Data Science for Processing Networks

A proposal to establish a Cornell-Technion Processing Networks Lab at Cornell Tech

Submitting team: Itai Gurvich, Shane Henderson and David Shmoys (Cornell Department of Operations Research and Information Engineering), Avishai Mandelbaum and Paul Feigin (Technion), Thorsten Joachims and Deborah Estrin (Cornell Department of Computer Science), Curtis Cole and Fei Wang (Cornell Weill), David Chokshi (NYU and NYC Health).

Date: May 7th, 2019

Data Science for Processing Networks

1 Introduction

Hospitals, courts and public transportation are service networks of central societal importance. Due to technological advances and public awareness, processes in such networks create (or could create) vast amounts of operational data, at unprecedented resolution and quality. Yet harnessing this data for publishable, reproducible and scalable research is an acknowledged and yet-to-be-overcome challenge. What we offer here is a blueprint for data-science labs, which will enable transformative steps to address this challenge: specifically, bringing together multiple scientific disciplines to the study of processing networks, which will support their design, planning, control, prediction and improvement along dimensions such as operational efficiency, quality of outcomes, and fairness of access.

This is therefore a proposal to develop infrastructure — physical (a *data lab*), human (a research team) and scientific (scholarly knowledge) — to advance data-science for processing networks in general, and healthcare systems in particular. The Processing-Network Lab, or PNLab for short, will serve as a data-based research hub which, among other things, will facilitate understanding, dissemination of and improvement upon best practices. PNLab activities will rely on analysis of its data and validation of models against this data; they will then foster the *creation from data* of novel models (e.g. statistical, mathematical, computational), scientific principles (e.g. congestion laws) and engineering solutions (e.g. of overcrowding problems in hospital emergency rooms or urban transportation).

Emergency Room (ER) overcrowding is but one manifestation of the challenge in accessing healthcare resources. This acknowledged access-crisis is the symptom of a severe mismatch between (or mismanagement of) service-demand and resource-capacity which, in fact, has plaguing much of the service sector: e.g. tele-services (phone, chat, internet), judicial ("justice delayed is justice denied"), financial (from queues in bank-branches to queues in high-frequency trading) and government (e.g. DMV and USCIS). Such congestion-prone services are all potential data-partners of our proposed PNLab, which will support evidence- and science-based improvements of their operations.

Our model for a data-based PNLab is also novel. Its *data* will serve as the "language" of multi-disciplinary research and the "bridge" between academia (research) and industry (applications); and it will focus on *processes* (e.g. customer journeys in congested service systems) and causalities (e.g. identifying and managing bottlenecks). Thus, together with PNLabs that will follow our model at other US institutions, and partnering with SEELab at the Technion in Israel (see SEELab website), we envision a game-change in how research, teaching and innovation in Operations Research and Management are practiced. Such a strong overarching statement requires concrete justification. To this end, we focus on one dimension of ER overcrowding (among many others), namely capacity planning, or specifically determining appropriate numbers of physicians and nurses (since salaries in service systems are typically the largest component of operational costs).

How many doctors and nurses? Common practice for determining ER capacity levels builds on experience or rules of thumb or ad-hoc improvements, all possibly originating in simple models. Many of those models appear in the scientific (predominantly operations research) literature; they were inspired by data but then do not "re-visit" the data, let alone the system that originated the data, so as to impact practice. A principled approach that starts from data, builds models, refines and validates them via simulation, and ultimately implements them and tests their recommendations against reality, would dramatically improve the state of affairs.

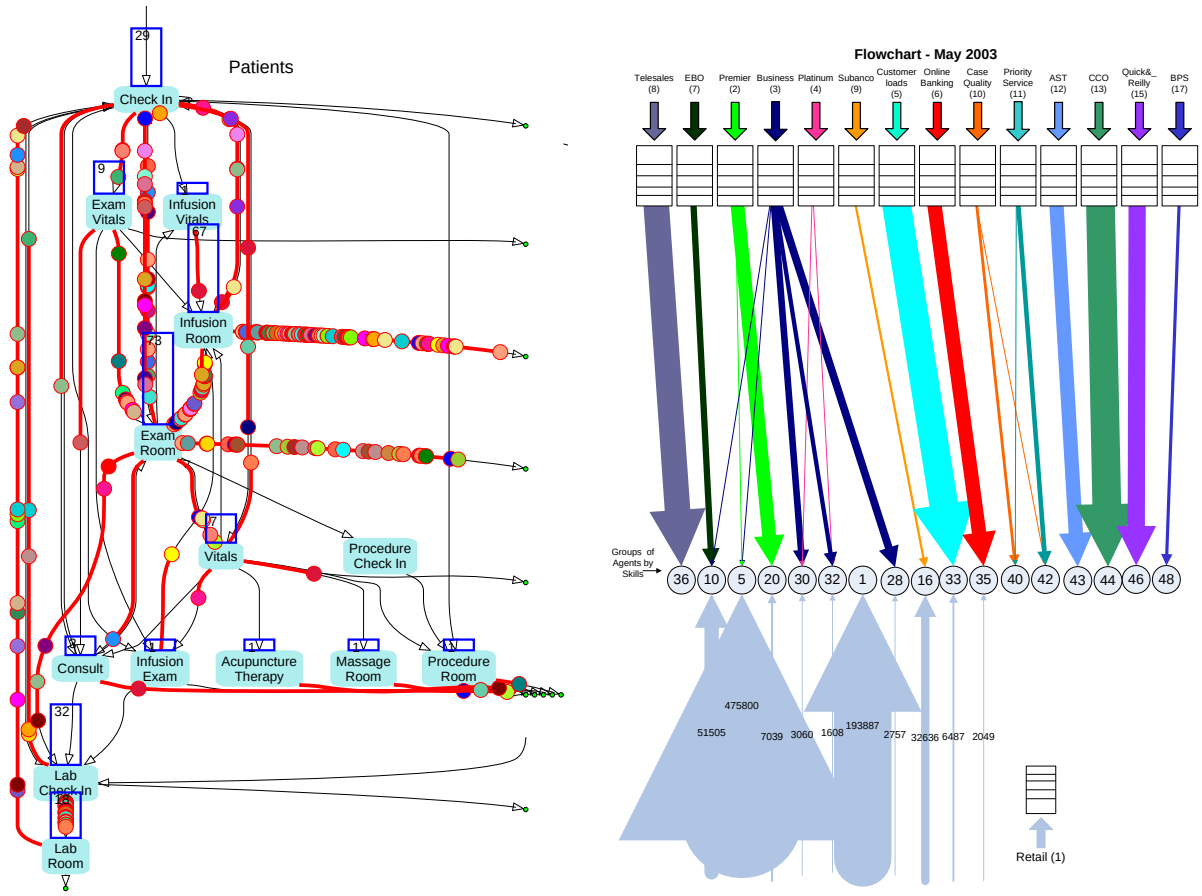

Figure 1: Data-based graphs of processing networks: (LEFT) Patients in an ambulatory cancer center, and (RIGHT) Customers in a bank call center.

In the specific context of ER staffing, [64, 98] demonstrated a 20% increase in the fraction of patients that see a physician within 30 minutes of their arrival to the ER. This demonstration is against a detailed validated simulation, but it stops short of validating the staffing rules against ER reality. Nevertheless, even such partial efforts are rare, and for a good reason: in isolation for a single facility, the efforts required (from gathering data to implementing research findings, or even to validating via simulation) are significant and hence not scalable. PNLab will offer the infrastructure to help make such efforts quite standard, and thus broadly feasible and more frequent. We shall achieve this by making research-ready data (or proxies thereof) and tools (e.g. data-based simulations) available to the broader community, of both researchers and practitioners. And the vision scales further: just as there are multiple labs, say, for solid-state physics, there should be multiple labs for processing network science, and here we are offering to create *a model lab* that others can replicate. PNLabs could share technologies and infrastructure (e.g. proprietary ER data that will be shared via corresponding ER simulations); while each lab will be of an intermediate size where scale suffices to support the required efforts but, at the same time, it is not too large to prevent the direct engagement with data and science by its individual researchers.

Data-based (models of) processing networks: In this proposal, it is convenient to have "processing

network" refer to both (the operations of) a service system as well as to models that capture it. More concretely, when referring to the management of a service operation, one often refers to managing its underlying processing network, or specifically to the orchestration of the activities/flows and resources of which it consists. For example, a processing network corresponding to an outpatient cancer institute could model patients flow: inputs are patients arriving for a chemotherapy session and outputs are treated patients; activities include blood test, exam by a physician and infusion. Figure 1(LEFT) is a *data-based* snapshot (it depicts real data) of the flow of patients in such a cancer institute. The resources — infusion nurses, physicians, blood-draw nurses, rooms — appear, from left to right, in the following data-animation. In call centers, an input could be a customer's call and the output is a served customer; see Figure 1(RIGHT). Here processing complexity arises from the multiplicity of customer types (rectangles at the top), groups of agents with different skills (numbered circles), and the algorithm that matches types with skills in real-time (which determines arcs, namely network topology).

