Data-driven Modelling and Optimization

Runyu Tang

Feburary 22, 2023

About the course

- This is my first time to teach this course. Feedbacks are extremely appreciated!
- The course syllabus is tentative, and we may need to adapt depending on time constraints.
- This course is just an introduction. Although it covers several interesting topics, you may dig more materials for researches or business applications.
- Some topics are very recent and we may not find one textbook to cover these materials. However, I'll provide some references at the end of each section.

What is data-driven optimization?

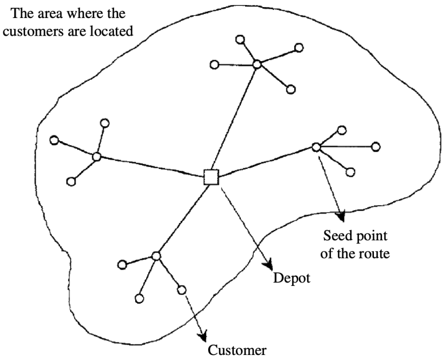
- Problem-driven research: the starting point is a problem identified by an academic or an industry professional, and the challenge is to find answers by developing new theory or new insights.
- Data-driven research: the starting point is a large amount of data, and the challenge is to find answers by analyzing the data and developing new models.

Simchi-Levi, David. OM forum—OM research: From problem-driven to data-driven research. Manufacturing & Service Operations Management 16.1 (2014): 2-10.

From theory to practice

- Theory: Location based heuristic to solve the capacitated vehicle routing problem.
- Practice: New York City school system, About 100,000 students rode around 1,150 buses.
- Used substantially fewer buses than what the city used.

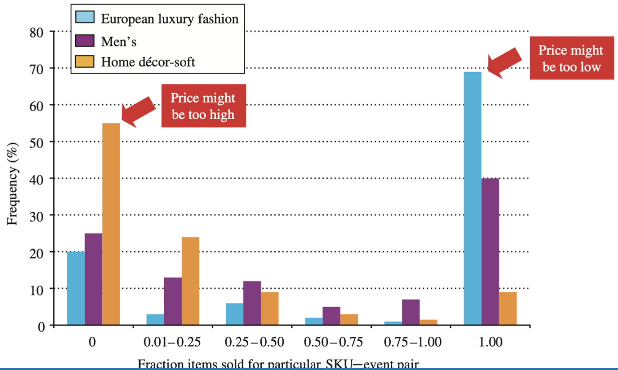
Figure 2 An Example of the Location-Based Heuristic



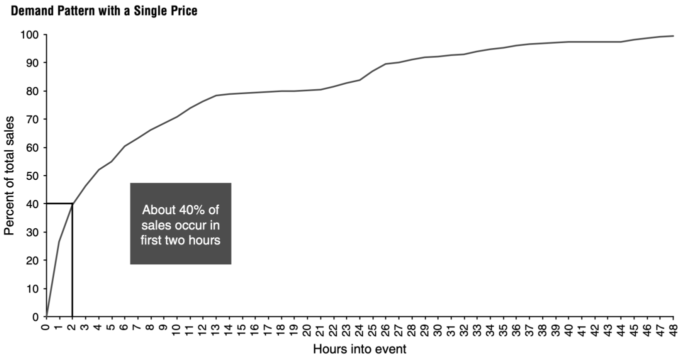
Tour used to contruct heuristic

An Example

- An retailer offers new products via online 48-hour sales events.
- The retailer has very little knowledge of customer demand as a function of price.
- It offers a single price during 48 hours.



What is data-driven optimization?



This suggests that once an event starts, the retailer faces an exploration–exploitation trade-off between learning to gather more information about the demand function versus optimizing price to maximize revenue.

The goal was to improve revenue, but it was not clear how to achieve that objective.
It was the **data** that suggested that the retailer had an opportunity to learn about consumer behavior in the first few minutes of an event and then optimize pricing decisions.

- Developed a learning and optimization method with very few price switches.
- Expected revenue increased by 6%-7%.
- Merging theory and practice.
- It started with data from our retailer and their interest in using data to improve their pricing process.
- If the retailer changes the price k times during the learning period, expected loss is proportional to $O(\log\log\cdots\log n)$.

