Event	Timestamp	Remaining Time
111	Reception_End	7:30:04	03:45:16
111	Triage_End	7:47:12	03:28:08
111	Treatment_End	9:10:10	02:05:10
222	Triage_End	7:35:52	00:15:32
222	Discharge	7:51:24	00:00:00



Model Learning

Learn:

1. Structure (Routing)
2. Dynamics (Arrivals, service times,…)
3. Scheduling policies



Analyze via:

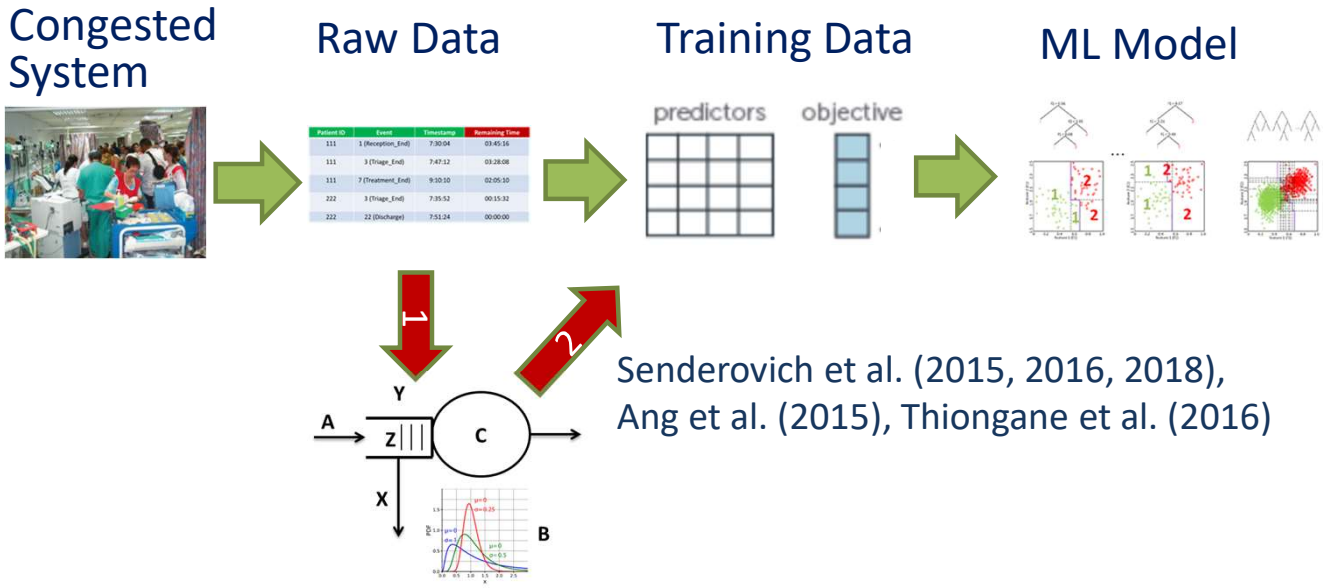
1. Closed-form expressions
2. Simulation
3. Machine learning

Senderovich et al. (2016). *Conformance checking and performance improvement in scheduled processes: A queueing-network perspective.*

Agenda

- Background on queue mining
- Queue mining for:
  - Predictive analytics
  - Prescriptive analytics
- Ongoing research in queue mining

Predictive Analytics & Queue Mining



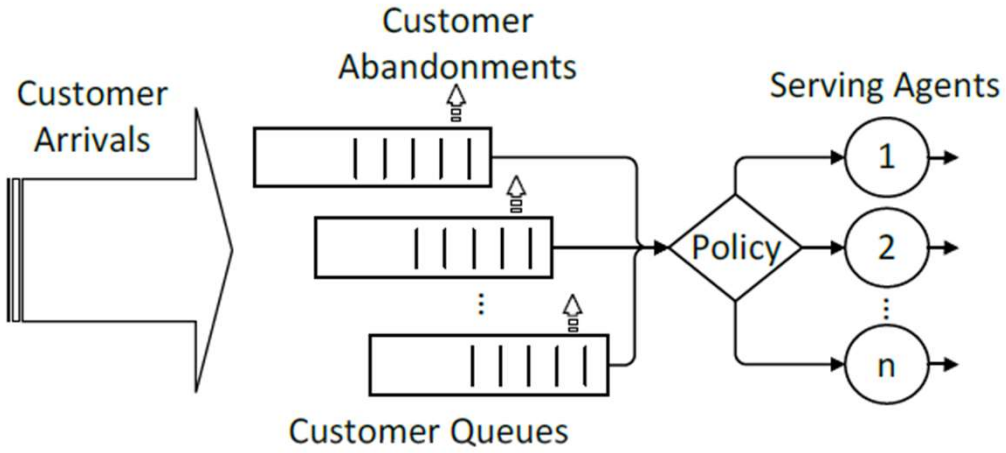
Senderovich et al. (2015, 2016, 2018),
Ang et al. (2015), Thiongane et al. (2016)

Ibrahim and Whitt (2009a, 2009b, 2010 ,2011)
Senderovich et al. (2015, 2018)

Green Path: Feature encoding (Leontjeva et al., 2015, Tax et al., 2017…)