Striving to cover the full cycle of service: Figure 1, we re-emphasize, is *data-based* — see the actual time-of-day at the bottom-left of the corresponding animation. This animation can be thought of as that of a single processing network with five resources or, alternatively, of five such networks. Indeed, the modelling scope can be as focused as *a* doctor's clinic or as broad as a public healthcare system where activities include appointments with doctors, urgent care visits, and re-admissions.

Scope- and data-wise, we shall strive to cover the full cycle of service. In call centers, the concept is *life-time experience/value of a customer*. In healthcare, this time-horizon is referred to as the *full cycle of care*: it starts with onset of symptoms, and it ends upon recovery (or other reasons for a system departure, possibly death); its data integrates in-hospital activities as in Figure 1(LEFT) with post-discharge activities that are recorded through traces in billing data or mobile-health devices (e.g. [1, 10, 25]). Similarly, life-time-experience data is formed from out-of- and in-call-center data. The approach clearly generalizes beyond call centers and hospitals, to form a complete processing network schema that would be relevant to most congestion-prone services. This generality calls for a proposal that is both abstract (refers to and uses the terminology of processing networks) as well as focused (demonstrating the abstract with concrete examples). Space constraints, however, render this ideal presentation infeasible — we thus opted to focus. We chose healthcare in view of its societal and economic significance (about 18% of US GNP), and the fact that many of this proposal's challenges (e.g. access to data, gaps between research and practice) are in fact magnified when viewed through "healthcare eyes". We hasten to emphasize that only our presentation is healthcare-focused, while our proposed PNLab will seek and maintain a wide variety of data-partnerships.

Data-based science & engineering of healthcare/processing networks: Challenges vary across healthcare settings: for example, it sometimes suffices to only vary capacity in order to improve access while, in other cases, a complete re-thinking of flows is required; see [16] for novel efforts in New York City's public health system to match the high demand for specialists with their scarce supply. Regardless, processing network data must be the basis of research, which will then support science, engineering and management; specifically measuring, modeling, analyzing, planning, controlling and accurately predicting operational performance. To this end, one must standardize data collection and infrastructure. And one must have benchmark data as those available for computer vision [2] and natural language processing (NLP) [4]. Like those other types of data, ours requires specialized tools. Importantly, the data does not come ready for use. The data and software that generates it (e.g. EHR or RTLS) are designed for practice and not for processing network research.

Transforming real data into processing network data is a non-trivial and (hugely) time-consuming task for which a framework and software infrastructure must be developed, particularly if analyses are to be done at scale.

The agenda we put forth thus has three meta goals:

i. Build a *physical and human infrastructure* for management (collecting, processing and disseminating) of *data* to support progress in the science and engineering of processing networks.

ii. Develop *frameworks* for data-based science and engineering of service/healthcare processes.

iii. Support data-based *engineering solutions* for fundamental operational problems in service / healthcare delivery.

Why be optimistic? The ultimate test will be the extent to which the proposed PNLab will inform management of say healthcare processes on how to improve access, reduce delays and, eventually improve outcomes. The figures and animations used throughout this proposal were generated via a platform developed at the Technion in Israel, specifically in its Service Enterprise Engineering Lab (SEELab). The success of SEELab, and our ongoing collaboration with its founders — they are key personnel on this proposal — provide a strong starting point. Then, we believe, scale and scope of *data science for processing networks* will significantly expand through access to ample new data sources, the engagement of local researchers *from multiple disciplines* (OR, medicine, and computer science), and an ultimately-created *network of collaborating PNLabs* that will follow our model. Our goal in this project is to provide the data schema and simulation models that will enable higher level data science modelling and engineering — see for example the divisions of data science as expounded in [30].

2 Data

Aggregated data (e.g. summary statistics) of processing networks are common and relatively easy to access. In contrast, it has proved a grand challenge to access data as required for our proposed PNLab, let alone using it to support publishable and reproducible research. This data constitutes time-stamped operational histories, features and protocols, at the resolution of individual events; and the challenges stem mostly from data heterogeneity, confidentiality (required NDAs or IRBs), acquisition technologies (e.g. EMR and RTLS in hospitals, ACD and CRM in call centers), and a lacking infrastructure to manage and analyze the data once acquired. PNLabs are our proposed solution to the latter challenge, exploiting the substantial commonality in the data footprint that different processing networks leave.

Indeed, ERs and call centers all have alike customer arrivals, service starts and ends, and queue-abandonment due to excessive waiting (ER patients leaving-without-being-seen or call-center customers hanging-up); and these events are typically endowed with *features* (clinical, financial, environmental, ...), and they are determined by *protocols* (e.g. scheduling, routing). One can hence abstract this data-commonality into processing-network models, that are capable of capturing the operational dimensions of any contemplated service systems. PNLabs will thus collect data (elaborated on in the present section), then use it to create basic processing-network models (Section 3), through its locally-developed technologies and tools (Section 4), for example automatic creation from data of simulation models.

A data model: Being more concrete and somewhat formal, the operational history of a processing network can be captured by event sequences (event logs) $\{(\tau(c,i,k,j), \text{act}(c,i,k,j) : j = 0, ..., J\}$ of successive times and activities of fork k of entity i of class c — times indicating the time of entry into an activity such as a queue, a service, or a transit area (e.g. corridor) at the completion of a service. Classes of entities may correspond to patients, providers or other resource types (e.g. a telemetry unit). A fork arises when an entity has "sub-units", such as when blood is drawn for a test: the patient remains in bed while the blood sample is routed elsewhere in the hospital. Creating such an event log from more basic, often noisy, data streams is often a challenge in itself and a subject of research.

In addition to these *observed* timed-events data, we may also have features data, $u(c,i)$, (clinical, operational, etc) for each entity, e.g. the timing of a scheduled appointment and/or the age and weight of a patient. We denote by X_t the collection of all timed-events data up until time t and let $X = X_T$ denote the complete events data for the whole period — say one day.

One typically has a collection of such data sets: $\{X^1, ..., X^m\}$ over m days. The days with their characteristics (day-of-the-week, holiday, weather) form the data $\{(X^i, d^i) : i = 1, ..., m\}$. Given X, we can compute performance measures $P(X)$ such as average waiting time at specific locations (possibly for each specific class of patients or providers); or the average length of stay (LoS) in the system for patients of specific types; or the utilization of particular resources; see Figure 2.

Often, the measurement and improvement of operational measures are linked directly to clinical ones. For STEMI patients who seek emergency cardiac care, clinicians measure the fraction of patients that exceed a threshold of 90 (or even 60) minutes from door-to-balloon (or door-to-treatment). This purely operational measure is used as proxy for clinical outcomes such as the likelihood and amount of cardiac muscle damage; see below for further discussion of clinical data.

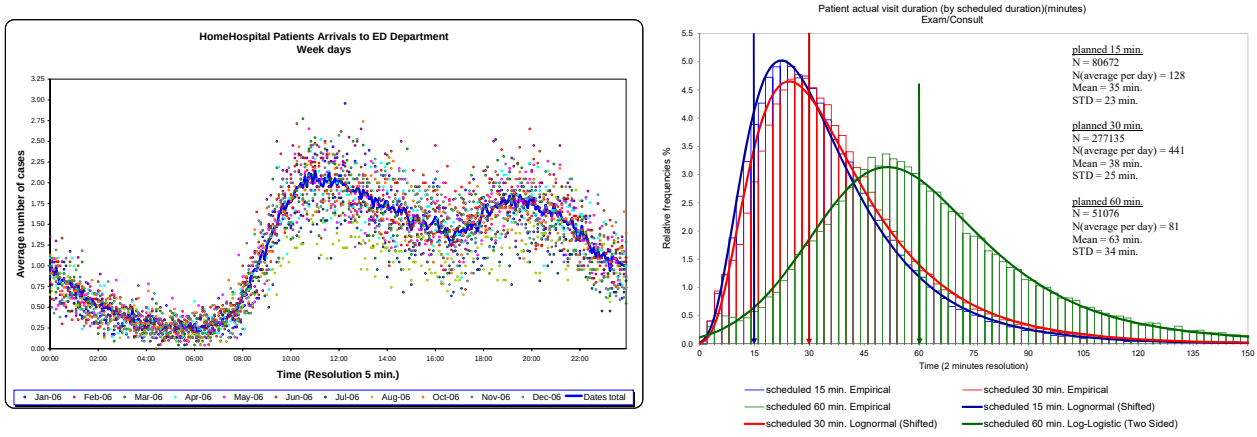

Figure 2: Processing network primitives: (LEFT) Patterns and variability of arrivals to an Emergency Room; (RIGHT) Duration of medical exam in a cancer institute, for different scheduled durations.