The traditional supply chain management will hit a wall sooner or later.

— David Simchi Levi.

- Organizations are changing the way they capture data, analyze information, and make decisions.
- We analyze systems that involve people. Such systems are difficult to understand unless we have access to behavioral data.
- Brought together statistics, machine learning, and OR techniques to not only deliver an effective pricing optimization tool, but also identify opportunities and develop a new mechanism.

Two types of data-driven researches:

- The focus is on a specific goal—increase revenue, decrease cost, reduce the spread of an epidemic—and the challenge is to **let the data identify** the specific issues, opportunities, and models that the organization should focus on.
- The second is an open-ended search for correlations and relationships without any clear goal in mind. This is typically the objective of **data mining**: uncover economic or other relationships by analyzing a huge mass of data.

Simchi-Levi, David. OM forum—OM research: From problem-driven to data-driven research. *Manufacturing & Service Operations Management* 16.1 (2014): 2-10.

Historically, research in operations management (OM) has focused on models... Today, there has been a shift in research in OM. This shift has been driven primarily by the increasing availability of data.

Data analytics (or analytics for short): using data to create models leading to decisions that create value.

- Supply Chain Management
  - Location and Omnichannel operations
  - Inventory Management
- Revenue Management
  - Choice Model and Assortment
  - Pricing and Promotion Planning
  - Personalized Revenue Management
- Healthcare operations

Mišić, V. V., Perakis, G. (2020). Data analytics in operations management: A review. *Manufacturing & Service Operations Management*, 22(1), 158-169

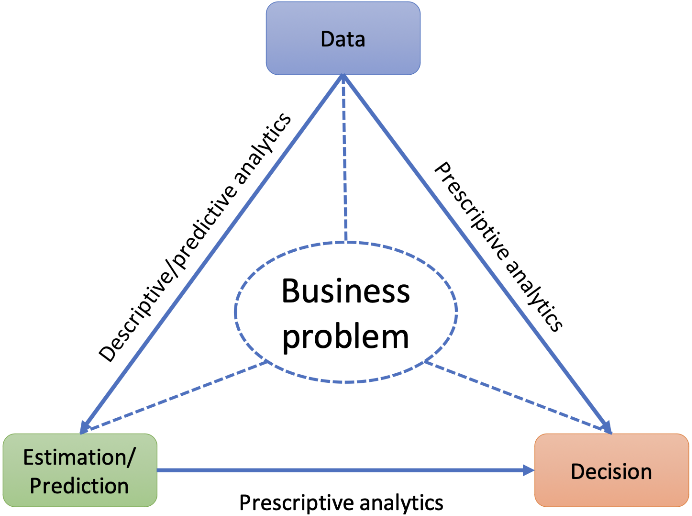
Analysis v.s. Analytics:

- Data analysis: the analysis restructures existing available information or data $\rightarrow$ the past trends.
- Data analytics: the analysis uses data to create models leading to decisions that create value $\rightarrow$ the future trends.
  The application of scientific & mathematical methods to the study & analysis of problems involving complex systems – Definition by INFORMS

Descriptive v.s. Predictive v.s. Prescriptive:

- Descriptive analytics: describes what is happening.
- Predictive analytics: forecasts potential future outcomes.
- Prescriptive analytics: helps you draw specific recommendations.

`https://www.informs.org/Explore/Operations-Research-Analytics`


Data
Descriptive/predictive analytics
Prescriptive analytics
Business problem
Estimation/ Prediction
Decision
Prescriptive analytics

"Motivated by the delivery operations and data of a food delivery service provider (Ele.me), we discuss a framework that integrates travel-time predictors with order-assignment optimization."

The order-assignment problem consists of two stages:

- Upon receiving the orders from customer locations $\mathcal{I}$, the service provider first assigns the locations $\mathcal{I}_k$ to driver k among a pool of drivers $\mathcal{K}$,
- Each driver k then chooses a route to visit the locations $\mathcal{I}_k$.

Sheng Liu, Long He, Zuo-Jun Max Shen (2020) On-Time Last-Mile Delivery: Order Assignment with Travel-Time Predictors. (previously Data-Driven Order Assignment for Last Mile Delivery) Management Science 67(7):4095-4119.