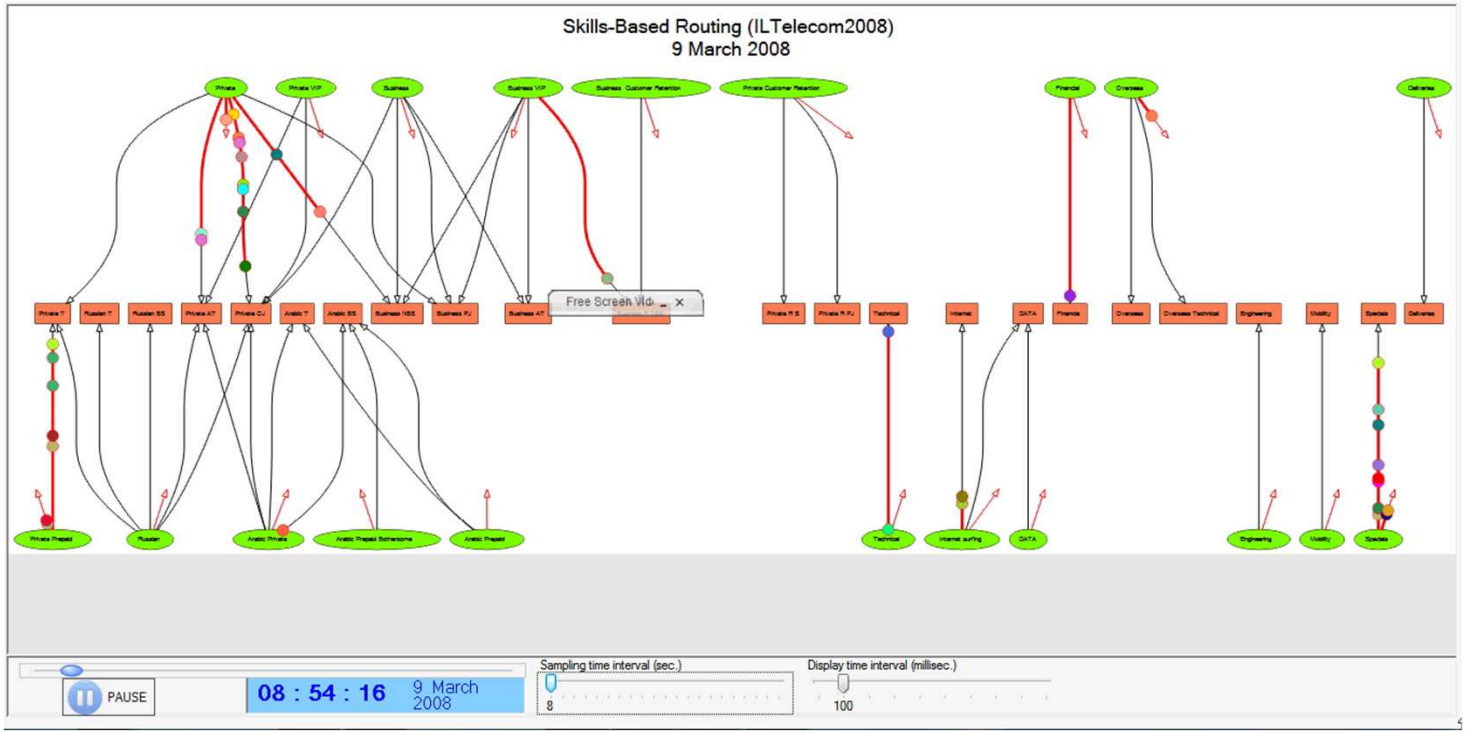
Predictive queue mining: paths 1 and 2

Single-Station Models:
Predictive Analytics in Call Centers

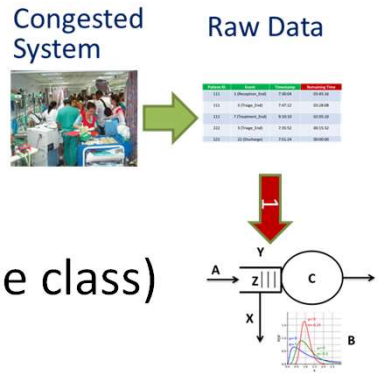


 Senderovich et al. (2014). Queue mining–predicting delays in service processes

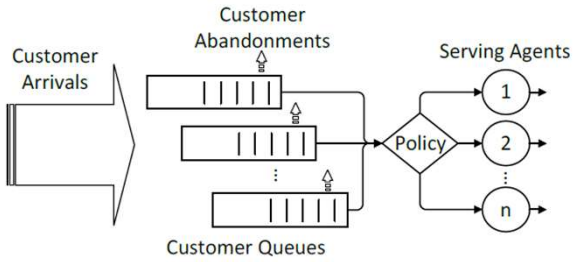
Routing in a Telecom Call Center



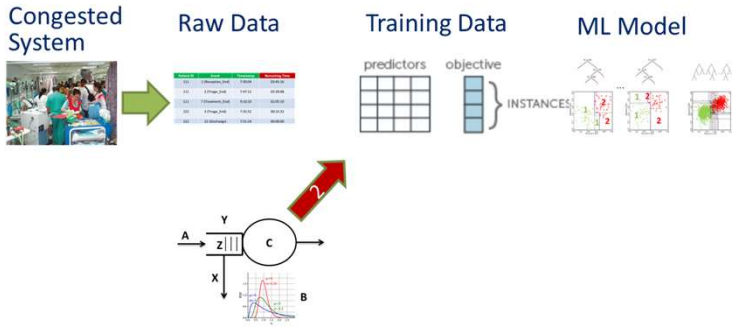
Snapshot Prediction and Bounds



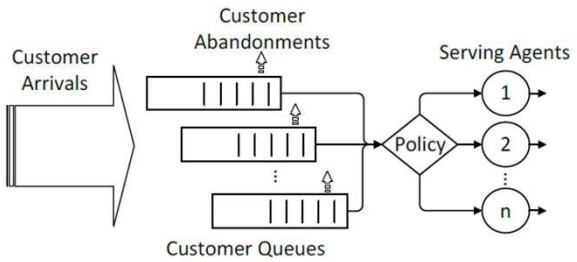
- Snapshot prediction: 1NN using the most recent to enter service (same class)
- Upper and lower bounds on **expected wait time per class**:
  - Upper bound – order of service completions is from "slowest" to "fastest" customer class; includes "overtaking"
  - Lower bound – "fastest" to "slowest" customer completes service in every iteration; no "overtaking"
  - Bounds coincide with each other and with the queue-length predictor for single-class queues



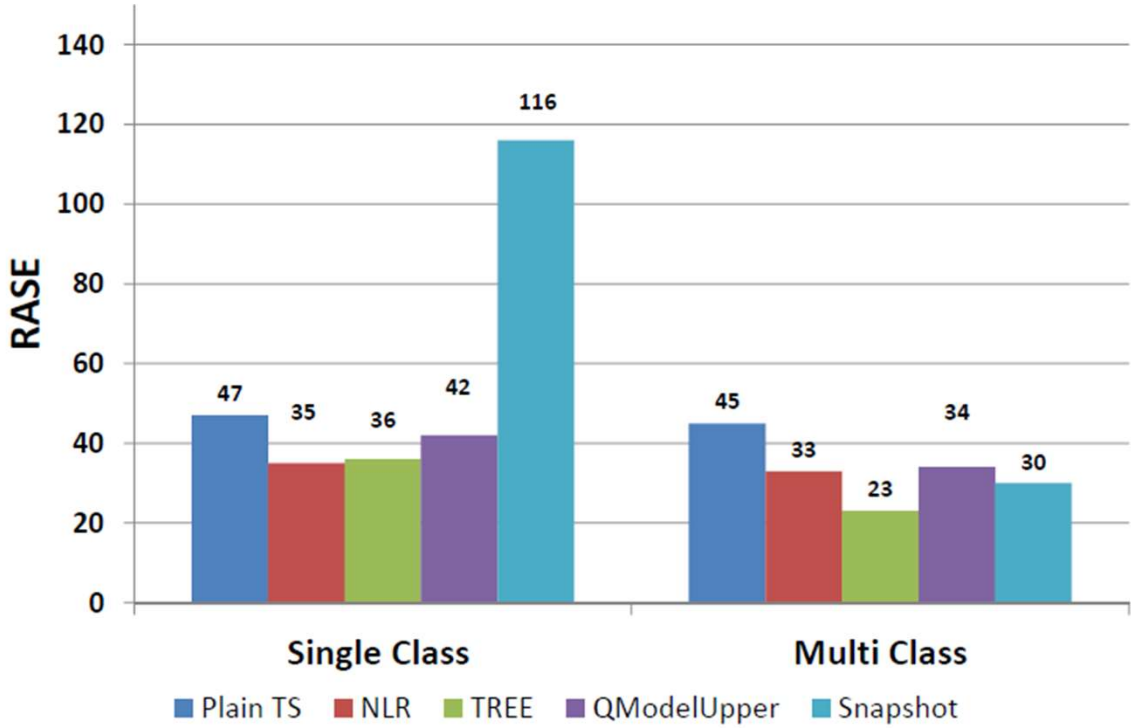
ML for Multi-Class Queues



- Feature encoding:
  - Queue-lengths per class
  - Number of customers in service
  - Class (customer type)
- Fed into nonlinear regression: Generalized Additive Models (GAMs) and regression trees

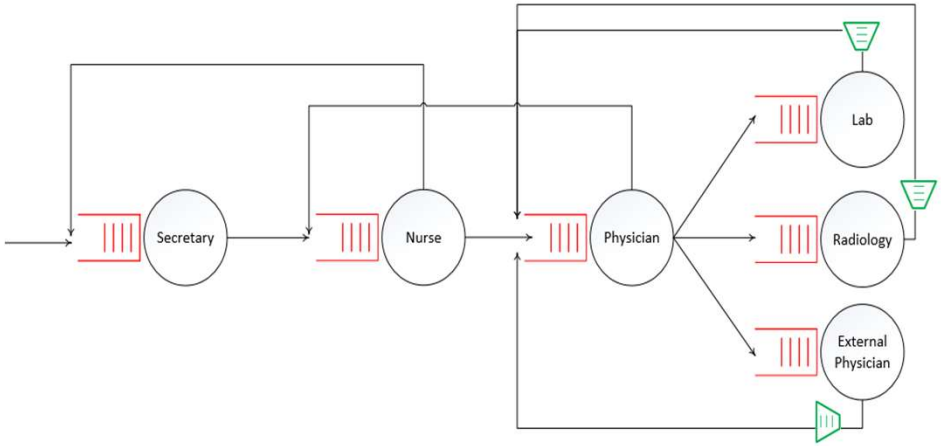


Results: Telecom Call Center Data (2008)



Snapshot predictor (1-NN) – no learning required.
QModel enables `what-if?' analysis, sensitive to assumptions
ML – most accurate and robust results

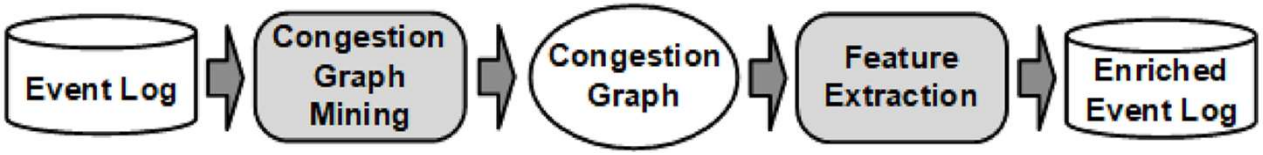
What about queueing networks?
Example: Hospital Processes



Congestion Graphs for Automated Time Prediction

Presented @ AAAI 2019

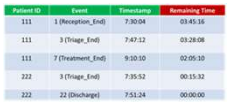
Arik Senderovich, J. Christopher Beck, Avigdor Gal, Matthias Weidlich



