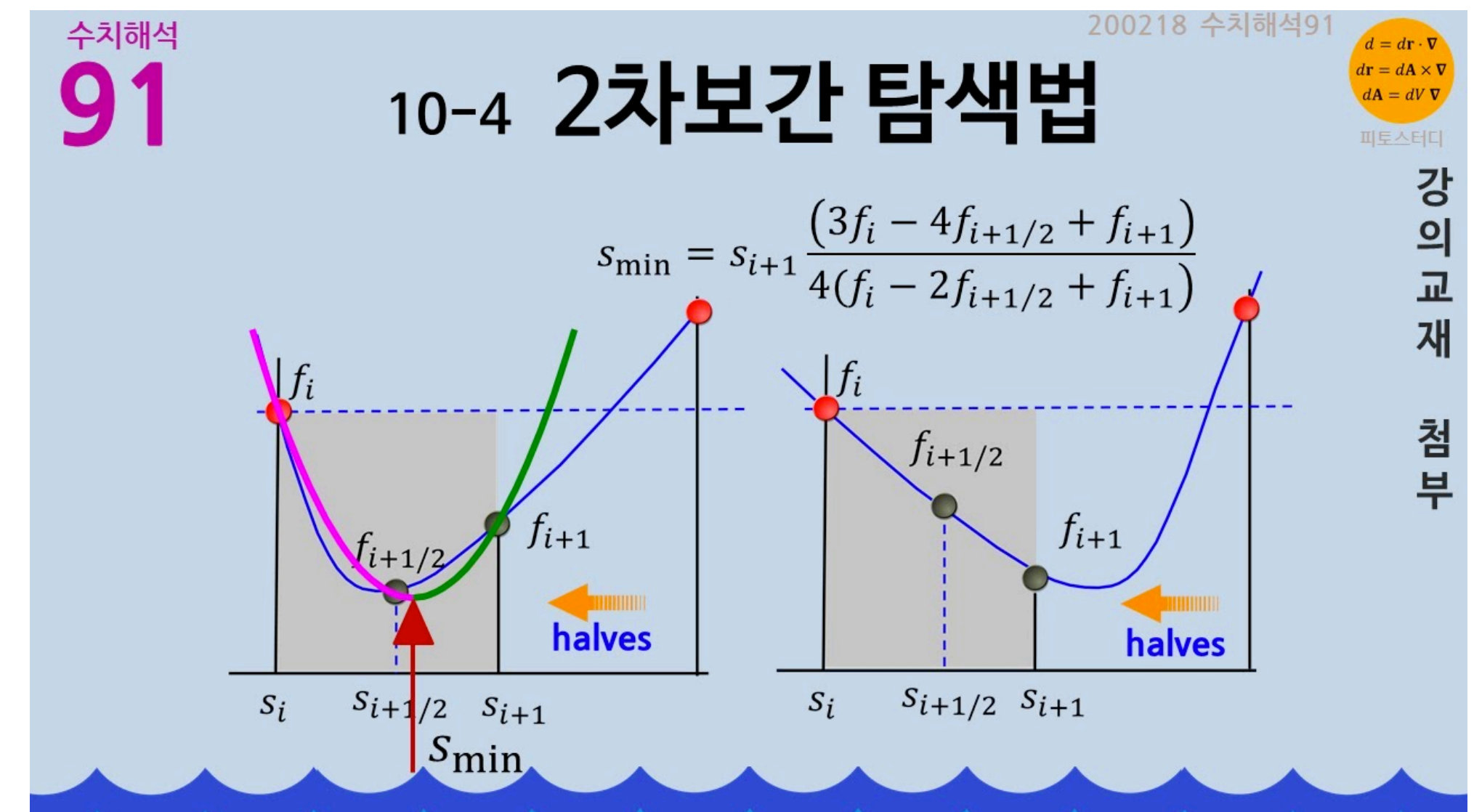
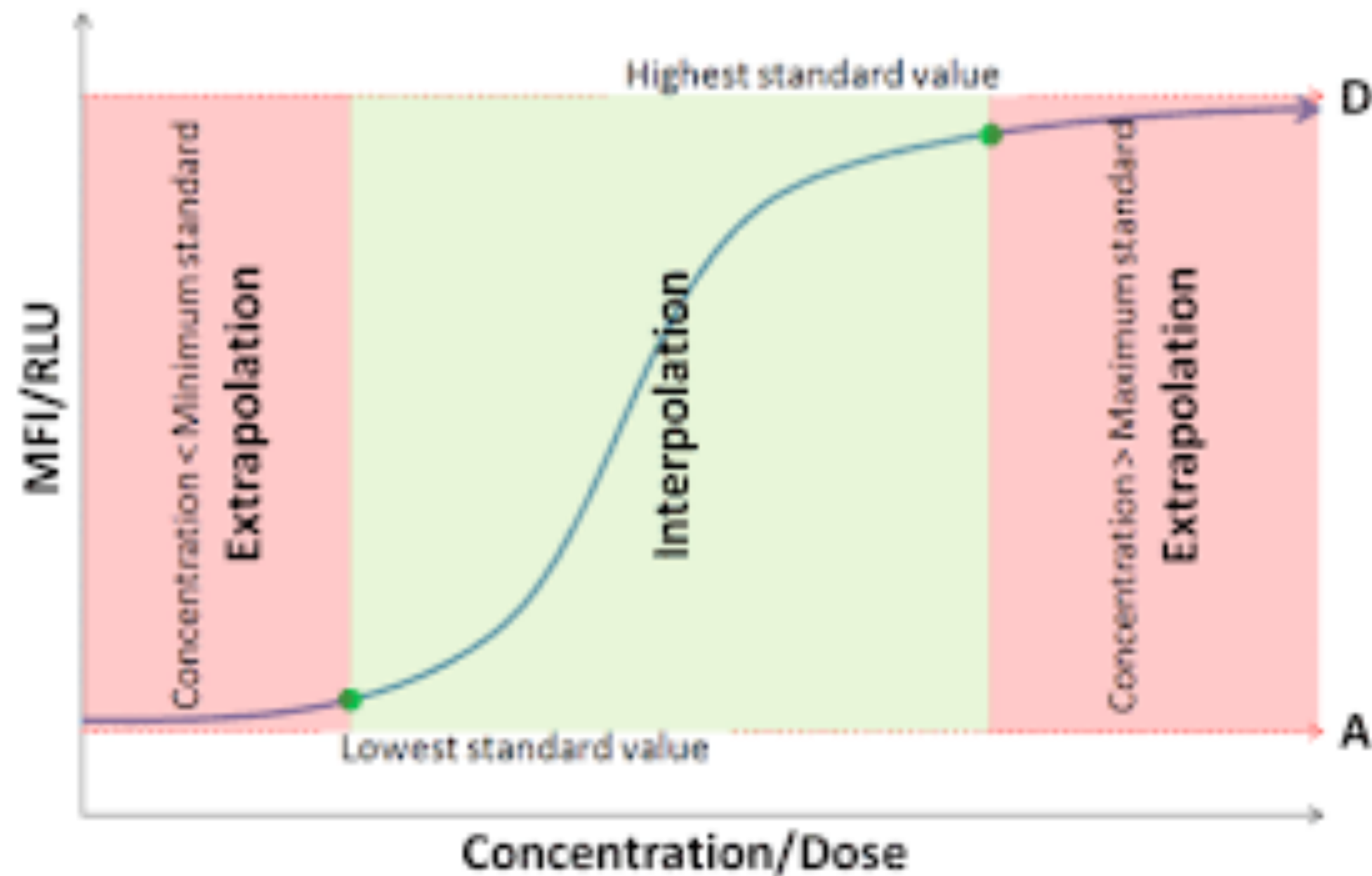


