

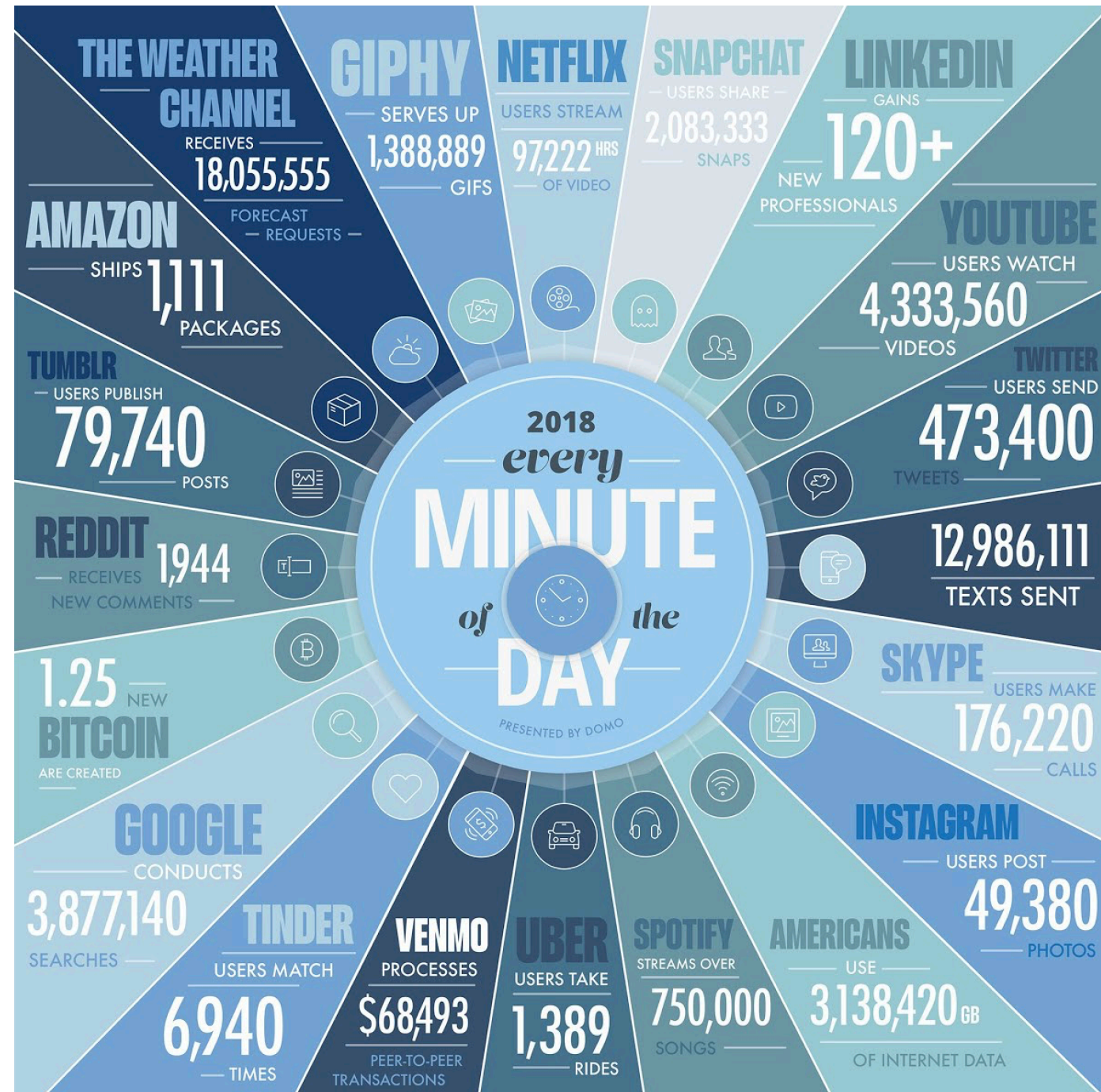
2020-1 Data Analysis with Applications

Lecture 01. Introduction to Analytics

Department of Industrial and Information Systems Engineering,
Soongsil University

Introduction

Data is Powerful and Everywhere



How much data is generated *every minute*?

Data is Powerful and Everywhere

- By 2020, it is estimated that for every person on earth, 1.7 MB of data will be created every second.
- New data is produced at an exponential rate.
- The ability to make **data-driven decision** is crucial to any business and industries.

Data and Analytics

- Survey of 1,300 companies worldwide found that over half **use data analytics in everyday decisions.**
- Estimated that there is a shortage of 140,000 - 190,000 people with deep analytical skills to fill the demand of jobs in the U.S. by 2018.
- IBM has invested over \$20 billion since 2005 to grow its analytics business.
- Companies will invest more and more on **analytics, hardware, software and services.**
- Critical in almost every industry
 - Healthcare, media, sports, finance, manufacturing, education, game, entertainmnet, e-commerce, government, etc.

What is Analytics?