Data set:

The dataset records the following information:

- (1) the order time: the time when the order is received by the provider;
- (2) the quantity: the number of items ordered;
- (3) the time window: the guaranteed delivery time;
- (4) the pickup time: the time when the order is collected by a driver;
- (5) the delivery time: the actual time when the order is delivered to the customer;
- (6) the longitude and latitude: the customer location;
- (7) the cutoff time: the provider has determined a sequence of cutoff times $\{t_1, t_2, ...\}$, and all orders placed in $(t_n, t_{n+1}]$ are guaranteed to be delivered by $t_{n+1} + 75$ minutes.

Figure 1. (Color online) Historical Demand Density Heat Map (Circle in the Center Represents the Depot)

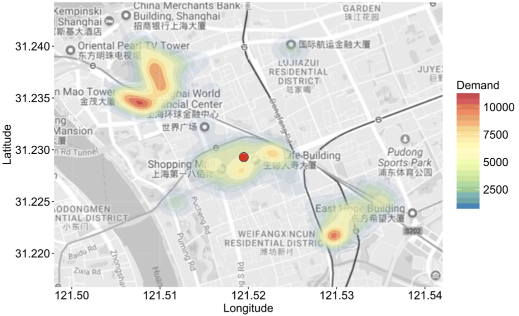
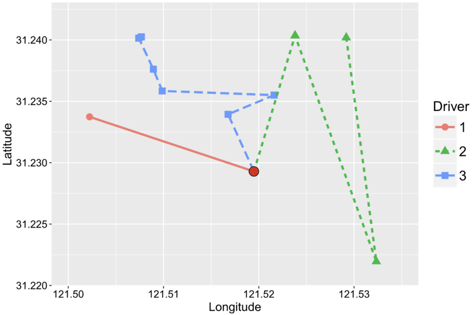


Figure 2. (Color online) Delivery Routes by Three Drivers (Circle in the Center Represents the Depot)



Two practical concerns in the order-assignment problem:

- The driver's routing behavior
  — The drivers do not always follow the planned delivery sequences from the routing tools
  — VRP, sVRP(stochastic), DOA(data-driven order assignment)
- Uncertain service time
  — The time a driver spends at a customer location is highly uncertain
  — SAA(sample average approximation) and DRO(distributionally robust optimization)

On-Time Last-Mile Delivery

Figure 9. Out-of-Sample Performance (Improvement over the VRP-SAA) of the DOA and VRP Models

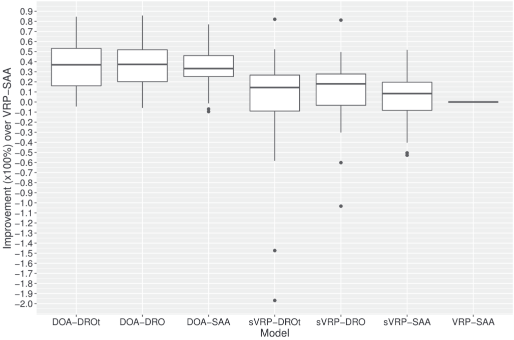
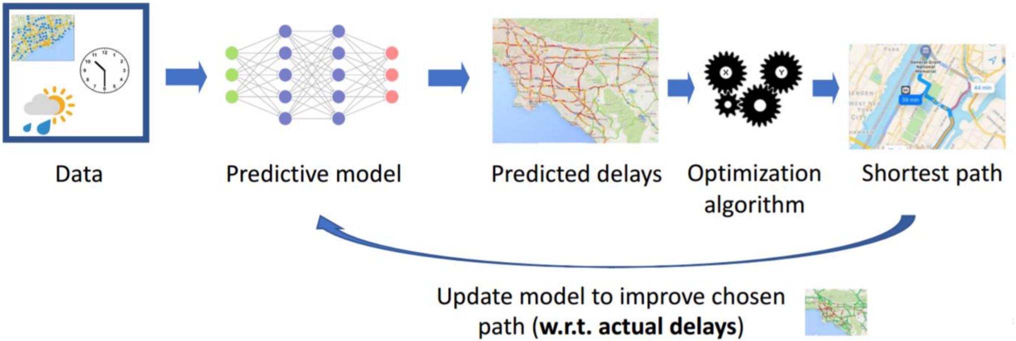