230924_7기_데이터분석기초

결측치 & 누락값 대체방법

보간법과 보외법 (수치해석)

- 보간법 (Interpolation)
- 내삽이라 하며 특정한 두 점 안쪽에 놓여있는 가능한 값을 구하려는 방법
- 보외법 (Extrapolation)
- 외삽 관찰 범위를 넘어서 다른 변수와의 관계에 기초하여 변수의 값을 추정하는 과정



E.g)라그랑주 보간법

Lagrange interpolation

수치해석에서 라그랑주 다항식은 라그랑주 형식에서 데이터 포인트의 주어진 집합으로부터 다항식을 보간하는 방법으로, [조제프루이 라그랑주](#)의 이름에서 왔다. 이것은 1779년 [에드워드 웨어링](#)에 의해 처음으로 발견되었고, 1783년에 [레온하르트 오일러](#)에 의해 마지막으로 재발견되었다.

정의 [편집]

$k + 1$ 데이터 포인트의 주어진 집합

$$(x_0, y_0), \dots, (x_j, y_j), \dots, (x_k, y_k)$$

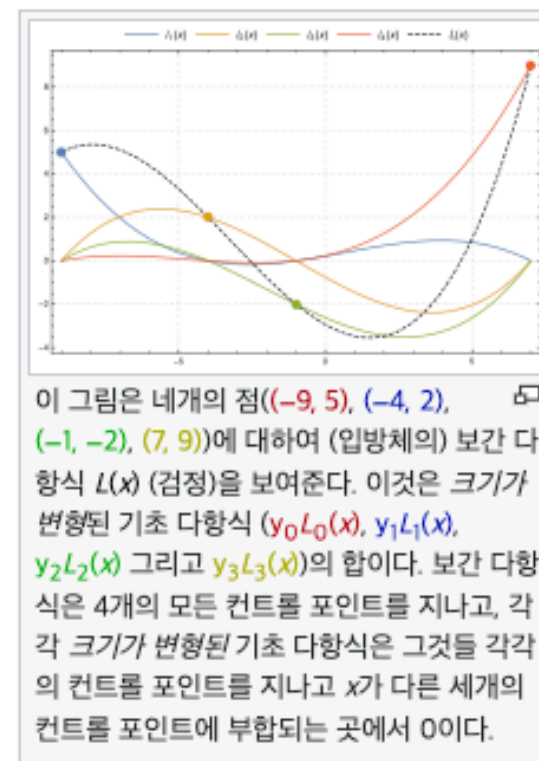
여기서 x_j 는 두 개의 같은 값이 존재하지 않고, 라그랑주 형식의 보간 다항식은 [선형 결합](#)