- The science of using **data** to build **models** that lead to better **decisions** that add **value** to individuals, to companies, to institutions.

What is Analytics?

- **Descriptive Analytics**
 - Finds patterns in the data
 - Summary statistics, Visualizations, etc.
- **Predictive Analytics**
 - Making predictions about future outcomes based on historical data and analytics techniques.
 - Regression, Decision tree, SVM, Neural Net, etc.
- **Prescriptive Analytics**
 - Focuses on finding the best course of action in a scenario given the available data.
 - Optimization, Decision-making.

This Course

- In this course, we study a variety of analytics techniques through **real-world examples and real data** originated from various industries.
- Theoretical part of analytics techniques is kept to a minimum.
- Focus on **descriptive & predictive analytics**.
- The statistical software **R** is used.
- Goal is to make you comfortable using data analytics in your career and your life.

Course Outline

- Introduction
- R Programming Basics
- Data Visualization
- Predictive Analytics Process
- Linear Regression
- Logistic Regression
- Decision Tree
- Recommendation System
- Clustering
- Social Network Analytics
- Text Analytics ...

- Each topic will be composed of :
 - One lecture on the basics of methodology
 - One or two lectures on practice in R using real-world cases
 - Homework Assignment

- 구성요소
 - 출석 10점 : 결석 시 1점 감점, 지각 시 0.5점 감점
 - 과제 30점 : Topic 당 1회의 R 활용 과제
 - Predictive modeling competition 15점 : 추후 공지
 - 기말시험 45점
- 학점은 A+B 80%, C이하 20% 로 부여함.
- 부정행위 시 해당 항목 0점 처리함.