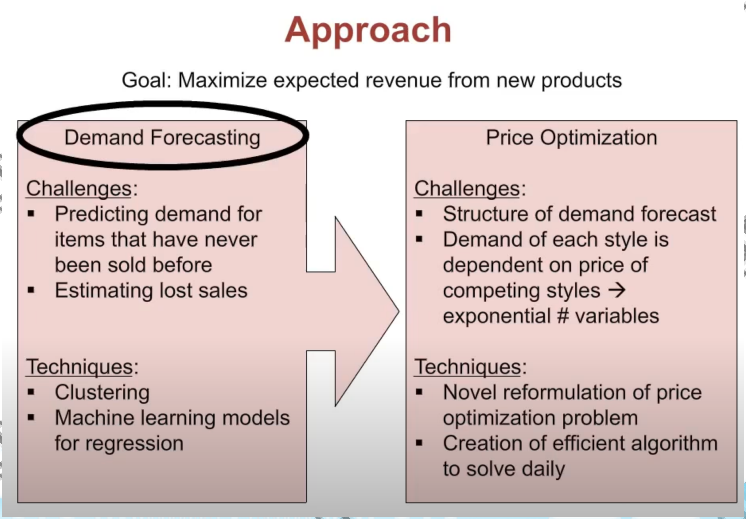
Table 6. Average Improvement of the DOA Models and Scaled VRP Models over the VRP-SAA

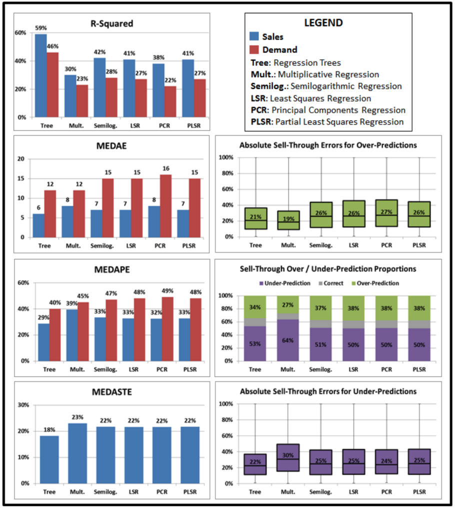
	DOA-DROt	DOA-DRO	DOA-SAA	sVRP-DROt	sVRP-DRO	sVRP-SAA
Average improvement over VRP-SAA, %	**36.3**	36.0	33.8	3.11	10.3	6.32


Data
Predictive model
Predicted delays
Optimization algorithm
Shortest path
Update model to improve chosen path (**w.r.t. actual delays**)

Data Analytics: Example

The data analytics process:





- Across all performance metrics evaluated, regression trees with bagging consistently outperformed the other regression models.
- Construct into an inter programming problem, andd design efficient algorithms.

Table 7 **Estimate of Percent Increase in Revenue Due to Raising Prices**

Category	Estimate of percent increase in revenue (%)	Estimate of 90% confidence interval (%)	Estimate of 95% confidence interval (%)
A	−3.4	[−11.5, 7.7]	[−13.5, 10.1]
B	11.4	[3.9, 19.2]	[1.1, 21.0]
C	12.5	[1.1, 23.4]	[−2.0, 26.6]
D	13.7	[3.4, 22.8]	[0.0, 25.2]
E	23.8	[5.4, 47.6]	[0.0, 56.7]
Overall	9.7	[2.3, 17.8]	[0.0, 20.2]

Departments in the major OR/MS journals:

- OR: Machine Learning and Data Science
  `https://pubsonline.informs.org/page/opre/editorial-statement/area-editors-statements`
- MS: Data Science (formerly Big Data Analytics)
  `https://pubsonline.informs.org/page/mnsc/editorial-statement`
- MSOM: Analytics in Operations
  `https://pubsonline.informs.org/page/msom/editorial-statement`
- POM: Revenue Management and Market Analytics
  `https://www.poms.org/journal/departments/`
- IJOC: Data Science & Machine Learning
  `https://pubsonline.informs.org/page/ijoc/editorial-statement`

Data Science (formerly Big Data Analytics)

Increased computational power and the explosion of data are rapidly changing the way and the extent to which organizations capture data, build models, and make decisions. A number of business decision problems impacted by this rapid change are germane to research in management science. The data science (DS) department (formerly, big data analytics) solicits research that advances our ability to solve complex business decision problems by learning from large datasets and complex environments.