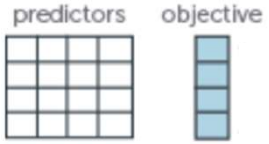
Prediction with Congestion Graphs

Congested System

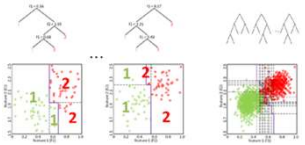
Raw Data



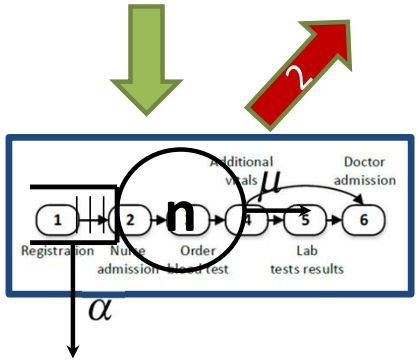
Training Data



ML Model



Congestion graphs

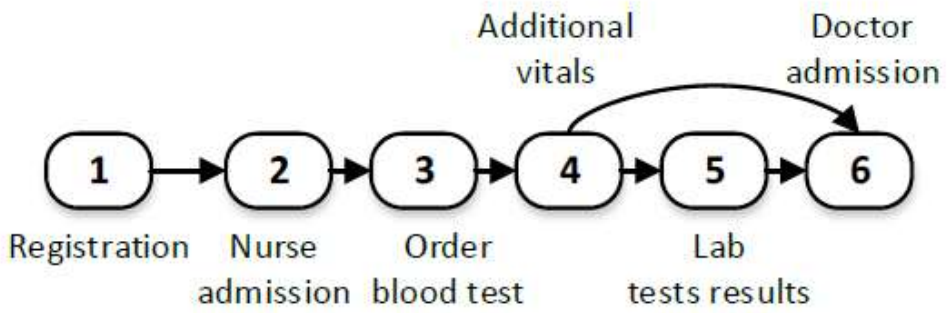


Patient ID	Event	Timestamp	Remaining Time
111	Reception_End	7:30:04	03:45:16
111	Triage_End	7:47:12	03:28:08
111	Treatment_End	9:10:10	02:05:10
222	Triage_End	7:35:52	00:15:32
222	Discharge	7:51:24	00:00:00

General model grounded in QT; no expert knowledge required

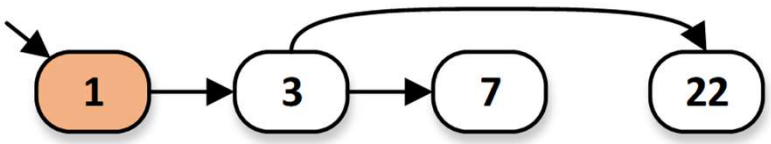
Defining Congestion Graphs

- Dynamic graphs that capture congestion features
- Construction:
  - **Vertices** are events from the log
  - **Edges** connect pairs of directly following events
  - **Entities** flow on the edges (time-on-the-arc)



Defining Congestion Graphs (Cont.)

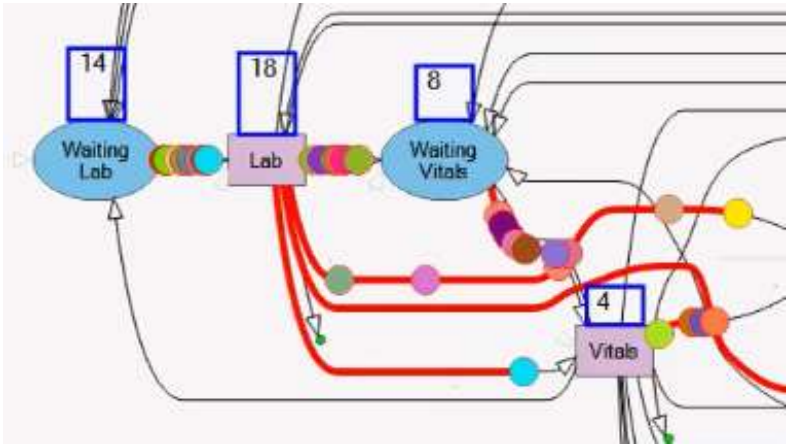
- Dynamic node labelling:
  - System state representation at time *t*
  - Mapped to congestion features
  - Technical details in the paper:
    - Queueing theoretic justifications for system state representation (via Jackson networks)
    - Approximating system from data



Node Labelling: Intuition

- Labelling of nodes with system state at time *t*:
  - Number of patients on outgoing edges (**14**)
  - Cumulative time patients spend on outgoing edges (**60 MIN**)
  - Node inter-arrival time (**5 MIN**)

State(t) = (14, 60min, 5min)

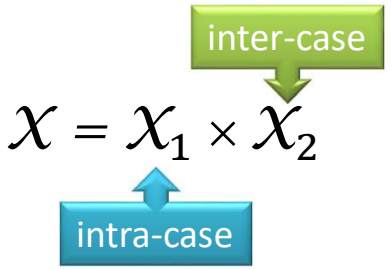


Generalizes to Inter-Case Encoding

Specific features to capture inter-case dependencies

Examples:

- Patients in system
- Patients at particular activity in the process
- Patients of a particular type



Also, feature engineering using
the performance spectrum,
e.g., identifying batching

[Senderovich, Di Francescomarino, Ghidini, Jorbina, Maria-Maggi, BPM 2017]
[Klijn, Fahland, ICPM 2020]

Congestion Feature Extraction

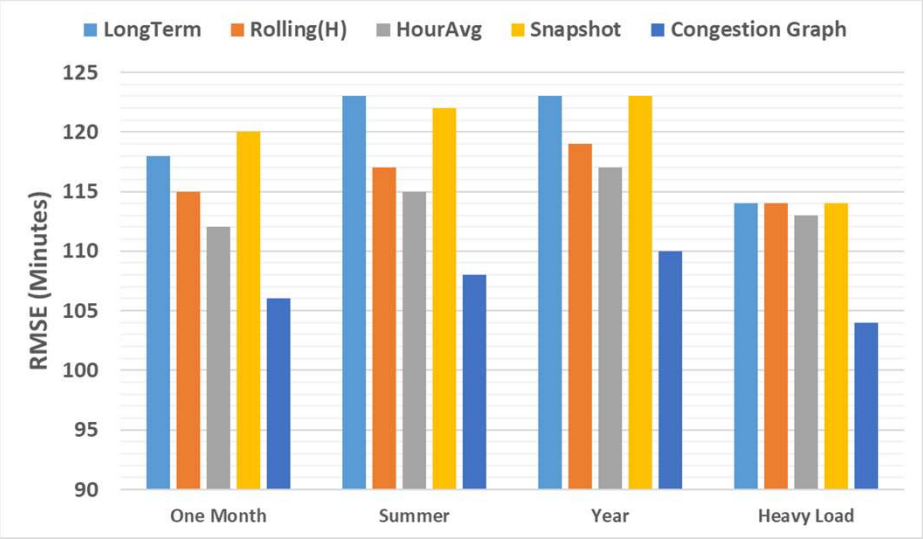
Features = node labels

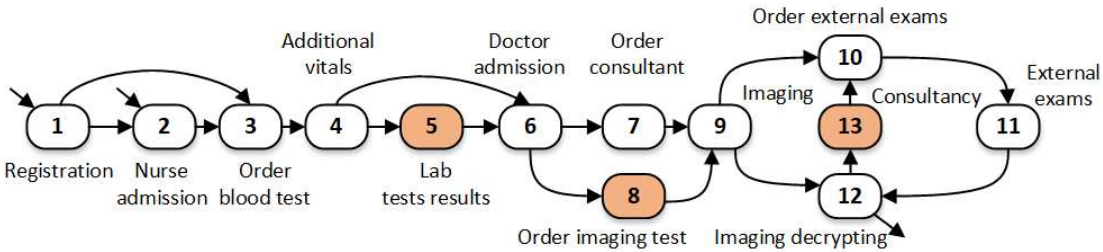
Patient ID	Event	Timestamp	CG[Lab Wait]	CG[Lab]	...
111	Reception_End	7:30:04	**14,60,5**	**6,35,15**	
111	Triage_End	7:47:12	**10,15,10**	**15,45,7**	
111	Treatment_End	9:10:10	…	…	
222	Triage_End	7:35:52	…	…	
222	Discharge	7:51:24	…	…	

Information available at test time

Number of features grows linearly in size of congestion graph

Results on a hospital dataset: Accuracy and Explainability





Agenda

- Background on queue mining
- Queue mining for:
  - Predictive analytics
  - Prescriptive analytics
- Ongoing research in queue mining

Mining Hybrid Machine Learning and Simulation Models

Presented @ BPO Workshop 2022

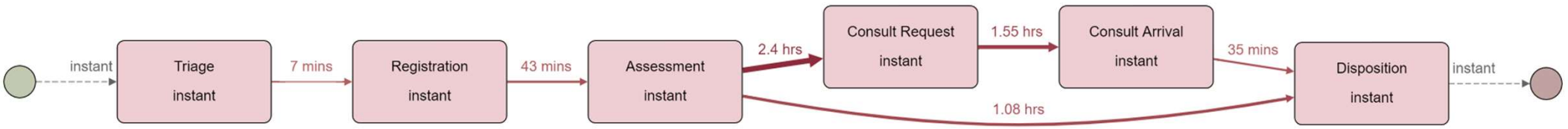
Arik Senderovich
School of Information Technology, York University (Toronto)