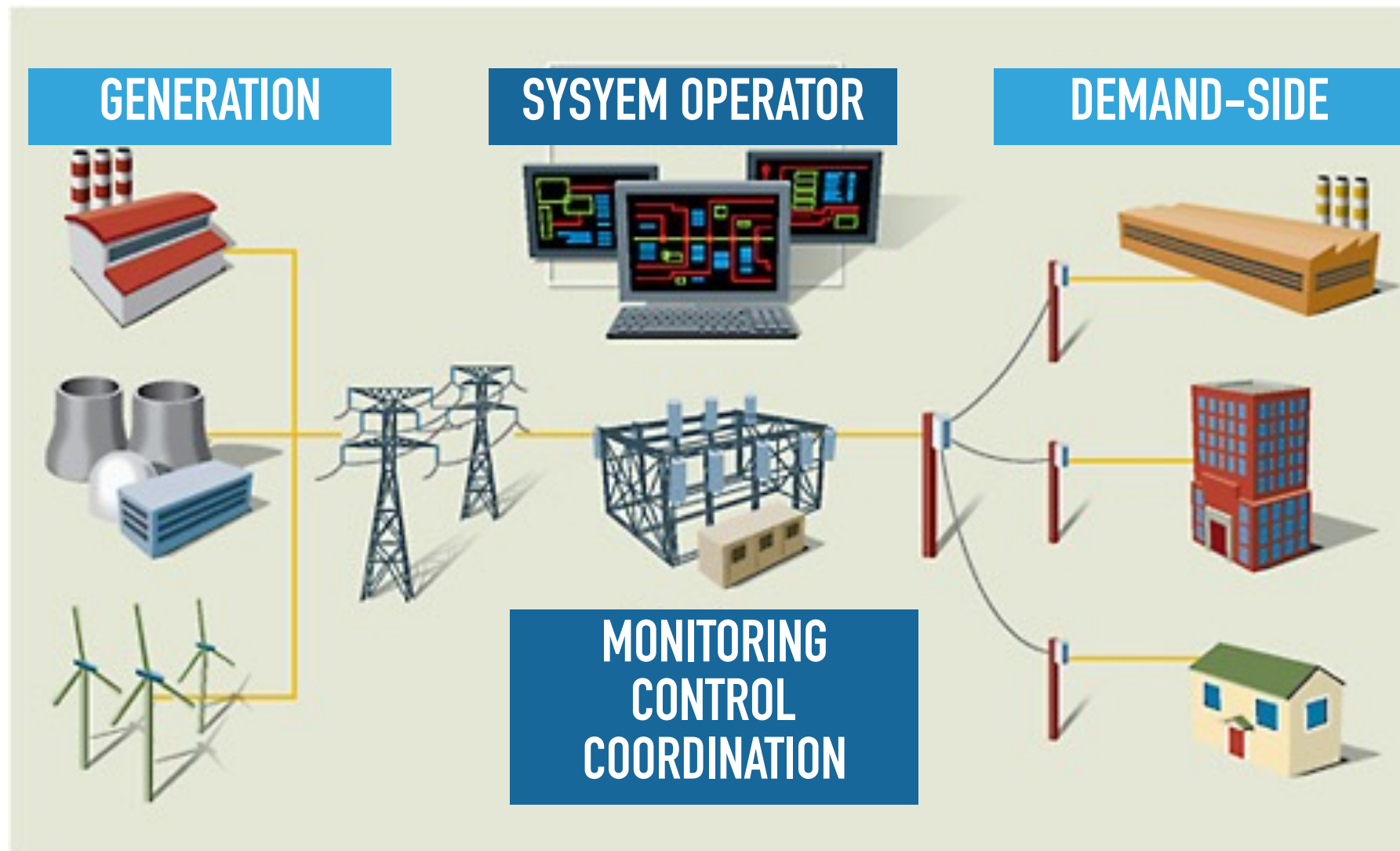
References

- *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R* by Galit Shmueli et al.
- *Machine Learning with R: Expert Techniques for Predictive Modeling* by Brett Lantz
- *The Analytics Edge* by Dimitris Bertsimas et al.
- *An Introduction to Statistical Learning with Applications in R* by Gareth James et al.
- *Hands-On Machine Learning with R* by Brad Boehmke
- *Data Visualization with R* by Rob Kabacoff
- and more..

Analytics in Power Generation Company

Descriptive, Predictive, Prescriptive Analytics

Power System Operation (전력시스템 운영)



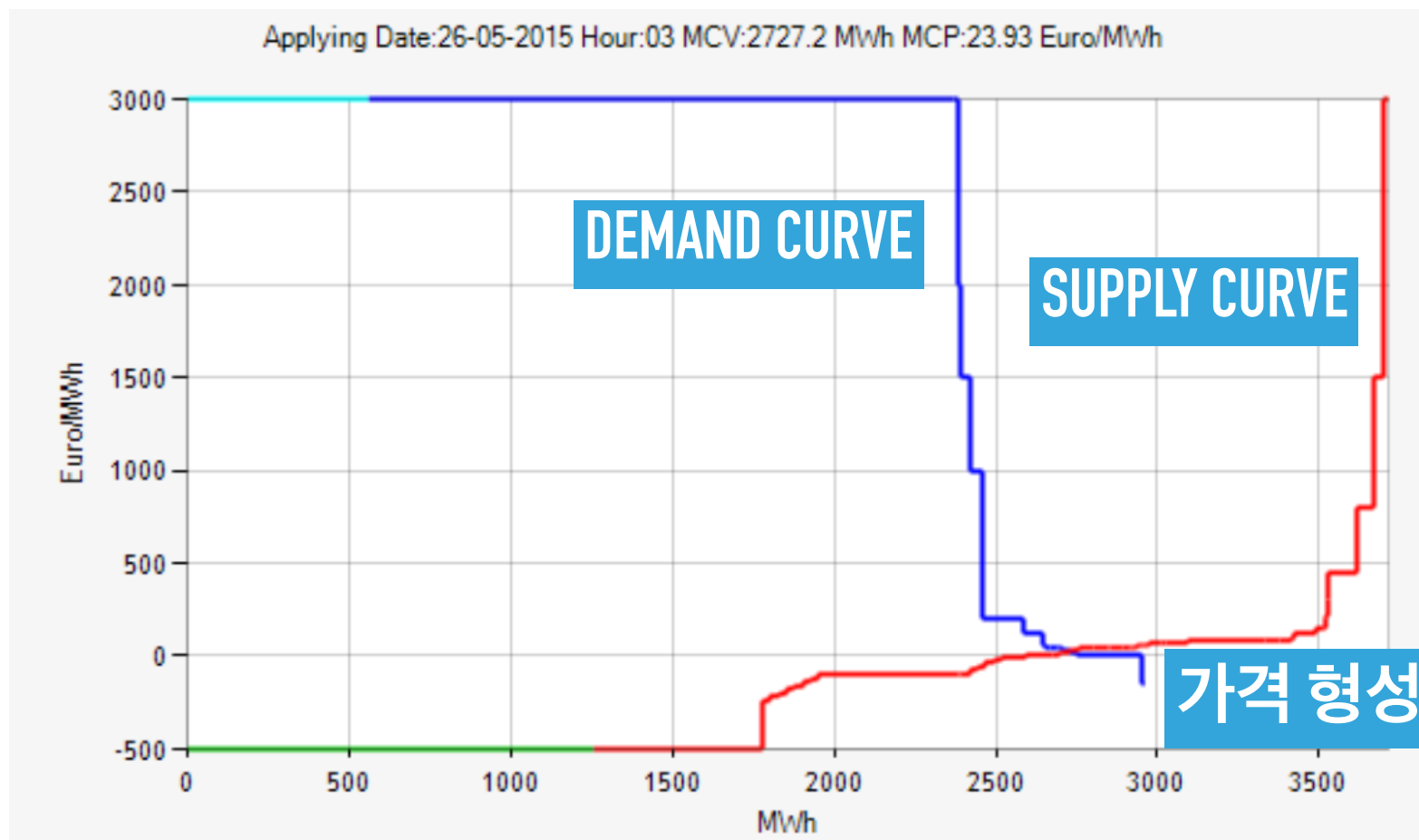
발전회사

한국전력
전력거래소

산업용, 가정용, 공공,..