Transforming the data generated by a hospital's or a clinic's software into one that allows for scientific use is a non-trivial task; see further discussion below. Doing this *at scale* requires the development and implementation of data-science infrastructures. The data acquisition, integration, and standardization operations will be hosted in the Service Enterprise Engineering New York City (SEENYC) lab that will be established as a prototype PNLab by the team submitting this proposal.

The lab's data infrastructure will provide the sandbox in which to try new machine learning methods, develop mathematical and simulation models and derive engineering solutions. The potential of this endeavor should be contrasted with the way much healthcare-operations research has been carried out up to this point: using a single data set which is acquired at significant effort but, at the end, it is one data set with its potentially idiosyncratic characteristics.

> Objective 1: Develop tools for the scalable acquisition, pre-processing and dissemination of service/healthcare processing network data.

A clarification is needed at this point concerning *data sensitivity*. The data we seek is *mostly* (but not purely) operational: it is hence less sensitive than clinical data and, with suitable precautions, can be made accessible. Nevertheless, experience suggests that institutions may be hesitant in making their data publicly available. It is precisely one of the scientific missions of the PNLab to provide mechanisms to share proxies of the real data on which research can be conducted, even if the original data is not shareable. This can be done, for example, by generating artificial data, or using simulation models which generate data that share sufficient fundamental characteristics with the original data to allow for developing and testing of engineering solutions.

Data Structures and Pre-Processing: Data for a service system would include multiple tables; [3] contains, for example, the description of the data structures for the call center in Figure 1(RIGHT). Tables are required that document both patient-relevant and resource-relevant time stamps (e.g. when does a nurse take a break). The data for the cancer institute in Figure 1 includes all the appointment-related time stamps (scheduled and actual), whether an appointment was cancelled or re-scheduled. As a patient cannot start treatment without the chemotherapeutic agents, a full mapping of the flows requires also all pharmacy related time stamps (when an order was placed, when was it ready, etc.). It takes significant human effort and computational power to pre-process these vast data sets, so as to create *analysis-relevant* data structures that support Exploratory Data Analysis (EDA). The pre-processing, in itself, may require the development of new inference tools. The RFID tags from the cancer institute, for example, make available the location of individuals (patients and providers) but not the activity being conducted. A translation (activity mining) step is needed; see [76]. Once the data is collected and processed into analysis-relevant structures, it can support the creation of informative visualizations and aggregate statistics for accessible EDA before any more detailed statistical analyses are executed; see Figure 2 and data-animation for a dynamic view of the flows in a hospital.

Integration of operational and clinical data: It is critical to understand how decisions with operational implications affect outcomes. How, for example, the number of nurses may affect outcomes for patients visiting the Emergency Room (ER)? A decision to run a test (say an ultrasound) consumes resources (a technologist, radiologist and ultrasound equipment) and increases the cost of care but may or may not improve outcomes for certain patients. In order to draw conclusions about the relationship between process and outcomes one must control for patients' clinical characteristics such as primary diagnoses (DRG), weight, age, co-morbidities, severity scores, etc.

The medical informatics community has also been exploring healthcare pathways and their connection to outcomes (see e.g. [59] by one of the co-PIs of this proposal). This pathways view is based on data from Electronic Health Records (EHR) and captures, in our network language, only those activities that leave a footprint in the EHR. Such data does not fully capture the resources

and/or the resource consumption (quantities and timing) and hence cannot then serve the important objective of exploring the *cost of care* and the operational implications of clinical decisions.

Clinical (specifically diagnostic) information, when combined with the operational data, allows for the the study of resource utilization patterns for different disease conditions and, in turn, of the *cost of care*, overall and for each diagnosis; see [11].

The connection between operational and clinical data goes both ways. In the age of precision medicine, clinical data is used to "customize" a process for a patient. Inferring prioritization protocols, requires then linking the operational data to the clinical and demographic data. Fairness and equity in care is an important area of study [12, 13, 14] where this integration can have significant impact. Measuring (in)equality is difficult. Biases of decision makers are implicit and are made visible only through the actual activities performed on/for a patients (e.g. performing an MRI on a patient). Comparing the "customer journey" of patients from different ethnicity groups while controlling for their clinical data may allow us to uncover unfairness (where it exists).

The mapping of operational and clinical data is non-trivial and constitutes a "barrier to entry" for the study of such relationships. The lab will be the integrator, whether it is for its hospital partners or for researchers (subject, naturally, to the necessary protocols of data-access permissions such as those determined by IRBs, NDAs, etc.).

3 Models: The mathematical language of processing networks

The natural language for modeling healthcare delivery processes is that of processing network models and queueing theory. The term and concept of processing networks was advanced by Harrison [43, 44, 42] for what we call here simply processes: a system that takes inputs of various kinds and uses various processing resources to produce outputs of various kinds. Such a network model provides a powerful abstraction of real-world systems. We take some freedom in extending this term to cover a variety of networks (and network features) that were since added to Harrison's original framework. Readers are reminded that we use "Processing Network" when referring to the original *system* of interest (e.g. the emergency room); and we shall refer to the processing-network model when referring to the processing network's modeling abstraction.

Let X be our processing-network data derived from the event-log data, and from which, readers may recall, one can compute (a vector of) performance measures of interest $Y = P(X)$. The available data would typically include clinical and/or demographic information U associated with the individual patients — information that would also determine each patient's class. It is useful to think of X as a function of (i) design variables ($\mathcal{D}$) (e.g. number of doctors/beds, routing and prioritization protocols), (ii) inputs (U) (e.g. clinical characteristic of the arriving patients on the focal day) and (iii) randomness (ω) (e.g. the arrival times of customers, their service requirements — each of which might be a random function of their clinical characteristics as captured by U). That is, $X = S(\mathcal{D}, U, \omega)$, where $S(\cdot)$ stands for the *true system* function.

We can consider possible data-generation models M as approximations for the true data generation process. The model, $M(\mathcal{D}, U, \omega')$, could be a simulation model, the outputs of which are based on generating a collective realization ω' for all the arrivals, service times, transition matrices, and possibly for U itself, and which are based on distributions fitted to data, presumably observed over many days. A detailed simulation model could produce streams of customer sequences, with times of arrival, service start and departure, mimicking the format of the event log X of the original data . The question of model adequacy is how well the performance measures $P(M)$ coincide with

$P(X)$. At another level, can we produce a model to predict for individual patients? For example, how long will the patient of class c, arriving at time t have to wait to be treated? See [75] for a machine learning approach to such prediction.

Another challenge is to use a model to evaluate the effect on the system of a change in the supply of resources (a design variable). The question would then be how well does the adjusted model $M(\hat{\mathcal{D}}, U, \omega')$ (with the adjusted design variables $\hat{\mathcal{D}}$), predict the performance of the adjusted reality $S(\hat{\mathcal{D}}, U, \omega)$ for relevant performance measures P. There are various modeling approaches of which simulation is just one. Machine learning is another (see further below) but it is often the case that interpretability is lost. Operations research models are simple but sometimes too simple to convince practitioners of their relevance. It is part of our objective, by taking data as the starting point, to support building models that are understandable enough to provide insight and confidence, yet complex enough to remain true to their reality. The modeling literature for healthcare processes is vast; see e.g. [40, 49, 27, 29] for books, book chapters and surveys on the topic.

In building a model one starts from the reality of the problem and tries to create a model that strikes a challenging balance between capturing the essential ingredients of said reality and being tractable enough to obtain useful insights and prescriptions. At the end of the modeling process one would (we, argue, one must) "close the loop" and explore the relevance of the model-based prescription *to actual data*. This "closing the loop" step has been largely absent from our literature and for good reason: adequate data sets have not been easily available to researchers.

Data should also be the starting point of model development. Part of the mission of this lab is to understand how *models can be built directly from data*. This is consistent (and will be methodologically supported by) the great advances in machine learning. Neural networks are models of the data whose structure they "learn". A data-based framework for processing networks will produce models of different levels of complexity that support predictions and counterfactuals.