Opher Baron and Dmitry Krass
Rotman School of Management, University of Toronto

Nancy Li
Industrial Engineering, University of Toronto

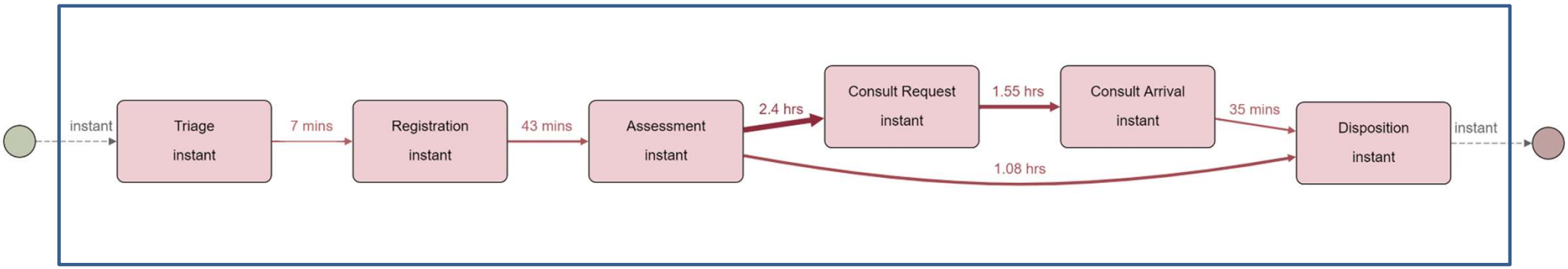
Motivation: Prescriptive Analytics in NYGH

- **Background**: NYGH is one of the largest EDs in Toronto
- **Goal**: improve patient flow by **reducing length-of-stay (LOS) of <u>consult</u> patients**
- **Task**: Estimate the effect of the intervention on the overall performance
- **Failed solution**: model LOS via ML and shorten LOS until convergence
  - Finds the "first-order" effect (individual reduction)
  - Doesn't capture "second-order" effect (system) – no convergence!
- **Required**: a model that would capture dependencies between cases (system)



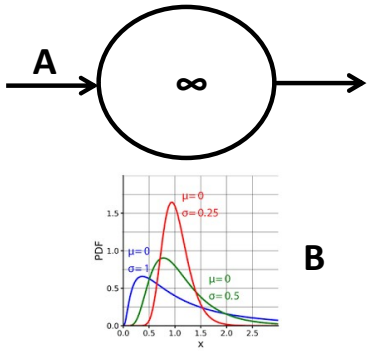
Problem: Missing Resource and Service Information

- In queueing models: nodes = resources
- However (as before) healthcare event logs are often missing:
  - Resource information (privacy consideration and measuring4billing)
  - Wait start-times
  - Most of the activities (only arrival, departure are trustworthy) – cannot use congestion graphs



Solution: Hybrid ML and Infinite-Server Model

- A - Arrivals = Poisson/appointment based/ML based/**data-driven**
- B - Service times = **ML model**
  - "Contextual" + **state** features
- C - Server capacity = infinity (missing data)



- Analysis: Simulation (+ML)

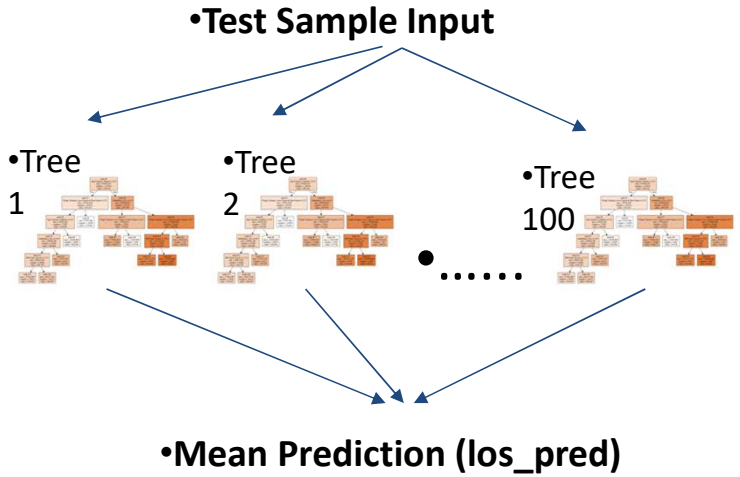
In Queueing Theoretic Notation

- "G(*t*)/G*(X,S)*/infinity" queueing model
- Service times depend on current **system state *S and context X***

  System state = static + dynamic + congestion contexts:
  - Static = medical history
  - Dynamic = temperature, time of day
  - Congestion = number in system (NIS) with variations

- Theoretical problem:
  - Does an M/G(*S*)/infinity => M/G/C?
  - Preliminary results are encouraging (under restrictions)

Service Times Sampling via Random Forests



Service times via Random Forest (RF) Regressor:

- **x** = Patient static info*, System State X, Season, Trend, Holidays*
- **y** = Patient's length of stay (LOS)
- **Hyperparameters**: min 30 samples in leafs and 100 trees in the RF model

Sampling Method from RF Model:

- Store a bucket of training errors
- For each sample (in **log** minutes):
  - los_pred = RF mean prediction
  - Error = random sample from the bucket of training errors
  - **Sampled LOS = exp(los_pred + error)**

•***Patient static info:** age, gender, ambulance, consult, initial zone, arrival hour, arrival day of week

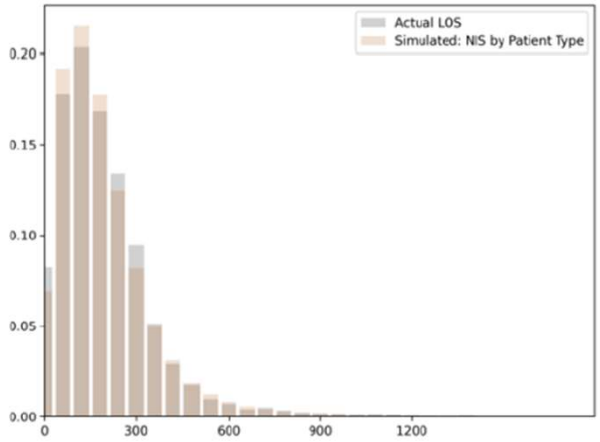
•***Season** (arrival week number, arrival month), **Trend** (number of weeks since beginning of training data), **Holidays** (Ontario public holidays)

Bringing it together…

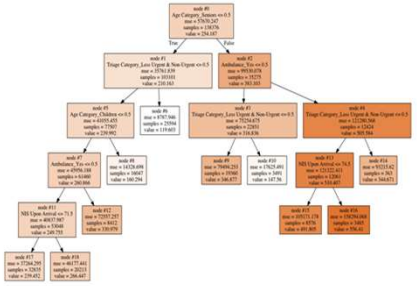
Patient Arrives

Sampled From Random Forest Models

1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)



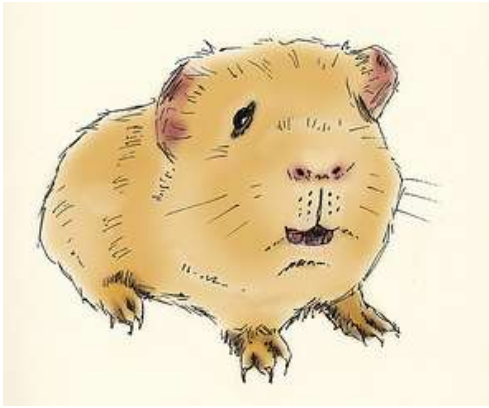
Patient's LOS Model



Testing on NYGH Data

- Hybrid modeling and feature selection
- "Goodness-of-fit" analysis
- Prescriptive analytics:
  - Is the model useful for intervention analysis?