Power Exchange (전력 거래소)

- 전력거래소에서 매일 매시간 전력의 가격과 생산량을 결정함.
- 발전회사들이 매일 매시간 발전량과 발전 비용을 하루 전에 입찰하면 전력거래소에서 예상되는 수요에 맞추어 가격과 발전회사별 생산량을 결정한다.

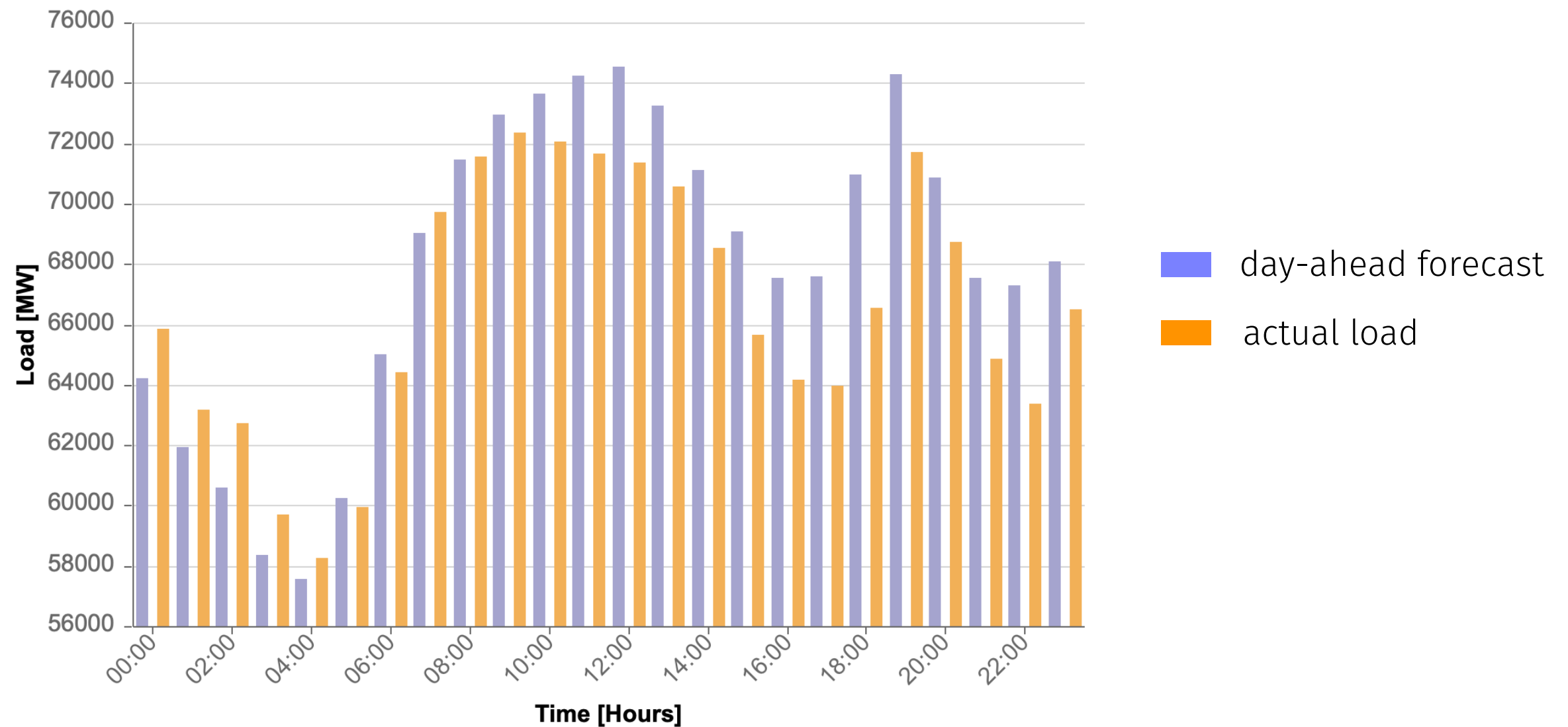


Generation Company (발전 회사)

- 미래의 전력 수요를 정확하게 예측하여야 수익을 최대화하는 발전량과 발전 비용을 입찰할 수 있다.
 - 전력가격이 높을 시간대에 발전기의 가동을 늘리고, 가격이 낮은 시간대에 가동을 줄이는 것이 전체 수익을 최대화하는 운영 전략이다.
 - 하지만 발전기의 유지보수 일정, 발전기의 효율, 연료의 도입시기 등의 제약을 고려하여 발전기의 가동 계획을 수립해야 한다.
- **Descriptive Analytics:** 과거 데이터 및 예측 결과 visualization
 - **Predictive Analytics:** 과거 데이터로 부터 전력가격을 예측하기 위한 통계적 모델링
 - **Prescriptive Analytics:** 예측된 전력가격을 기반으로 최적의 발전기 운영 스케줄 수립