> Objective 2: Develop the theoretical foundation and the software infrastructure for the creation of process models from healthcare process data.

This goal differs fundamentally from the prevalent practice where models are remote from data and it differs from the practice of proposing a model whose *primitives* are estimated from the data — modelling the triage step as a single-server $M/G/1$ queue with service-time distribution estimated from the data is different from "asking the data" what is the best model for the triage step. The development and validation of models (for prediction and prescription) directly from data require advances in stochastic modeling theory and in supporting statistical theory.

Simple models in the service of a complex reality. Given the complexity of processing networks, the approach is often to "collapse" the complexity into simpler models, approximations that, the modeler believes, capture essential ingredients of reality. In this way, the basic Erlang-A queue — a queue with Poisson arrivals, exponential service times and exponential customer patience — has been valuable in the study of complex call centers as the one in Figure 1; see [37]. An extension of the Erlang model to retrials, referred to as the Erlang-R model, has been useful in the study of ER's; see [28, 99] as well [74] for priorities in emergency rooms.

In addition to discrete-state-space models, the theory covers "fluid" and "diffusion" approximations (the process analogues of the strong law of large numbers and the central limit theorem)

that can accommodate service characteristics (e.g. randomness, transience, heterogeneity, fork-join, fairness) and resolutions (from mass customization to flow aggregates), which are otherwise intractable; see [95, 26, 97] for some notable examples. These can be used for bottleneck analysis in complex processes (see e.g. [41]) as well as more refined analyses targeted at understanding the structure of waiting times and their optimization; see e.g. [51].

Simple models facilitate, through the dimensionality reduction that they bring, the *inference, analysis and control* of processing networks; see e.g. [38] and [63, 84] for network inference problems.

A beautiful illustration of the asymptotic approximations' value for inference, is Reiman's "snapshot principle" [72]. In a complex network of queues, consider predicting the sojourn time of a customer/patient—the time from the moment the patient enters the network until the patient leaves. This is a complex prediction task. The focal patient's sojourn time depends on the future paths of all patients currently present in the network. A prediction based on the snapshot principle uses only the likely path of the focal patient and the number of patients waiting in each buffer (waiting for blood test, waiting for imaging etc.) on that path. No other information is required; see [36] for an implementation to travel-time prediction in a transportation network.

4 Science and Engineering

The total vision of this proposal, as it pertains to data acquisition, management and analysis, is summarized in Figure 3. The framework proceeds with: 1. the acquisition of data and its conversion to a processing network format; 2. use of a statistical engine and a graphical engine to model the building blocks of the network; 3. creation of a data-based simulation that will serve as a virtual reality of that network; 4. creation of corresponding mathematical model(s); 5. use of the data to validate the mathematical models.

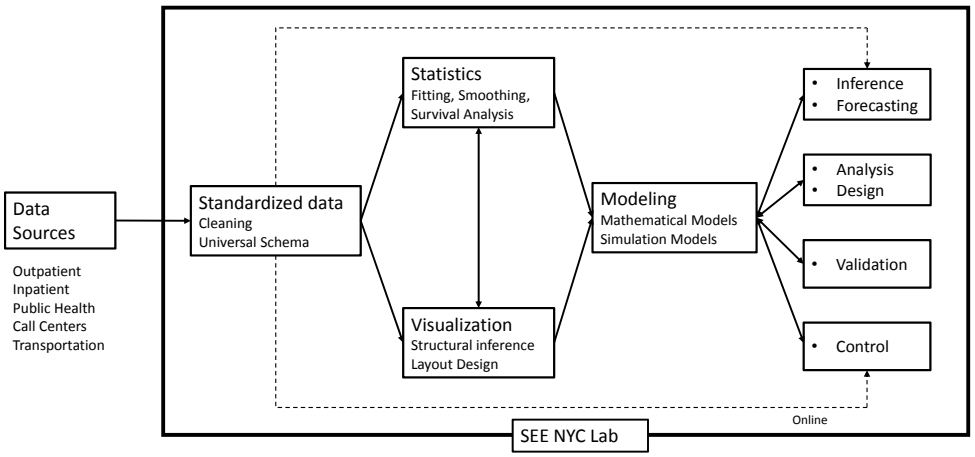

Figure 3: A schematic for a framework: Data Science for Processing Networks

The classical scientific paradigm is to experiment, measure, model, validate, analyze, refine and so on. It is traditional and routine in Biology, Chemistry and Physics, and recently also adopted in Economics and Network Science. We are proposing to do the same with processing networks, which calls for the development of novel theory which interfaces with measurements and experiments.

The science should consist of algorithms to mine processes from data, statistical methods to test models against data, machine learning and causal inference methods to build models for performance prediction. The engineers should then use these models to develop solutions (flow designs, capacity planning rules etc.). To make this happen we need (a) methods to test models against data and (b) ways to generate new models from data as discussed above.

This scientific pursuit will inevitably lead to new bridges between machine learning (ML) and Operations Research (OR). In contrast to mathematical queueing models, those generated ML algorithms are conceptually closer to reality as they arise directly from the data. These models must be regularized to prevent over-fitting and may often be limited in the operational insight that they can provide: they suffer from, what is often referred to as, the "curse of dimensionality".

The process models of OR are mostly mathematical and hence, necessarily, simple relative to the complex realities that they seek to approximate. They suffer from the "curse of simplicity". They do provide ample insights but it remains the responsibility of the researcher to establish the practical relevance of this insight. These models if shown to practitioners (say ER managers), without data-based evidence, will be deemed as irrelevant.

The gap between machine learning and operations research gives rise to ample opportunities. ML can help OR with its "curse of simplicity" by validating and hence establishing relevance of OR models. OR can help ML address the "curse of dimensionality" by offering robust models that reveal structural properties that have prescriptive implications. In order to be more concrete, let us revisit the problem of emergency room staffing. The length of stay, an important performance metric (that has clinical implications), depends also on the patient's waiting times for tests, doctor visits etc. which in turn are functions of how well the emergency room is staffed. The operations research version of tackling the staffing problems is one that tries to (1) approximate the complex reality by a simpler model and (2) propose staffing levels based on this approximate model. A researcher trained in statistical learning might approach this question in a completely different way by trying to learn a "good" staffing algorithm directly from the data; an algorithm that — given forecasts of demands for a day — identifies how many nurses, residents etc. to assign to each shift in order to achieve a certain performance in terms of waiting time of patients. To what extent can one learn a good decision by just considering past data of these systems? Can the structural understanding from OR models help in this learning?

> Objective 3: Develop statistical tools for forecasting, model construction and counterfactuals on processing networks.

Forecasting: Methods and testing. Forecasting methods have a long history and significant practical impact. Call-centers served here, again, as a sandbox for the development of methods [52, 9] and there are now papers that focus specifically on healthcare; see e.g. [96].

To develop and test forecasting methods requires data sets (more than one) against which to test these methods. In the context of service systems this testing has been historically done on a single, or at most a handful, of data sets. A comprehensive approach to processing network models will allow one to measure the quality of forecasting, not just against the input data, but also in a *performance-relevant way*. The quality of a forecasting method for ER arrivals should be judged, for example, by measuring the patient waiting times when staffing is done based on the proposed forecasting method; see e.g. [65, 102, 93, 94] and references therein for papers that study forecasting for emergency medical services (EMS) calls and ambulance travel times. The first in this list, [65], tests its proposed forecasting method against both standard quality-of-forecast metrics such as root

mean square error (RMSE) and via performance-relevant metrics using a queueing system to model the responsiveness of ambulances. An infrastructure like ours will support executing the latter test against data rather than models.

Model construction and validation: Constructing models directly from data raises both philosophical and complex scientific/engineering questions.

Mathematical models: It is our objective to develop (and support the development of) methods to build mathematical and simulation models directly from data. Given emergency room (ER) data, one should be able to build a processing-network model that reflects the ER's essential ingredients and that is a good "predictor", say, for the waiting time for a physician.

Queueing theory provides mechanisms to restrict attention to a subset of tractable and interpretable models. Even when "venturing" outside the world of queueing models, this theory highlights certain properties, or "congestion laws", that any good model must obey. One can think of these laws as a meta model — the rules of queueing's "programming language". It is known that, in great generality, as the arrival rate increases (with all else being fixed, and without customer abandonment) the waiting times grow as $1/(1-\rho)$ where ρ is the utilization; see Figure 4. This implies, in particular, that linear models are not the "right" family to use here. Early examples of inferring a network from data can be found in [48, 63]. Recent work [50, 76] focuses on the inference of the network from location data like the one underlying the data in Figure 1(LEFT).