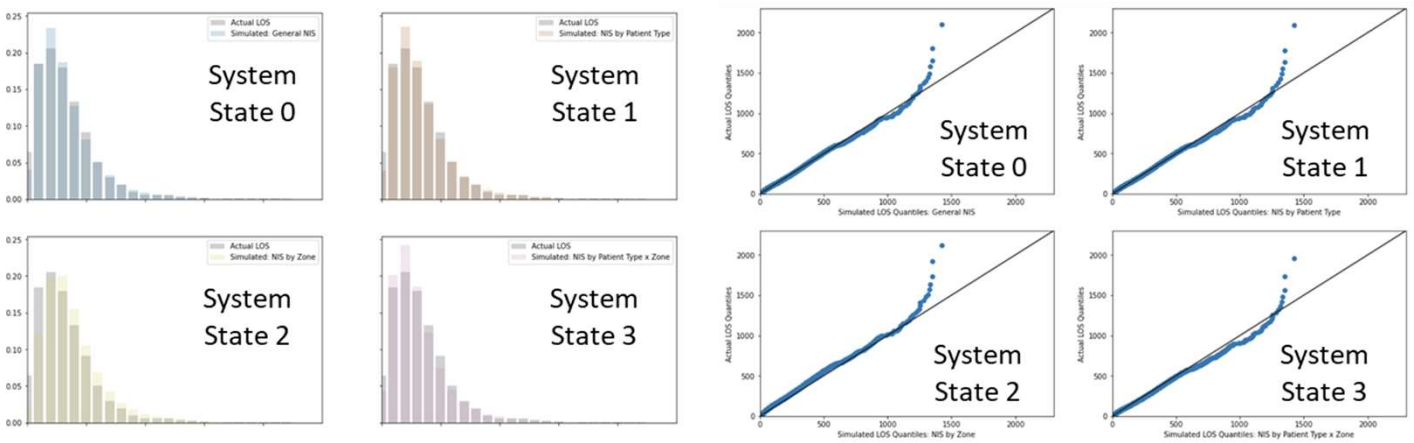
  i.e., can the model capture congestion effects?



Additional datasets tested: Israeli ED, Singapore ED

“Goodness-of-Fit” Results

1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)



- Performance of the 4 models **cannot be differentiated** via visual inspection
- Their LOS tends to be **higher in simulated distribution** than in actual distribution (except for the tail)
- At the tails (highest values in the distribution), all models **deviates significantly** from the 45-degree line
- The **model of System State 0** with General NIS is the best (model 1 is a close second)

Intervention on Consult Patients

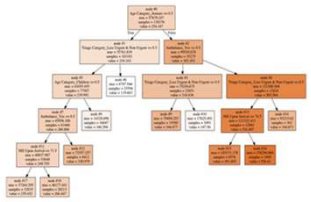
Patient Arrives

Baseline — "**Naive**" (From Data)

Sampled From Random Forest Models

1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)

Patient's LOS Model

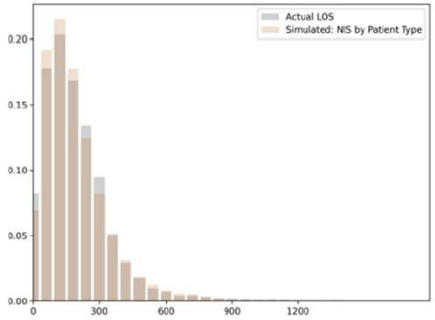


Is this a **Consult** patient? If yes...

Apply Intervention*

If intervention is 10% cutdown:
LOS = LOS x 0.9

***Interventions**: 10%, 20%, 30%, 40%, 50% Cutdown in Consult Patient's LOS



Intervention Results

- **For T123 Admitted patients:**
  - The *% of reduction* in both mean and 90th percentile LOS are **comparable** to the (*% LOS cutdown x % of consult patients)*
  - Reason: queueing system is "transparent" for these patients
- **For both T123 Not Admitted and T45 patients:**
  - **Baseline** captures the **1st order effect** of reducing consult patient's LOS
  - Our **state-aware models**, also capture **2nd order effect**
  - The 2nd order congestion effect can be up to 10 times stronger than the 1st order

Ongoing Research in ~ Queue Mining

- Comparing simulation/queue mining approaches (with Dumas et al.)
- Relating stochastic Petri nets and queueing models to process mining (with Montali)
- The price of misprediction: on scheduling with predictions (with Dijkman)
- Queues with invisible customers (with Shaposhnik et al.)
- Prescriptive analytics (with Baron, Krass, Hu):
  - Developing theoretical model for convergence to finite-server model (for many-servers)
  - Applying to a model with more data (gathered recently)
  - Employing causal models in queue mining for better intervention estimates
  - Considering additional interventions (e.g., admission control)

Commercial Pause:

Friday morning talk @ AAAI2023

Privacy Attacks on Schedule-Driven Data

Stephan A. Fahrenkrog-Petersen, Arik Senderovich, Alexandra Tichauer, Ali Kaan Tutak, J. Christopher Beck, Matthias Weidlich

AAAI 2023

Acknowledgements

- The work was developed with my wonderful supervisors and co-authors including (left-to-right): Avigdor Gal, Matthias Weidlich, Avishai Mandelbaum, Chris Beck, Opher Baron, Dmitry Krass

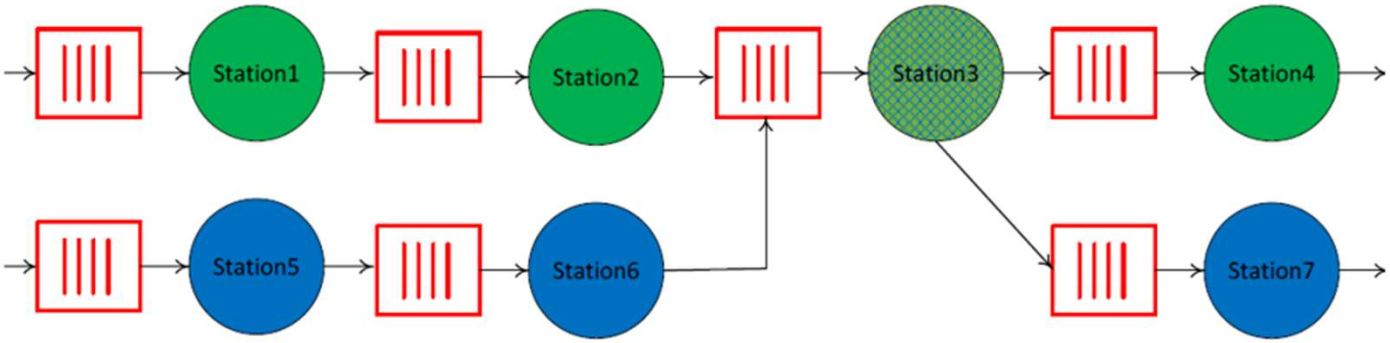




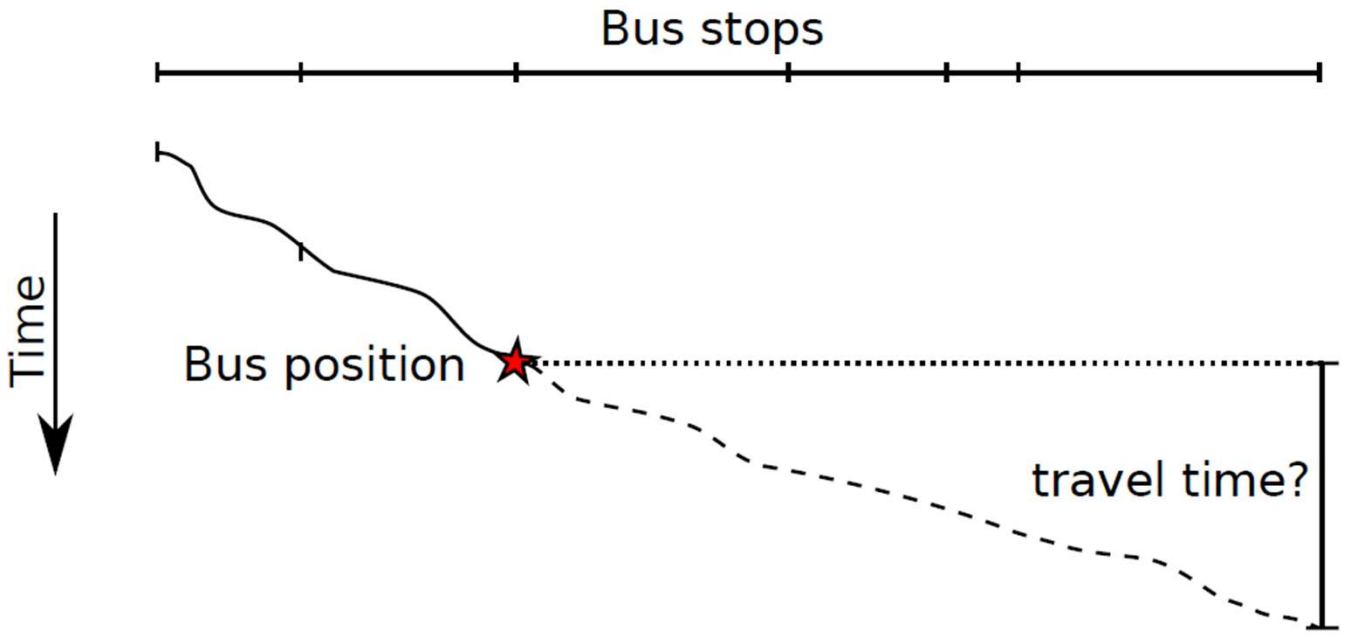

Thank you!