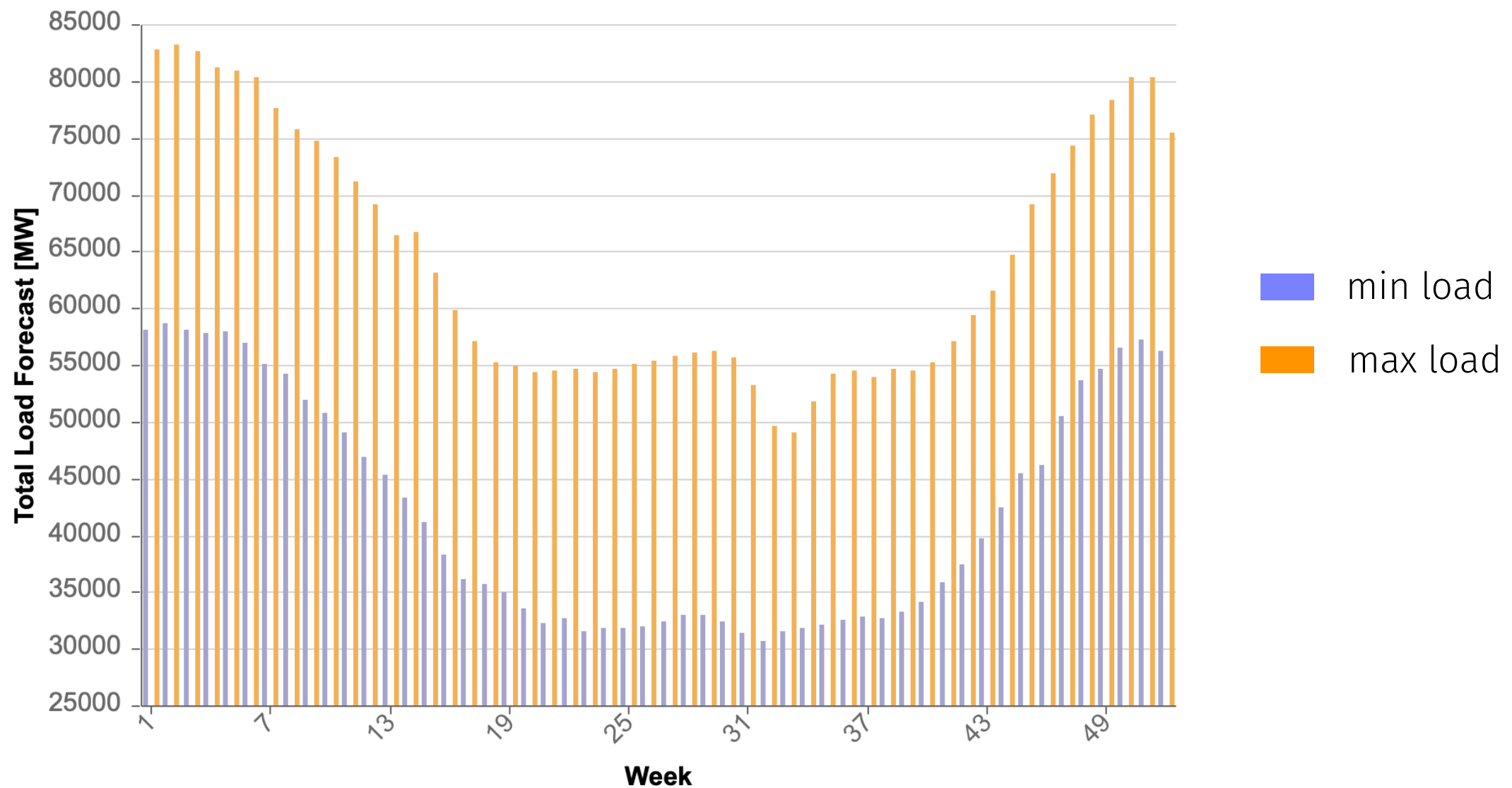
Descriptive Analytics

- Day-ahead Load forecast vs. Actual Load (2020/02/27, France)