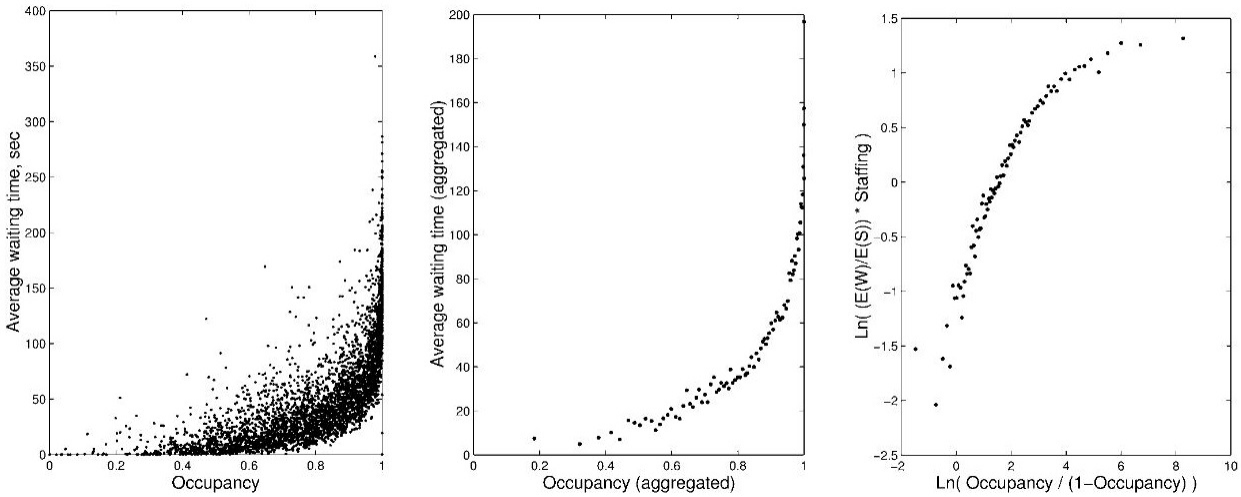

Figure 4: Congestion laws in data: Average wait increases as $1/(1-\rho)$

Simulation models: Often, to convince practitioners, a credible simulation model is needed that is more detailed than the mathematical model. A simulation model can also be used to validate insights, and compute performance metrics. Inherent to building simulations is the "difficulty of representing the complexity of hospital activity within a simulation model that must, like all models, be a simplification" [39]. Some papers propose or hint at a general approach; [79] is a good example for emergency rooms. It would be of great practical value if such simulation models could be built reliably from data to support operational improvement in a timely way. Such a model could follow the same basic logic of the mathematical model, so it can be built using the same techniques used to

produce the mathematical model. The modeling primitives (patient interarrival times, processing delays) can then be used exactly as they appear in the data and various performance metrics can be computed. This is called "trace-driven simulation;" see [60, Chapter 5].

Trace-driven models are not always suitable for predictions. For example, it is not clear how one can use trace-driven simulation to model bursts in the arrival process of patients? To that end, we intend to also explore the development of *parametric* simulation models that replace data sequences with parametric families of stochastic processes. For example, a sequence of processing times can be modeled as i.i.d. and coming from a given parametric family such as the gamma family. Recent work by Xiaowei Zhang, Zeyu Zheng and Peter Glynn that is yet to be made broadly accessible suggests that such modeling, in conjunction with careful injection of appropriate macro structure related to dependence structures in the data, can accurately reproduce the key performance measures we hope to model. Our methodologies would enable us to test such conjectures in healthcare data sets.

Models (mathematical and simulation) are only useful if their predictions are aligned with the effects seen in practice. An informative example of such comparison in healthcare appears in [100]. However, there is a difference between *a* simulation model that validates well against a given data set *and building generalizable knowledge and methods for constructing simulation models from processing network data.* The latter requires the access to multiple data sets.

Machine learning for model acquisition. Machine learning has shown great success in modeling complex dependencies in a wide range of real-world applications, and we will employ Machine Learning to fit more realistic models of processing networks for both direct optimization and simulation. On the one hand, we conjecture that this will lead to models that capture the true processes better than existing models. On the other hand, however, it also implies new challenges to the stochastic optimization procedures that now have to work on top of models as complex as a deep network, without a concise mathematical model that can be studied analytically. Nevertheless, recent demonstrations of the progress in reinforcement learning (e.g. [77]) have shown that deriving high-performing policies optimizing such models is feasible, breaking some of the long-standing challenges in artificial intelligence.

Machine Learning algorithms and methodology will play a role both in model acquisition, as well as in the evaluation of the learned models and the policies derived from them. One key aspect of a data-driven approach is the need to repeatedly collect data, since the data we collect is dependent on the current policy that operates the system. For example, by changing the scheduling policy in an ER department, we may get into states of the system that never occur under a different policy. This highlights that we are dealing with a learning problem that is fundamentally causal and where the optimization of policies is inherently counterfactual to the data we have collected in the past. This likely leads to a cycle of modeling and policy optimization, with deployment and new data collection as part of the process. This iterative process is well understood in other industrial applications (e.g. search engines, recommender systems) [87, 88, 57], where actions are atomic and limited to the contextual bandit model. However, the complexity and sequential nature of actions in processing networks pushes the state of the art, while still not reaching the full complexity of reinforcement learning as common in autonomous robotics. This opens the way for extending existing and tractable techniques from the contextual bandit setting to processing networks.

Machine learning for policy generation. Machine learning also offers a fundamentally different approach to optimizing policies that does not require an explicit model that provides the

basis for optimization. Such model-free reinforcement learning [85] does not estimate the model, but directly optimizes the policy and typically performs some form of stochastic gradient descent in the policy space. Policy-gradient approaches are particularly useful when the system is highly complex and difficult to learn, but where there may be a comparably simple policy that performs well. For example, learning a model of the utility that a user of a movie recommendation service derives from any particular movie is typically more challenging than learning a good recommendation policy [87], partly because even small errors and bad extrapolations in the estimated model can lead to large errors in the policies that are derived from them.

Particularly relevant are off-policy policy-gradient methods that reuse existing stochasticity in the data without the need for adaptive and interactive experimental control. By correcting the sampling bias in the logged data that is used for training, off-policy learning approaches have been successful in the recommendation and search-engine domains [57, 56, 6]. We anticipate that similar approaches can also fundamentally change how we derive policies for processing networks, in particular on well-defined questions that fit a contextual-bandit framework (e.g. how many nurses to staff based on context features like weekday, weather, holidays, flu season, etc.). Correction of sampling biases again relates to causal and counterfactual inference questions. However, the actions that are taken in processing networks are largely under the control of a known policy during logging, which makes the data more experimental than observational and allows - at least in principle - control for confounding through explicit randomization.

5 Intellectual merit and broader impact

5.1 Intellectual merit

We propose an infrastructure to advance the *science* of processing networks. Centered on data, it will bring together researchers from different disciplines (operations research, medicine, computer science), not just to leverage existing knowledge, but also to highlight gaps in said knowledge. For operations research, the data (and its science) will inform the building of new models, indeed, it will inform the very mechanisms through which models (mathematical or simulation) are built and tested — originating them in data and validating them against data. For clinicians and healthcare-policy researchers, this infrastructure offers a unique opportunity to investigate cost of care — more broadly the relationship between resource utilization and outcomes — using data that brings together the operational and the clinical. For machine-learning and statistics, processing-network data with its complex inputs (e.g. trajectories of patients) offers opportunities to re-visit and re-think learning methods for predictions and counterfactuals or reinforcement learning.

To achieve this, we built an interdisciplinary team of researchers with expertise that spans the mathematical modeling language of processing networks and their simulation, clinical expertise, ownership of and familiarity with Electronic Health Records covering inpatient, outpatient and public health, patient-generated data, and, finally, expertise in the theory of statistical learning with specific application to health-related data.

5.2 Broader impact

Hospitals, courts and public transportation are service networks of central societal importance. Managing these processes depends on the integration of individual patient/customer, operational, psychological and financial viewpoints. Such assimilation must be based on reliable patient-level

data and its evaluation calls for measuring quality, efficiency and outcomes of care processes, often in real-time, and ideally covering the full time-line from the onset of symptoms through hospitalization and readmissions all the way to recovery/mortality.

Our model for a data-based PNLab is novel. Its *data* will serve as the "language" of multi-disciplinary research and the "bridge" between academia (research) and industry (applications); and it will focus on *processes* (e.g. customer journeys in congested service systems) and causalities (e.g. identifying and managing bottlenecks). It offers the opportunity to change the way research, teaching and innovation are conducted in healthcare delivery (more broadly, in service enterprises).