Areas of interest include problems of dynamic optimization (e.g., how might we exploit structure in high dimensional exploration?); data-driven decision making (e.g., how might we make optimal decisions in the face of high-dimensional contextual data?); inference (e.g., how do we draw causal inferences from rich observational data? How might we design experiments on commerce platforms?); the interface with optimization (e.g., how can machine learning techniques benefit traditional online algorithms? How can integer programming techniques certify the robustness of deployed DS models?); and fairness and equity (e.g., how can bandit models increase efficiency and equity in hiring?). In all cases, we care particularly that the problem studied is soundly motivated by a relevant business or application context. This could range from problems in transportation, to healthcare, to social science contexts, and beyond. The department will be welcoming of any broad impactful application area. In exceptional research, the connection of the research to the motivating application context will be evident from an empirical study with data from the motivating context.

Recognizing the pace of research in the broader DS community, we particularly welcome submissions for which preliminary abridged versions of the research appeared recently in selective archival conferences (such as NeurIPS, ICML, COLT, and ICLR). Authors submitting such papers can submit, at their own discretion, all (anonymized) reviewer feedback from such conference submissions.

Common techniques in data-driven optimization:

- Machine learning
- (Data-driven Distributionally) Robust optimization
- Smart predict then optimize (featured data)
- Demand learning (multi-armed bandit)
- ...

- In total 32 hours
- I will teach the first 17 hours
- The other 15 hours will be taught by **Prof. Zhang Huili**

Schedule

Time	Content
Day 1	Introduction
Day 1, 2	Common machine learning methods
Day 3, 4	Data-driven robust optimization
Day 5	Data-driven with featured data
Day 6	Presentation

We want to introduce you to the common techniques and their applications in Data-driven optimization, such that

- **Get interested** in the increasingly popular field of data-driven optimization.
- You can **read the related papers and understand the key ideas** on your own.
- **Handle some problems** adopting the methods introduced in the course.
- Maybe, in the future, **do some researches** in this field.

Where to find data?

- Kaggle, `https://www.kaggle.com/`
- M&SOM data driven research challenge,
  - 2018: Cainiao data
  - 2020: JD transaction level data
  - 2020: NetEase Cloud music data
  - 2021: RiRiShun Logistics data (Haier)
- Alibaba Tianchi, `https://tianchi.aliyun.com/competition/activeList`
- COAP `https://www.coap.online/competitions/1`
- Self-collected data...

- Simchi-Levi, David. OM forum—OM research: From problem-driven to data-driven research. *Manufacturing & Service Operations Management* 16.1 (2014): 2-10.
- Feng, Q., Shanthikumar, J. G. (2018). How research in production and operations management may evolve in the era of big data. *Production and Operations Management*, 27(9), 1670-1684.
- Mišić, V. V., Perakis, G. (2020). Data analytics in operations management: A review. *Manufacturing & Service Operations Management*, 22(1), 158-169
- M. Qi, H.Y. Mak and Z.J.M. Shen (2020). Data-Driven Research in Retail Operations - A Review. *Naval Research Logistics*, 67(8), 595-616.

- Kris Johnson Ferreira, Bin Hong Alex Lee, David Simchi-Levi (2015) Analytics for an Online Retailer: Demand Forecasting and Price Optimization. *Manufacturing & Service Operations Management* 18(1):69-88.
- Sheng Liu, Long He, Zuo-Jun Max Shen (2020) On-Time Last-Mile Delivery: Order Assignment with Travel-Time Predictors. *Management Science* 67(7):4095-4119.
- Guodong Lyu, Chung-Piaw Teo (2022) Last Mile Innovation: The Case of the Locker Alliance Network. *Manufacturing & Service Operations Management* 24(5):2425-2443.
- Siddharth Arora, James W. Talor, Ho-Yin Mak (2023). Probabilistic Forecasting of Patient Waiting Times in an Emergency Department. *Manufacturing & Service Operations Management*, Accepted for Publication.