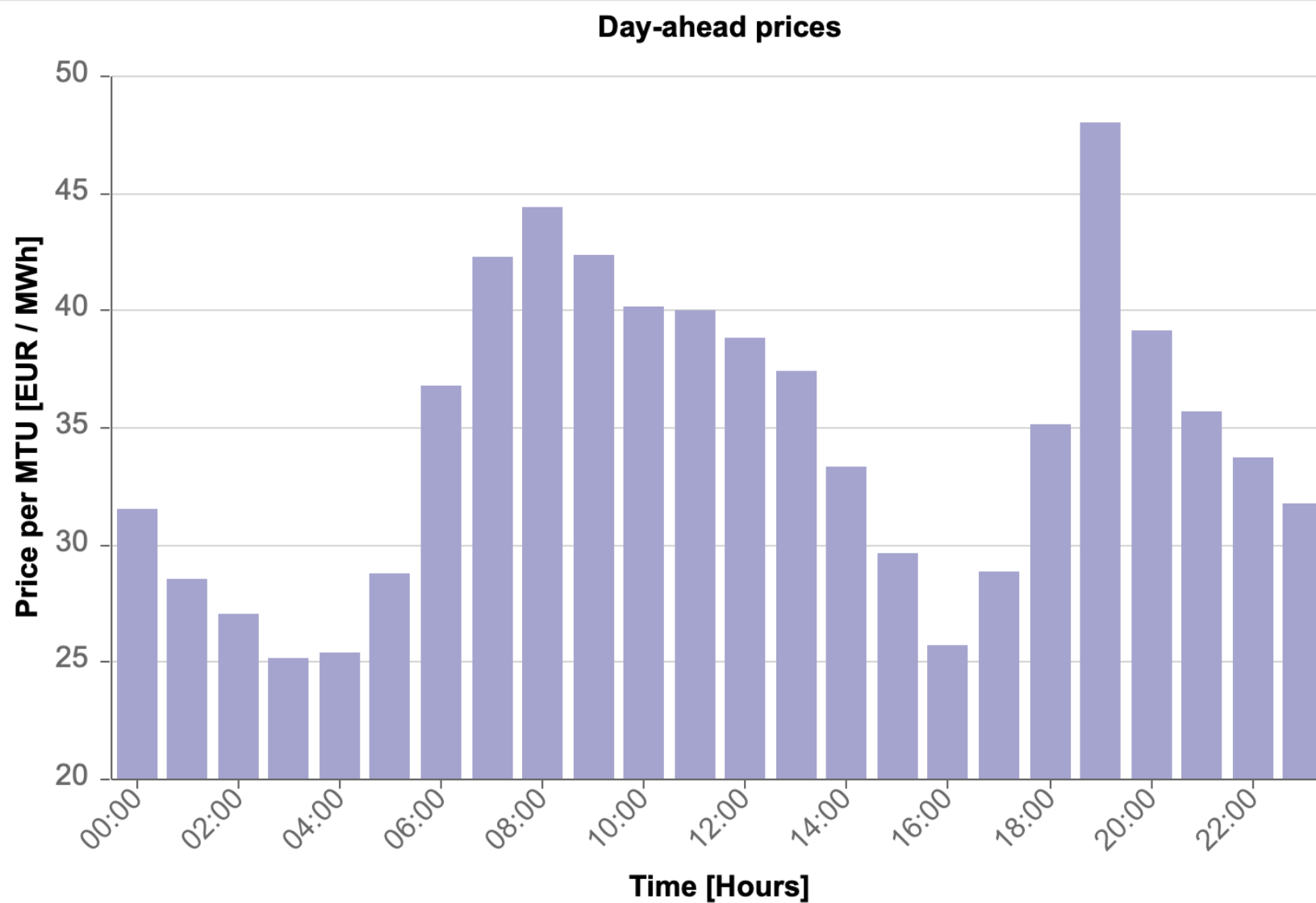
Descriptive Analytics

- Year-ahead weekly load forecast (2019, France)