Technology transfer: The proposed lab has the mission to support research, education and, not least, entrepreneurship and innovation for healthcare-process optimization. The PI is faculty at Cornell Tech where engagement with practice and practical innovation are embedded in the research and educational cultures. The start-up studio, a central part of the curriculum, is where students come up with new engineering ideas. A non-negligible number of the studio teams turn into actual start-ups upon the students' graduation. The proposed lab will support the identification of opportunities, the development of ideas and, finally their testing on *real data*.

Our collaboration with healthcare practitioners will create opportunities to validate solutions in actual clinical settings and to disseminate successful ones to the wider healthcare community as products and services.

Creation of data and software resources. In addition to making data sets (or proxies thereof) available to the research and practice communities, we will — after collecting enough data — share best-practices that we observe in the data. Hospitals, for example, can then compare their performance against those best practices. We will also provide data and simulation environments in which people can test algorithms and models (e.g. staffing, prioritization etc.).

Curriculum development. Processing-network theory has always been an integral part of the education of operations research and industrial engineering practitioners. Data, however, has been largely absent from most curricula. It is our view that data should become an integral part of said education. This lab, with its mission to make such data and simulations thereof accessible to the community, can serve this important objective. As evidence, this data has been integrated into service engineering and service-operations courses at Cornell and, before that, at the Technion.

Graduate education. The PI and co-PIs teach in institutions with sizable and high-quality PhD programs. This lab, through its scientific infrastructure and its unique set of partners and members, offers a unique opportunity to educate PhD students that are truly multi-disciplinary, bringing together data-science (machine learning, statistics, etc.) with operations research and operations management. The result would be a novel type of data scientist/engineer, one who is also familiar with processing-network theory and its application in a variety of important contexts.

Our PhD programs are diverse. More than 30% of Cornell ORIEs doctoral students are female, which provides an opportunity to promote women's (and other under-represented groups) work and visibility in the field. This diversity will be eventually reflected in the faculty in academic departments.

6 Results from Prior NSF Support

PI **Itai Gurvich:** A recent award from the NSF is CMMI-1662294 "Taylor Expansion Approximations for Dynamic Programming Problems," $349,971, 6/1/2017 - 6/1/2020.

Intellectual Merit. Dynamic Programming (DP) is the fundamental operations-research tool for solving optimization problems where decisions are subject to changes in real time. The research proposed here expands the decision makers toolbox by offering a new structured, implementable approach to the approximation of, otherwise intractable, dynamic decision problems; see [15].

Broader Impact. The proposed agenda seeks the development of practical algorithm and their implementation in domains where dynamic optimization is needed. Training includes one PhD student and one undergraduate student.

Co-PI **S. G. Henderson** Two previous awards with a start date in the last 5 years: CMMI 1537394 Stochastic Optimization Models and Methods for the Sharing Economy, $200,000, 9/1/15 – 8/31/18; DMS 1839346 TRIPODS+X:RES: Collaborative Research: The Future of the Road - A Data-Driven Redesign of the Urban Transit Ecosystem, $425,000, 10/1/2018 – 9/30/2020. Of these, the first is more relevant here.
Intellectual Merit: Advances in simulation [54, 55, 58, 19, 20, 17, **?**, 18, 35, 70, 71, 80, 91, 92], with applications in bike sharing [78, 46, 34, 31, 32, 33, 53] and emergency services [22, 23, 24, 62, 67, 66, 21, 73].
Broader Impact: We have a close ongoing relationship with Motivate, a provider of bike-sharing systems across the USA. A repository [47] of simulation-optimization problems and algorithms has been developed for empirical comparison of algorithms. "The Optima Corporation" has adopted some of our ambulance redeployment methodology. Training includes 1 postdoc (Susan Hunter, now on Purdue IE faculty), 9 PhD students, 15 M. Eng. students, 12 undergraduate researchers, and 10 workshops for female/URM prospective undergraduates.

Co-PI **F. Wang** is the PI of "CAREER: Interpretable Deep Modeling of Discrete Time Event Sequences" (IIS-1750326, 7/01/18-6/30/23, $539,642).

Intellectual Merit: The project includes a series of research on developing interpretable models (e.g. knowledge injection, knowledge distillation and knowledge transfer) for sequence based deep learning. The specific focusing application area of this project is analyzing longitudinal patient electronic health records.

Broader Impact: The proposed research will be incorporated in to the contents of the course the PI is teaching. The PI will also reach out to K-12 and undergraduate/graduate students to involve them in the proposed research. The project has already result in several publications [45, 69, 83, 81, 90, 101].

Co-PI **T. Joachims** is the PI of "III: Small: Collaborative Research: RUI: Batch Learning from Logged Bandit Feedback" (IIS-1615706 / IIS-1615679, 07/01/16-06/30/19, $516,000).

Intellectual Merit: Develops the learning theory and machine learning foundations for batch learning with contextual bandit feedback.

Broader Impact: Extended megsradio.fm, POEM and BanditNet software, Criteo dataset, organized NIPS16, NIPS17, ICML18, RecSys18 workshops, research with 5 undergrad and master students at Cornell and 12 undergrad students at Ithaca College (see [68] credits). [5, 89, 61, 86, 57, 56, 8, 7, 6, 82] ICTIR 2016 Best Presentation Award, WSDM 2017 Best Paper Award.

References

[1] Apple's new mHealth project takes on remote patient monitoring. https://mhealthintelligence.com/news/apples-new-mhealth-project-takes-on-remote-patient-monitoring. Accessed March 25th 2019. 4

[2] CVDF common visual data foundation. http://www.cvdfoundation.org/. Accessed March 25th 2019. 4

[3] DATA-MOCCA: Data Model for Call Center Analysis. volume 1: Model description and introduction to user interface. 7

[4] LDC: Linguistic Data Consortium. https://www.ldc.upenn.edu/. Accessed March 25th 2019. 4

[5] A. Agarwal, S. Basu, T. Schnabel, and T. Joachims. Effective evaluation using logged bandit feedback from multiple loggers. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2017. 16

[6] A. Agarwal, K. Takatsu, I. Zaitsev, and T. Joachims. A general framework for counterfactual learning-to-rank. In *ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 2019. 14, 16

[7] A. Agarwal, I. Zaitsev, and T. Joachims. Consistent position bias estimation without online interventions for learning-to-rank. In *ICML Workshop on Machine Learning for Causal Inference, Counterfactual Prediction, and Autonomous Action (CausalML)*, 2018. 16

[8] A. Agarwal, I. Zaitsev, and T. Joachims. Counterfactual learning-to-rank for additive metrics and deep models. In *ICML Workshop on Machine Learning for Causal Inference, Counterfactual Prediction, and Autonomous Action (CausalML)*, 2018. 16

[9] S. Aldor-Noiman, P. D. Feigin, and A. Mandelbaum. Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics*, 3(4):1403–1447, 2009. 11

[10] M. S. H. Aung, F. Alquaddoomi, C-K Hsieh, M. Rabbi, L. Yang, J. P. Pollak, D. Estrin, and T. Choudhury. Leveraging multi-modal sensing for mobile health: a case review in chronic pain. *IEEE Journal of Selected Topics in Signal Processing*, 10(5):962–974, 2016. 4

[11] J. S. Banthin, P. Cunningham, and D. M. Bernard. Financial burden of health care, 2001–2004. *Health Affairs*, 27(1):188–195, 2008. 8

[12] J. R. Betancourt, A. R. Green, J. E. Carrillo, and E. R. Park. Cultural competence and health care disparities: key perspectives and trends. *Health affairs*, 24(2):499–505, 2005. 8

[13] P. A. Braveman. Health disparities and health equity: concepts and measurement. *Annu. Rev. Public Health*, 27:167–194, 2006. 8

[14] P. A. Braveman, S. Kumanyika, J. Fielding, T. LaVeist, L. N. Borrell, R. Manderscheid, and A. Troutman. Health disparities and health equity: the issue is justice. *American journal of public health*, 101(S1):S149–S155, 2011. 8

[15] A. Braverman, I. Gurvich, and J. Huang. On the taylor expansion of value functions. *arXiv preprint arXiv:1804.05011*, 2018. 16

[16] H. Byrnes-Enoch, J. Singer, and D. A. Chokshi. Improving access to specialist expertise via econsult in a safety-net health system. *NEJM catalyst*, 2017. 4