sariks@yorku.ca

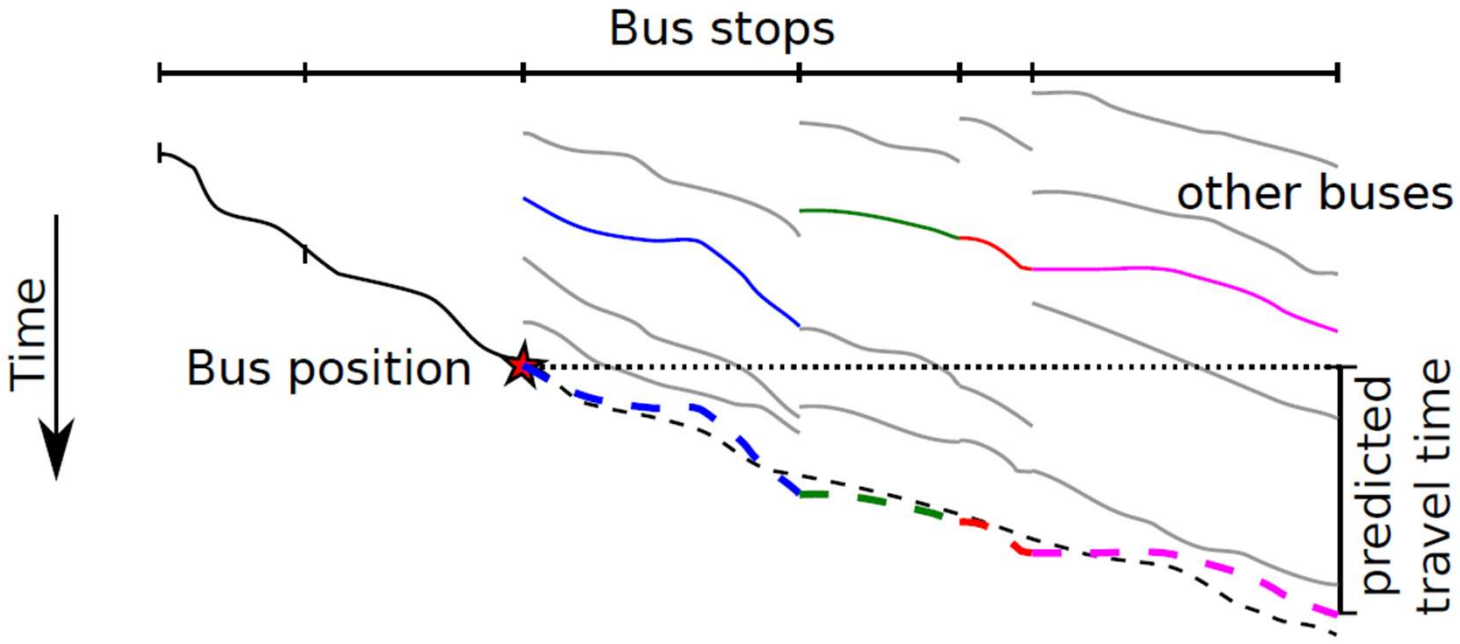
Networks of Queues with Pre-Defined Routes (Transportation Systems)



Snapshot Prediction



Snapshot Prediction

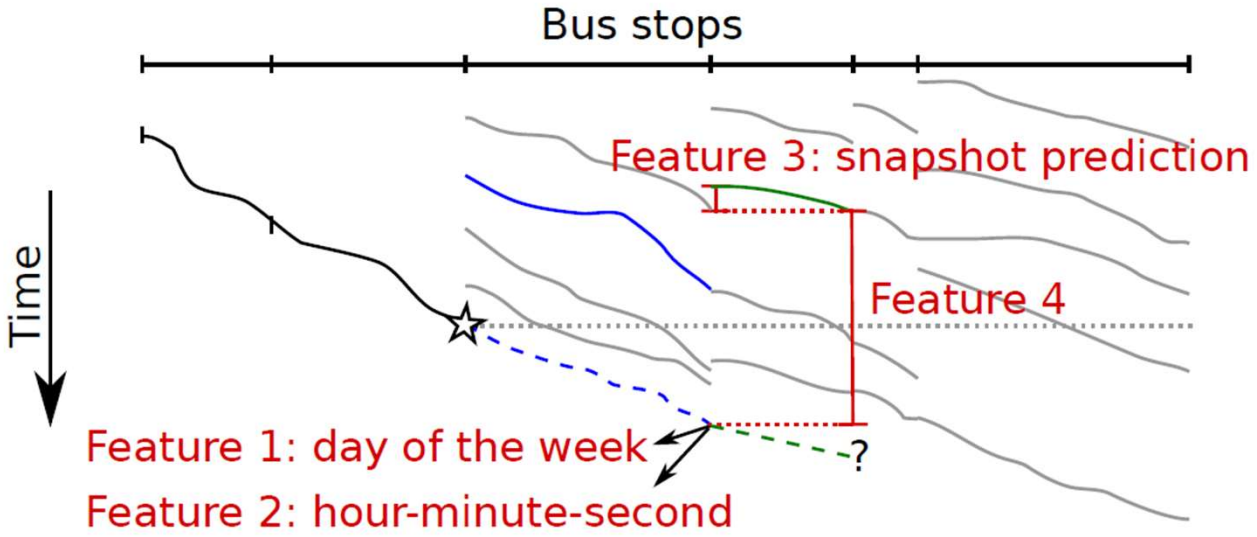


Learning from Historical GPS Data

Event Id	Journey Id	Timestamp	Bus Stop	Journey Pattern
1	36006	1415687360	Leeson Street Lower (846)	046A0001
2	36012	1415687365	North Circular Road (813)	046A0001
3	36009	1415687366	Parnell Square (264)	046A0001
4	36006	1415687381	Leeson Street Lower (846)	046A0001
5	36009	1415687386	O'Connell St (6059)	046A0001
6	36012	1415687386	North Circular Road (814)	046A0001
7	36006	1415687401	Leeson Street Upper (847)	046A0001
8	36009	1415687406	O'Connell St (6059)	046A0001

- Snapshot predictor lacks context (day of week, time of day, future traffic state…)
- We would like to **learn** from similar historical situations to predict the current travel time

ML Approach: Feature Construction

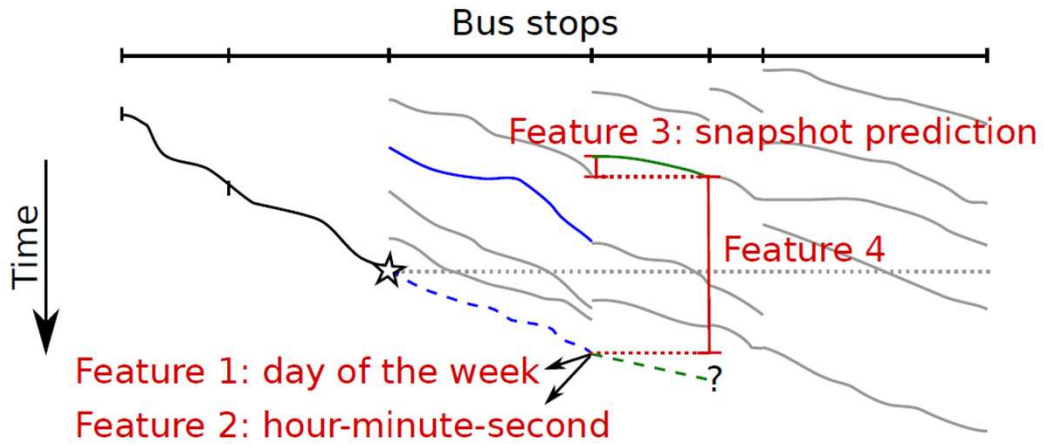


Snapshot predictor is one of the features

Feature Construction

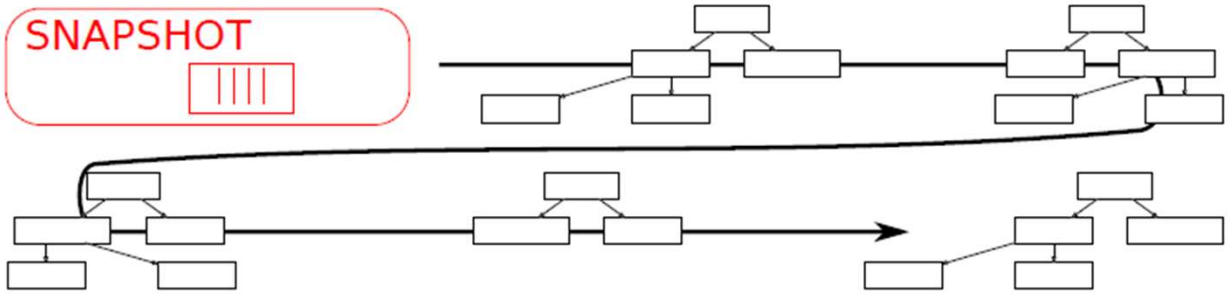
X Y

Journey ID	Segment	HR:MM:SS	Snapshot	Age	Travel Time
36006	846-847	7:00:02	2.5MIN	10MIN	2 MIN
36012	813-814	7:47:12	1.2MIN	30MIN	1 MIN