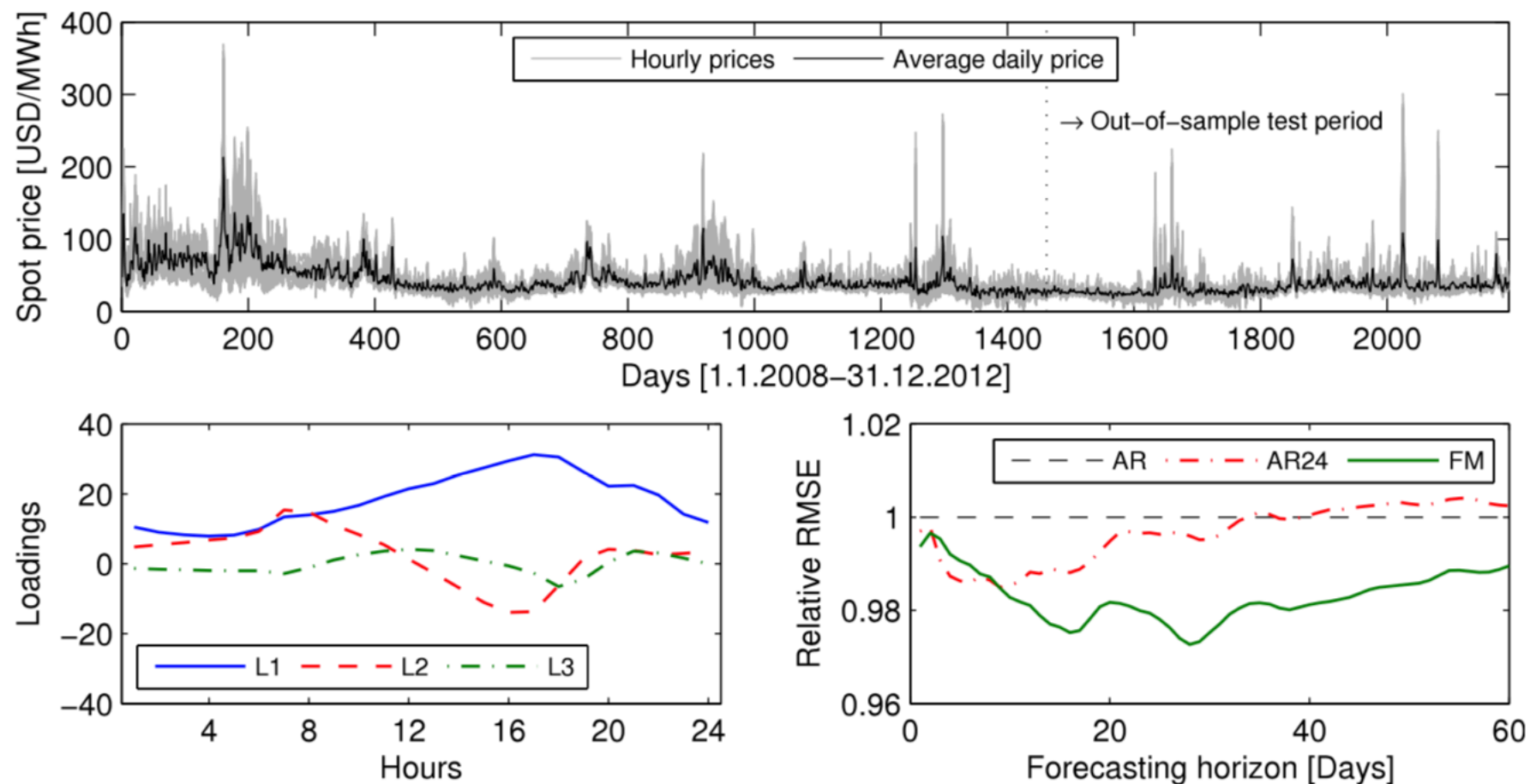
Descriptive Analytics

- Day-ahead prices (2020/02/27, France)



Predictive Analytics

- 전력가격 예측은 발전회사, 송배전 회사, 전력거래소 등의 에너지 관련 회사들의 의사결정에 필요한 가장 중요한 요소 중의 하나이다.
- 모델링 기법: regression, time-series modeling, neural network, simulation, etc.

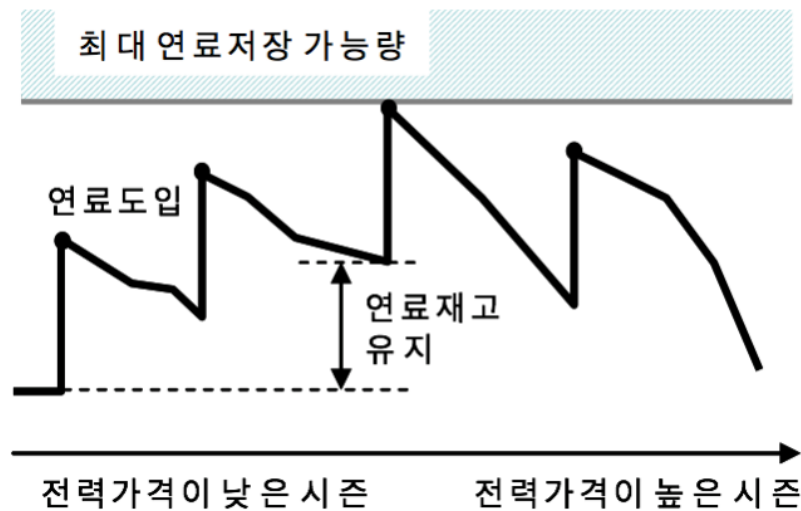


Prescriptive Analytics

- 전력 가격 예측, 연료 재고, 발전기 특성 등을 고려하여 최적의 발전기 가동 스케줄을 결정해야 한다.
- 선형계획법, 정수계획법, 시뮬레이션과 같은 OR 및 최적화 기법이 주로 활용된다.

DATE	1h	2h	3h	4h	5h	6h	7h	8h	9h	10h
20111101	125.14_	121.16_	120.86_	120.69_	120.88_	123.84_	127.39_	130.64_	145.20_	163.83_
20111102	125.14_	121.16_	120.86_	120.69_	120.88_	123.84_	127.39_	130.64_	150.98_	165.00_
20111103	123.85_	120.86_	120.32_	120.19_	120.69_	121.40_	125.14_	128.87_	138.91_	150.98_
20111104	121.40_	120.69_	120.00_	120.00_	120.19_	120.88_	124.13_	128.81_	138.91_	150.98_
20111105	123.85_	120.86_	120.00_	120.00_	120.00_	120.86_	121.40_	121.53_	127.39_	130.64_
20111106	120.69_	119.91_	115.58_	83.94_	115.58_	119.28_	120.00_	119.28_	118.29_	119.28_
20111107	119.91_	76.35_	69.96_	69.95_	72.99_	119.28_	120.19_	121.16_	130.64_	144.50_
20111108	121.40_	120.69_	120.19_	120.00_	120.19_	120.86_	124.13_	128.81_	144.50_	154.72_
20111109	121.40_	120.69_	120.19_	120.00_	120.19_	120.86_	124.13_	128.81_	144.50_	155.00_
20111110	121.40_	120.69_	120.19_	120.00_	120.19_	120.88_	124.13_	127.39_	144.50_	155.00_
20111111	121.40_	120.69_	120.19_	120.00_	120.19_	120.88_	124.13_	128.81_	144.50_	150.98_
20111112	123.84_	120.69_	120.00_	120.00_	120.00_	120.69_	121.40_	121.40_	124.13_	127.39_