[17] E. Cao Ni, D. F. Ciocan, S. G. Henderson, and S. R. Hunter. Comparing message passing interface and mapreduce for large-scale parallel ranking and selection. In L. Yilmaz, W. K. V. Chan, T. M. K. Roeder, C. Macal, and M.D. Rosetti, editors, *Proceedings of the 2015 Winter Simulation Conference*, pages 3858–3867, Piscataway NJ, 2015. IEEE. 16

[18] E. Cao Ni and S. G. Henderson. How hard are steady-state queueing simulations? *ACM Transactions on Modeling and Computer Simulation*, 25(4):Article 27, 2015. 16

[19] E. Cao Ni, S. R. Hunter, and S. G. Henderson. Ranking and selection in a high performance computing environment. In R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, editors, *Proceedings of the 2013 Winter Simulation Conference*, pages 833–845, Piscataway NJ, 2013. IEEE. 16

[20] E. Cao Ni, S. R. Hunter, and S. G. Henderson. A comparison of two parallel ranking and selection procedures. In A. Tolk, D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 3761–3772, Piscataway NJ, 2014. IEEE. 16

[21] E. Cao Ni, S. R. Hunter, S. G. Henderson, and H. Topaloglu. Exploring bounds on ambulance deployment policy performance. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, editors, *Proceedings of the 2012 Winter Simulation Conference*, pages 497–508. IEEE, 2012. 16

[22] T. A. Carnes, S. G. Henderson, D. B. Shmoys, M. Ahghari, and R. D. MacDonald. Mathematical programming guides air-ambulance routing at Ornge. *Interfaces*, 43:232–239, 2013. 16

[23] K. C. Chong, S. G. Henderson, and M. E. Lewis. The vehicle mix decision in emergency medical service systems. *Manufacturing & Service Operations Management*, 18(3):347–360, 2016. 16

[24] K. C. Chong, S. G. Henderson, M. E. Lewis, and H. Topaloglu. A bound on the performance of optimal ambulance redeployment policies in loss systems. Submitted, 2016. 16

[25] A. Coravos, S. Khozin, and K. D. Mandl. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *NPJ digital medicine*, 2(1):14, 2019. 4

[26] J. G. Dai and W. Lin. Maximum pressure policies in stochastic processing networks. *Operations Research*, 53(2):197–218, 2005. 10

[27] T. Dai and S. R. Tayur. Healthcare operations management: A snapshot of emerging research. *Manufacturing & Service Operations Management, Forthcoming*, 2019. 9

[28] F. de Véricourt and O. B. Jennings. Nurse staffing in medical units: A queueing perspective. *Operations Research*, 59(6):1320–1331, 2011. 9

[29] B. T. Denton. Handbook of healthcare operations management. *New York: Springer*, 10:978–1, 2013. 9

[30] D. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. 5

[31] D. Freund, S. G. Henderson, and D. B. Shmoys. Minimizing multimodular functions and allocating capacity in bike-sharing systems. In F. Eisenbrand and J. Koenemann, editors, *Integer Programming and Combinatorial Optimization Proceedings*, volume 10328 of *Lecture Notes in Computer Science*, pages 186–198. Springer, 2017. 16

[32] D. Freund, S. G. Henderson, and D. B. Shmoys. Minimizing multimodular functions and allocating capacity in bike-sharing systems. *Submitted*, 2017. 16

[33] D. Freund, S. G. Henderson, and D. B. Shmoys. Bike sharing. In Ming Hu, editor, *Sharing Economy: Making Supply Meet Demand*, volume 6 of *Springer Series in Supply Chain Management*, pages 435 – 459. Springer, 2019. 16

[34] D. Freund, A. Norouzi-Fard, A. Paul, S. G. Henderson, and D. B. Shmoys. (even) smarter tools for (citi)bike sharing. In *Association for the Advancement of Artificial Intelligence Proceedings*, page Submitted, 2017. 16

[35] M. C. Fu, G. Bayraksan, S. G. Henderson, B. L. Nelson, W. B. Powell, I. O. Ryzhov, and B. Thengvall. Simulation optimization: A panel on the state of the art in research and practice. In A. Tolk, D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 3696–3706, Piscataway NJ, 2014. IEEE. 16

[36] A. Gal, A. Mandelbaum, F. Schnitzler, A.rik Senderovich, and M. Weidlich. Traveling time prediction in scheduled transportation with journey segments. *Information Systems*, 64:266–280, 2017. 10

[37] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227, 2002. 9

[38] A. Goldenshluger. Nonparametric estimation of the service time distribution in the M/G/∞ queue. *Advances in Applied Probability*, 48(4):1117–1138, 2016. 10

[39] M. M. Günal and M. Pidd. Discrete event simulation for performance modelling in health care: a review of the literature. *Journal of Simulation*, 4(1):42–51, 2010. 12

[40] D. Gupta. Queueing models for healthcare operations. In *Handbook of Healthcare Operations Management*, pages 19–44. Springer, 2013. 9

[41] I. Gurvich and J. A. Van Mieghem. Collaboration and multitasking in networks: Architectures, bottlenecks, and capacity. *Manufacturing & Service Operations Management*, 17(1):16–33, 2014. 10

[42] J. M. Harrison. Stochastic networks and activity analysis. *AMERICAN MATHEMATICAL SOCIETY TRANSLATIONS*, pages 53–76, 2002. 8

[43] J. M. Harrison. A broader view of brownian networks. *The Annals of Applied Probability*, 13(3):1119–1150, 2003. 8

[44] J. M. Harrison. Brownian models of open processing networks: Canonical representation of workload. *The Annals of Applied Probability*, 13(1):390–393, 2003. 8

[45] L. He, K. Chen, W. Xu, J. Zhou, and F. Wang. Boosted sparse and low-rank tensor regression. In *Advances in Neural Information Processing Systems*, pages 1017–1026, 2018. 16

[46] S. G. Henderson, E. O'Mahony, and D. B. Shmoys. (citi)bike sharing. Submitted, 2016. 16

[47] S. G. Henderson and R. Pasupathy. Simulation optimization library, 2017. `http://www.simopt.org` Accessed April 26, 2017. 16

[48] C. Hong, A. Harrison, J. M.and Mandelbaum, A. Van Ackere, and L. M. Wein. Empirical evaluation of a queueing network model for semiconductor wafer fabrication. *Operations Research*, 36(2):202–215, 1988. 12

[49] W. J. Hopp and W. S. Lovejoy. *Hospital operations: Principles of high efficiency health care*. FT Press, 2012. 9

[50] T-C Horng, N. J. Dingle, A. Jackson, and W. J. Knottenbelt. Towards the automated inference of queueing network models from high-precision location tracking data. *ECMS*, 9:664–674, 2009. 12

[51] J. Huang, B. Carmeli, and A. Mandelbaum. Control of patient flow in emergency departments, or multiclass queues with deadlines and feedback. *Operations Research*, 63(4):892–908, 2015. 10

[52] R. Ibrahim, P. Regnard, N.and L'Ecuyer, and H. Shen. On the modeling and forecasting of call center arrivals. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2012. 11

[53] N. Jian, D. Freund, H. Wiberg, and S. G. Henderson. Simulation optimization for a large-scale bike-sharing system. In T. M. K. Roeder, P. I. Frazier, R. Szechtman, and E. Zhou, editors, *Proceedings of the 2016 Winter Simulation Conference*, pages 602–613, Piscataway NJ, 2016. IEEE. 16

[54] N. Jian and S. G. Henderson. An introduction to simulation optimization. In L. Yilmaz, W. K. V. Chan, T. M. K. Roeder, C. Macal, and M.D. Rosetti, editors, *Proceedings of the 2015 Winter Simulation Conference*, pages 1780–1794, Piscataway NJ, 2015. IEEE. 16

[55] N. Jian, S. G. Henderson, and S. R. Hunter. Sequential detection of convexity from noisy function evaluations. In A. Tolk, D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 3892–3903, Piscataway NJ, 2014. IEEE. 16

[56] T. Joachims, A. Swaminathan, and M. de Rijke. Deep learning with logged bandit feedback. In *International Conference on Learning Representations (ICLR)*, 2018. 14, 16

[57] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *ACM Conference on Web Search and Data Mining (WSDM)*, pages 781–789, 2017. 13, 14, 16

[58] S. Kim, R. Pasupathy, and S. G. Henderson. A guide to sample average approximation. In Michael C. Fu, editor, *Handbook of Simulation Optimization*, volume 216 of *International Series in Operations Research & Management Science*, chapter 8, pages 207–244. Springer New York, 2015. 16

[59] G. T. Lakshmanan, S. Rozsnyai, and F. Wang. Investigating clinical care pathways correlated with outcomes. In *Business process management*, pages 323–338. Springer, 2013. 7