Learner adaptation

Boosting over the Snapshot Predictor



Results

Accuracy of the prediction over all trips. Worse, Best, **Best of S+xx and xx**

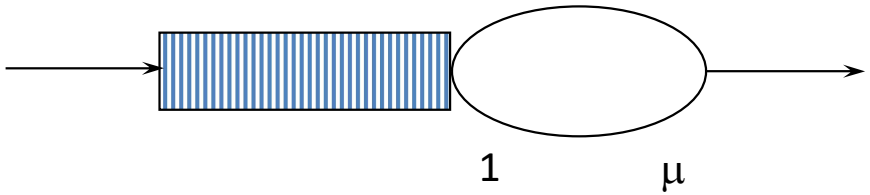
	RMSE	MARE (%)	MdARE (%)
S	539	23.37	16.15
RF	539	24.11	16.37
ET	519	22.05	15.23
AB	512	27.08	18.05
S+AB	**504**	**26.32**	**16.84**
GB	508	20.46	13.84
S+GB	**494**	**19.95**	**13.53**
GBLAD	520	19.38	13.86
S+GBLAD	**514**	**19.06**	**13.65**

Common Performance Measures

- L - number of customers at the service station (sometimes L_s);
- L_q - number of customers in the queue;
- W - sojourn time of a customer at the service station (W_s);
- W_q - waiting time of a customer in the queue.

If λ – arrival rate to the system, Little's formula implies:

$$\mathrm{E}[L] = \lambda \cdot \mathrm{E}[W]; \qquad \mathrm{E}[L_q] = \lambda \cdot \mathrm{E}[W_q]\,.$$



Performance Measures for M/M/1

Traffic intensity $\rho = \frac{\lambda}{\mu} < 1$ (assumed for stability).

Steady-state distribution $L \stackrel{d}{=} Geom(\text{p} = 1\text{-}\rho)$ (from 0):

$$\pi_i = (1-\rho)\rho^i, \; i \geq 0.$$

Properties:

- Sojourn time is exponentially distributed:

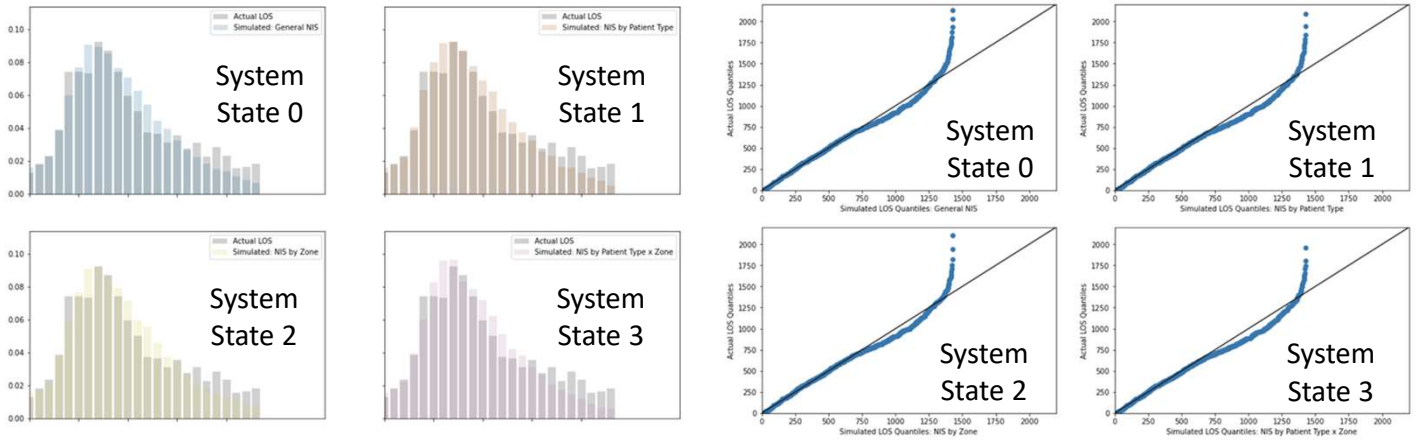
$$W \sim \exp\left(\text{mean} = \frac{1}{\mu(1-\rho)} = \frac{1}{\mu}\left[1 + \frac{\rho}{1-\rho}\right]\right).$$

- Number-in-system: $\text{E}[L] = \frac{\rho}{1-\rho}$; $\text{E}[L_q] = \frac{\rho^2}{1-\rho}$.
- Server's utilization (occupancy) is $\rho = \lambda/\mu$.
  (Little's formula, system = server.)

Final Train-test split – 3 months in 2018

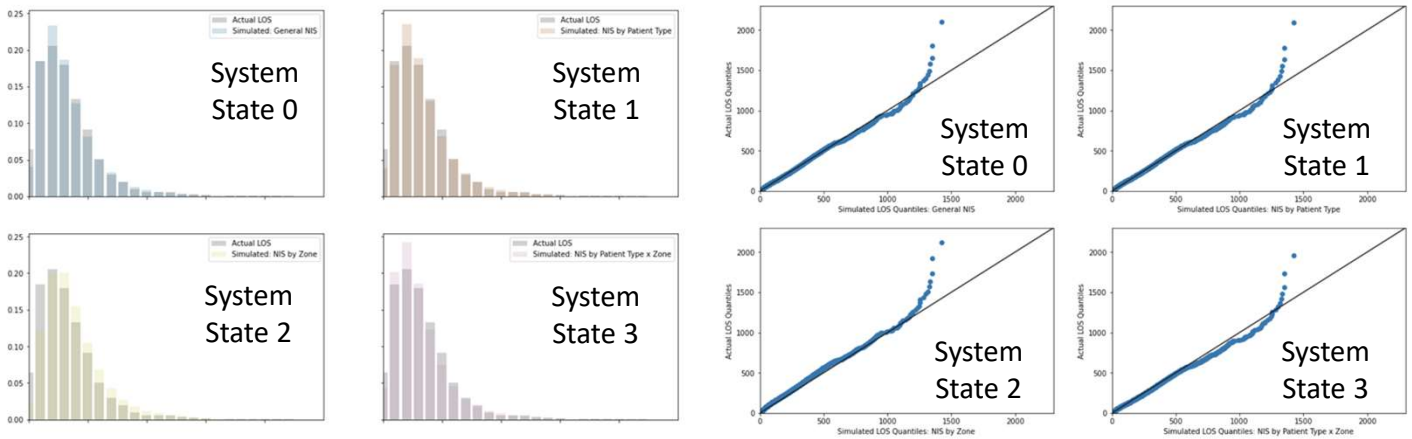
	Training (May, Jun, Jul)	In-sample Test (July)	Out-of-sample Test (August)
T123 Admitted	• 3,422 in total • 97.0% consults	• 1,171 in total • 97.8% consults	• 1,160 in total • 95.2% consults
T123 Not Admitted	• 19,281 in total • 6.17% consults	• 6,550 in total • 5.6% consults	• 6,645 in total • 5.3% consults
T45	• 6,458 in total • 1.33% consults	• 2,168 in total • 1.2% consults	• 1,939 in total • 1.2% consults

LOS Histograms and Q-Q Plots (T123 Admitted, August 2018)