[Electricity price forecast]



[Fuel inventory]

Maximize

$$\sum_{i \in U} \sum_{k \in M(i)} \sum_{t \in T} PROFIT_{ikt} \cdot x_{ikt} - COST_{OM}(x_{ikt}) \quad (2)$$

$$\sum_{k \in M(i)} x_{ikt} \leq 1 \quad (3)$$

$$u_t = \sum_{i \in U} \sum_{k \in M(i)} FU_{ikt} \cdot x_{ikt} \quad (4)$$

$$r_{t-1} + y_t = u_t + r_t \quad (5)$$

$$r_{t-1} + y_t \leq TANK^{max} \quad (6)$$

$$CARGO^{min} \cdot z_t \leq y_t \leq CARGO^{max} \cdot z_t \quad (7)$$

$$FUFL^{min} \leq \sum_{t \in T} y_t \leq FUEL^{max} \quad (8)$$

$$\sum_{k \in MEb(i) \cup Mlo(i)} x_{ikt} \leq \sum_{k \in MoE(i)} x_{ikt-1} \quad (9)$$

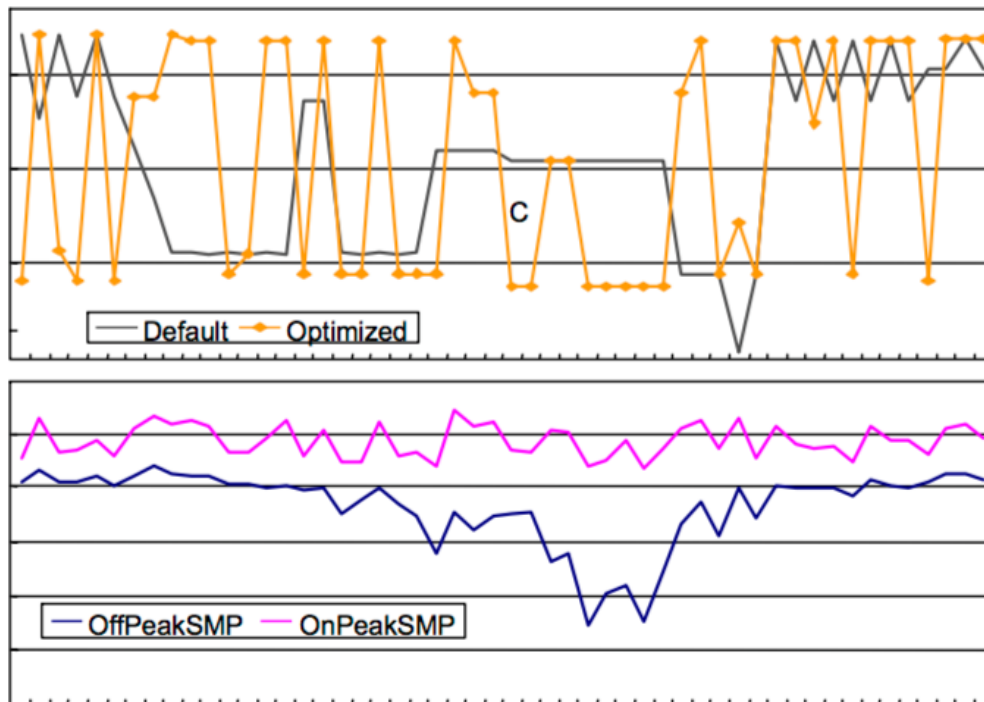
$$\sum_{k \in MAo(i) \cup Mlo(i)} x_{ikt} \leq \sum_{k \in MoA(i)} x_{ikt-1} \quad (10)$$

$$x_{ikt}, z_t \in B, u_t, y_t, r_t \in R^+$$

[Optimization model]

Prescriptive Analytics

- 최적화 모형으로부터 얻어지는 발전기 가동 스케줄과 기존 가동 스케줄의 성능을 비교한다.
- 일반적으로 Analytics의 도입으로 이익의 2~5% 증가 (or 비용의 2~5% 감소)를 달성할 수 있다면, 이것은 매우 큰 성과라고 할 수 있다.



[Optimized Schedule]

표 1. 비교 결과

	Default	Optimized	Ratio
연간매출	50,210,895	51,492,440	2.55%
연간이익	24,168,793	25,483,575	5.44%
잔존연료	0.0716%	0.0003%	0.0713%

[Performance comparison]