[60] A. M. Law. *Simulation Modeling and Analysis*. McGraw Hill, New York, 5th edition, 2013. 13

[61] D. Lefortier, A. Swaminathan, Xiaotao Gu, T. Joachims, and M. de Rijke. Large-scale validation of counterfactual learning methods: A test-bed. In *NIPS 2016 What-If Workshop*, 2016. 16

[62] R. D. MacDonald, M. Ahghari, L. Walker, T. A. Carnes, S. G. Henderson, and D. B. Shmoys. A novel application to optimize aircraft utilization for non-urgent air transfers. *Air Medical Journal*, 33(1):34–39, 2014. 16

[63] A. Mandelbaum and S. Zeltyn. Estimating characteristics of queueing networks using transactional data. *Queueing systems*, 29(1):75–127, 1998. 10, 12

[64] Y. Marmor. *Emergency-Departments Simulation in Support of Service-Engineering: Staffing, Design, and Real-Time Tracking*. PhD thesis, Technion, 2010. 3

[65] D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson. Forecasting emergency medical service call arrival rates. *Annals of Applied Statistics*, 5(2B):1379–1406, 2011. 11

[66] M. S. Maxwell, E. Cao Ni, C. Tong, S. R. Hunter, S. G. Henderson, and H. Topaloglu. A bound on the performance of an optimal ambulance redeployment policy. *Operations Research*, 62(5):1014–1027, 2014. 16

[67] M. S. Maxwell, S. G. Henderson, and H. Topaloglu. Tuning approximate dynamic programming policies for ambulance redeployment via direct search. *Stochastic Systems*, 3(2):322–361, 2013. 16

[68] MegsRadio Internet Radio Station. `http://megsradio.fm`. 16

[69] X. Min, B. Yu, and F. Wang. Predictive modeling of the hospital readmission risk from patients? claims data using machine learning: A case study on copd. *Scientific reports*, 9(1):2362, 2019. 16

[70] S. Pallone, P. I. Frazier, and S. G. Henderson. Multisection: Parallelized bisection. In A. Tolk, D. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, editors, *Proceedings of the 2014 Winter Simulation Conference*, pages 3773–3784, Piscataway NJ, 2014. IEEE. 16

[71] S. Pallone, P. I. Frazier, and S. G. Henderson. Coupled bisection for root ordering. *Operations Research Letters*, 44(2):165–169, 2016. 16

[72] M. I. Reiman. Open queueing networks in heavy traffic. *Mathematics of operations research*, 9(3):441–458, 1984. 10

[73] M. Restrepo, S. G. Henderson, and H. Topaloglu. Erlang loss models for the static deployment of ambulances. *Health Care Management Science*, 12(1):67–79, 2009. 16

[74] S. Saghafian, W. J. Hopp, M. P. Van Oyen, J. S. Desmond, and S. L. Kronick. Patient streaming as a mechanism for improving responsiveness in emergency departments. *Operations Research*, 60(5):1080–1097, 2012. 9

[75] S. Schönig, R. Jasinski, L. Ackermann, and S. Jablonski. Deep learning process prediction with discrete and continuous data features. In *Proceedings of the 13th International Conference on Evaluation of Novel Approaches to Software Engineering - Volume 1: ENASE,*, pages 314–319. INSTICC, SciTePress, 2018. 9

[76] A. Senderovich, A. Rogge-Solti, A. Gal, J. Mendling, and A. Mandelbaum. The road from sensor data to process instances via interaction mining. In *International Conference on Advanced Information Systems Engineering*, pages 257–273. Springer, 2016. 7, 12

[77] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017. 13

[78] D. Singhvi, S. Singhvi, P. I. Frazier, S. G. Henderson, E. O' Mahony, D. B. Shmoys, and D. B. Woodard. Predicting bike usage for new york city's bike sharing system. In *Association for the Advancement of Artificial Intelligence Proceedings*, 2015. 16

[79] D. Sinreich and Y. N. Marmor. A simple and intuitive simulation tool for analyzing emergency department operations. In *Proceedings of the 36th conference on Winter simulation*, pages 1994–2002. Winter Simulation Conference, 2004. 12

[80] S. G. Steckley, S. G. Henderson, D. Ruppert, R. Yang, D. W. Apley, and J. Staum. Estimating the density of a conditional expectation. *Electronic Journal of Statistics*, 10:736–760, 2016. 16

[81] C. Su, J. Tong, Y. Zhu, P. Cui, and F. Wang. Network embedding in biomedical data science. *Briefings in bioinformatics*, 2018. 16

[82] Y. Su, A. Agarwal, and T. Joachims. Learning from logged bandit feedback of multiple loggers. In *ICML Workshop on Machine Learning for Causal Inference, Counterfactual Prediction, and Autonomous Action (CausalML)*, 2018. 16

[83] M. Sun, F. Tang, J. Yi, F. Wang, and J. Zhou. Identify susceptible locations in medical records via adversarial attacks on deep predictive models. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 793–801, 2018. 16

[84] C. Sutton and M. I. Jordan. Bayesian inference for queueing networks and modeling of internet services. *The Annals of Applied Statistics*, pages 254–282, 2011. 10

[85] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 1998. 14

[86] A. Swaminathan. *Counterfactual Evaluation and Learning from Logged User Feedback*. PhD thesis, Cornell University, 2017. 16

[87] A. Swaminathan and T. Joachims. Batch learning from logged bandit feedback through counterfactual risk minimization. *Journal of Machine Learning Research (JMLR)*, 16:1731–1755, Sep 2015. Special Issue in Memory of Alexey Chervonenkis. 13, 14

[88] A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems (NIPS)*, 2015. 13

[89] P. Frazier T. Joachims T. Schnabel, A. Swaminathan. Unbiased comparative evaluation of ranking functions. In *ACM International Conference on the Theory of Information Retrieval (ICTIR)*, pages 109–118, 2016. 16

[90] F. Tang, C. Xiao, F. Wang, and J. Zhou. Predictive modeling in urgent care: a comparative study of machine learning approaches. *JAMIA Open*, 2018. 16

[91] R. Waeber, P. I. Frazier, and S. G. Henderson. A framework for selecting a selection procedure. *ACM Transactions on Modeling and Computer Simulation*, 22(4):Article 16, 2012. 16

[92] R. Waeber, P. I. Frazier, and S. G. Henderson. Bisection search with noisy responses. *SIAM Journal on Control and Optimization*, 51(3):2261–2279, 2013. 16

[93] B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson. Travel time estimation for ambulances using bayesian data augmentation. *Ann. Appl. Statist.*, 7(2):1139–1161, 2013. 11

[94] B. S. Westgate, D. B. Woodard, D. S. Matteson, and S. G. Henderson. Large-network travel time distribution estimation for ambulances. *European Journal of Operational Research*, (1):322–333, 2016. 11

[95] W. Whitt. The queueing network analyzer. *Bell System Technical Journal*, 62(9):2779–2815, 1983. 10

[96] W. Whitt and X. Zhang. Forecasting arrivals and occupancy levels in an emergency department. *Operations Research for Health Care*, 2019. 11

[97] R. J. Williams. Diffusion approximations for open multiclass queueing networks: sufficient conditions involving state space collapse. *Queueing systems*, 30(1-2):27–88, 1998. 10

[98] G. B. Yom-Tov. *Queues in Hospitals: Queueing Networks with ReEntering Customers in the QED Regime*. PhD thesis, Technion, 2010. 3

[99] G. B. Yom-Tov and A. Mandelbaum. Erlang-r: A time-varying queue with reentrant customers, in support of healthcare staffing. *Manufacturing & Service Operations Management*, 16(2):283–299, 2014. 9

[100] S. Zeltyn, Y. N. Marmor, A. Mandelbaum, B. Carmeli, O. Greenshpan, Y. Mesika, S. Wasserkrug, P. Vortman, A. Shtub, and T. Lauterman. Simulation-based models of emergency departments:: Operational, tactical, and strategic staffing. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 21(4):24, 2011. 13

[101] X. Zhang, L. He, K. Chen, Y. Luo, J. Zhou, and F. Wang. Multi-view graph convolutional network and its applications on neuroimage analysis for parkinson's disease. *AMIA Annual Symposium*, 2018. 16

[102] Z. Zhou, D. S. Matteson, D. B. Woodard, S. G. Henderson, and A. C. Micheas. A spatio-temporal point process model for ambulance demand. *Journal of the American Statistical Association – Applications and Case Studies*, 110(509):6–15, 2015. 11