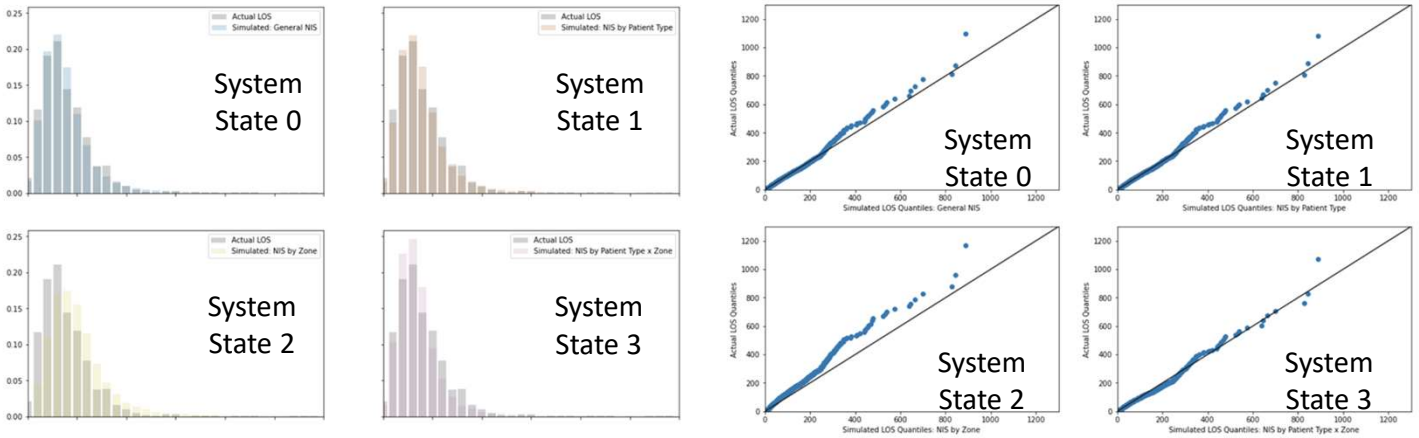
1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)

LOS Histograms and Q-Q Plots (T123 Not Admitted, August 2018)



1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)

LOS Histograms and Q-Q Plots (T45, August)



1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)

KS Test Statistics

	System State 0	System State 1	System State 2	System State 3
T123A (July)	**0.0289**	0.0423	0.0385	0.0490
T123NA (July)	**0.0381**	0.0395	0.0839	0.0737
T45 (July)	0.0584	**0.0541**	0.139	0.115
T123A (August)	**0.0447**	0.0545	0.0496	0.0641
T123NA (August)	**0.0379**	0.0429	0.118	0.0441
T45 (August)	**0.0315**	0.0351	0.217	0.0884

1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)

“Goodness-of-Fit” Results

- Performance of the 4 models **cannot be differentiated** via visual inspection
- Their LOS tends to be **higher in simulated distribution** than in actual distribution (except for the tail)
- At the tails (highest values in the distribution), all models **deviates significantly** from the 45-degree line
- The **model of System State 0** with General NIS is the best (model 1 is a close second)

Modeling and Intervention Analysis

Patient Arrives

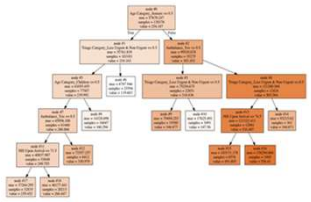
Baseline

"**Naive**" (From Data)

Sampled From Random Forest Models

1. **System State 0** (General NIS)
2. **System State 1** (NIS by Patient Type)
3. **System State 2** (NIS by Zone)
4. **System State 3** (NIS by Patient Type x Zone)

Patient's LOS Model

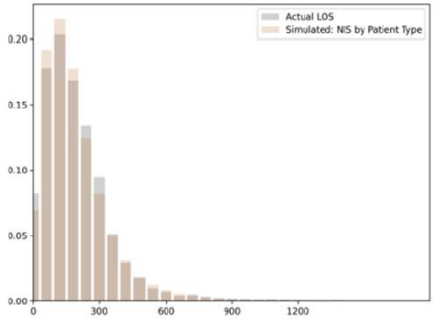


Is this a **Consult** patient? If yes...

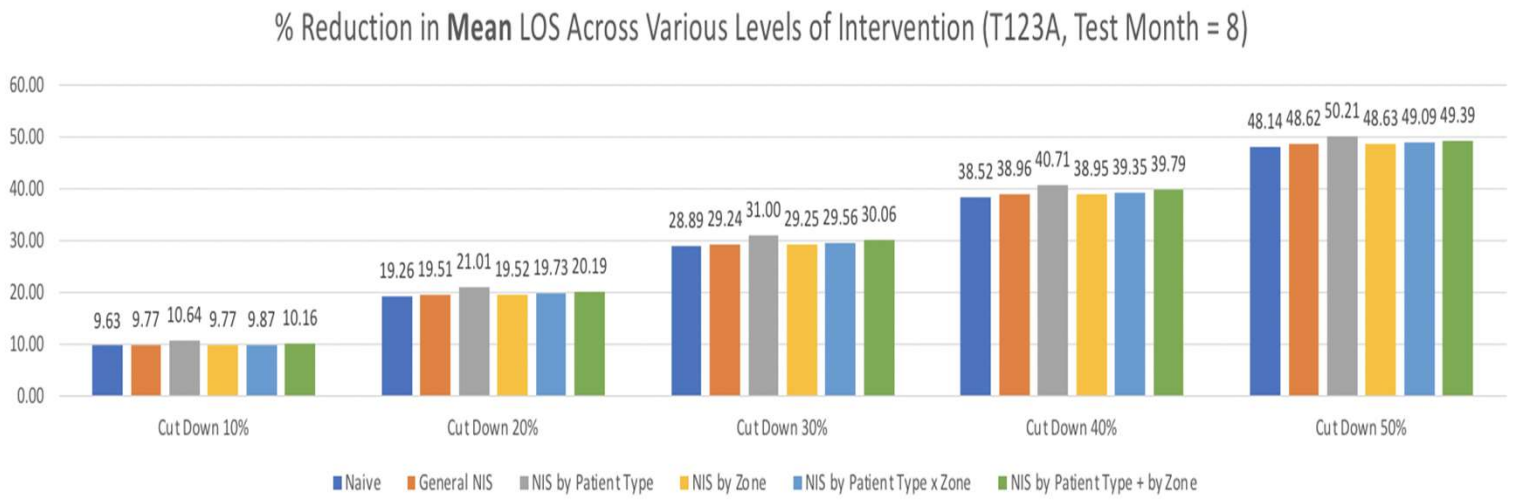
Apply Intervention*

If intervention is 10% cutdown:
LOS = LOS x 0.9

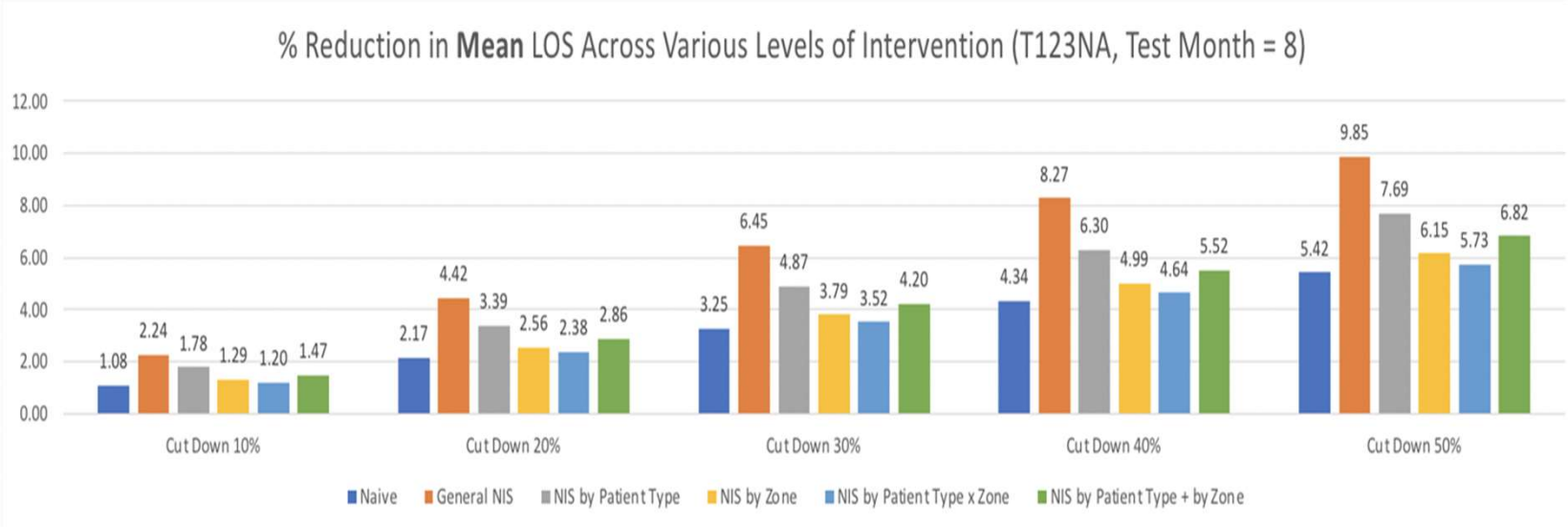
***Interventions**: 10%, 20%, 30%, 40%, 50% Cutdown in Consult Patient's LOS



% Reduction in Mean LOS (August, T123A)



% Reduction in Mean LOS (August, T123NA)



% Reduction in Mean LOS (August, T45)