$$L(x) := \sum_{j=0}^k y_j \ell_j(x)$$

이다. 이것의 라그랑주 기초 다항식은 다음과 같다.

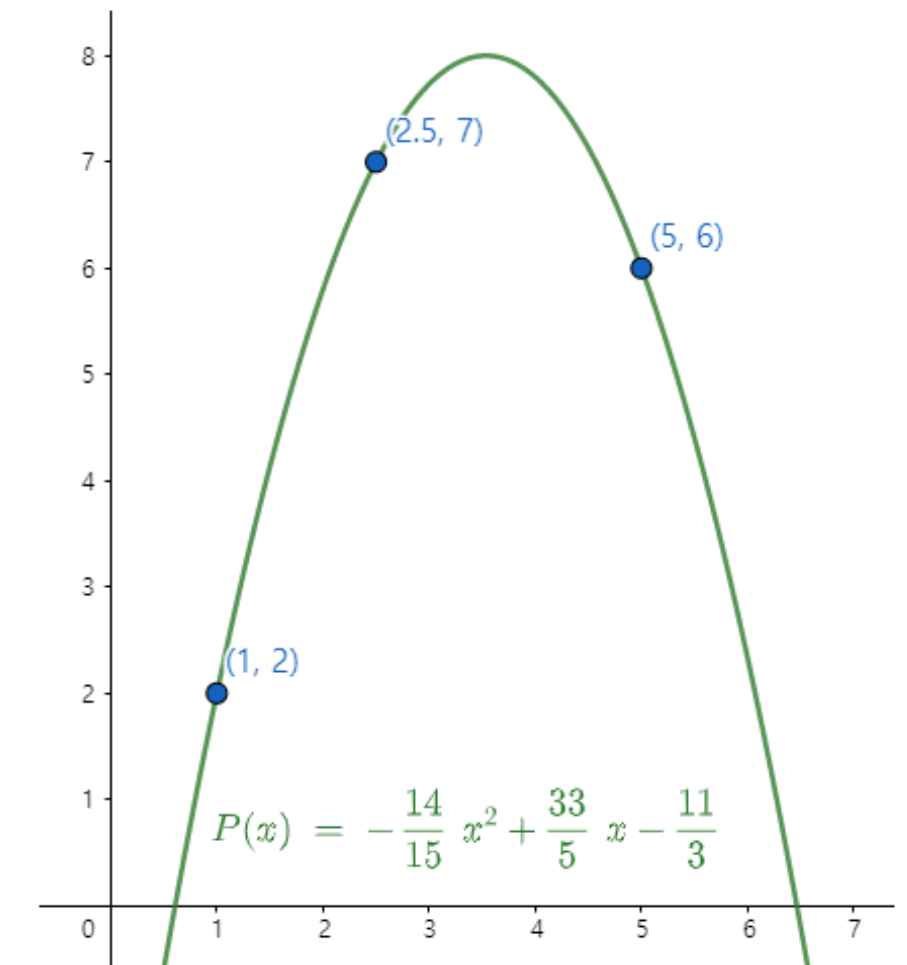
$$\ell_j(x) := \prod_{f=0, f \neq j}^k \frac{x - x_f}{x_j - x_f} = \frac{(x - x_0)}{(x_j - x_0)} \dots \frac{(x - x_{j-1})}{(x_j - x_{j-1})} \frac{(x - x_{j+1})}{(x_j - x_{j+1})} \dots \frac{(x - x_k)}{(x_j - x_k)}.$$

x_i 는 두 개의 같은 값이 존재하지 않기 때문에(그리고 존재할 수도 없다, 그렇지 않으면 데이터 집합의 의미가 모순된다), $x_j - x_f \neq 0$ 이 표현이 잘 정의 된다.



(1,2), (2.5,7), (5,6) 을 가지고 Ln,k 를 세우면

$$L_{2,0} = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{(x - 2.5)(x - 5)}{(1 - 2.5)(1 - 5)}$$
$$L_{2,1} = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = \frac{(x - 1)(x - 5)}{(2.5 - 1)(2.5 - 5)}$$
$$L_{2,2} = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{(x - 1)(x - 2.5)}{(5 - 1)(5 - 2.5)}$$

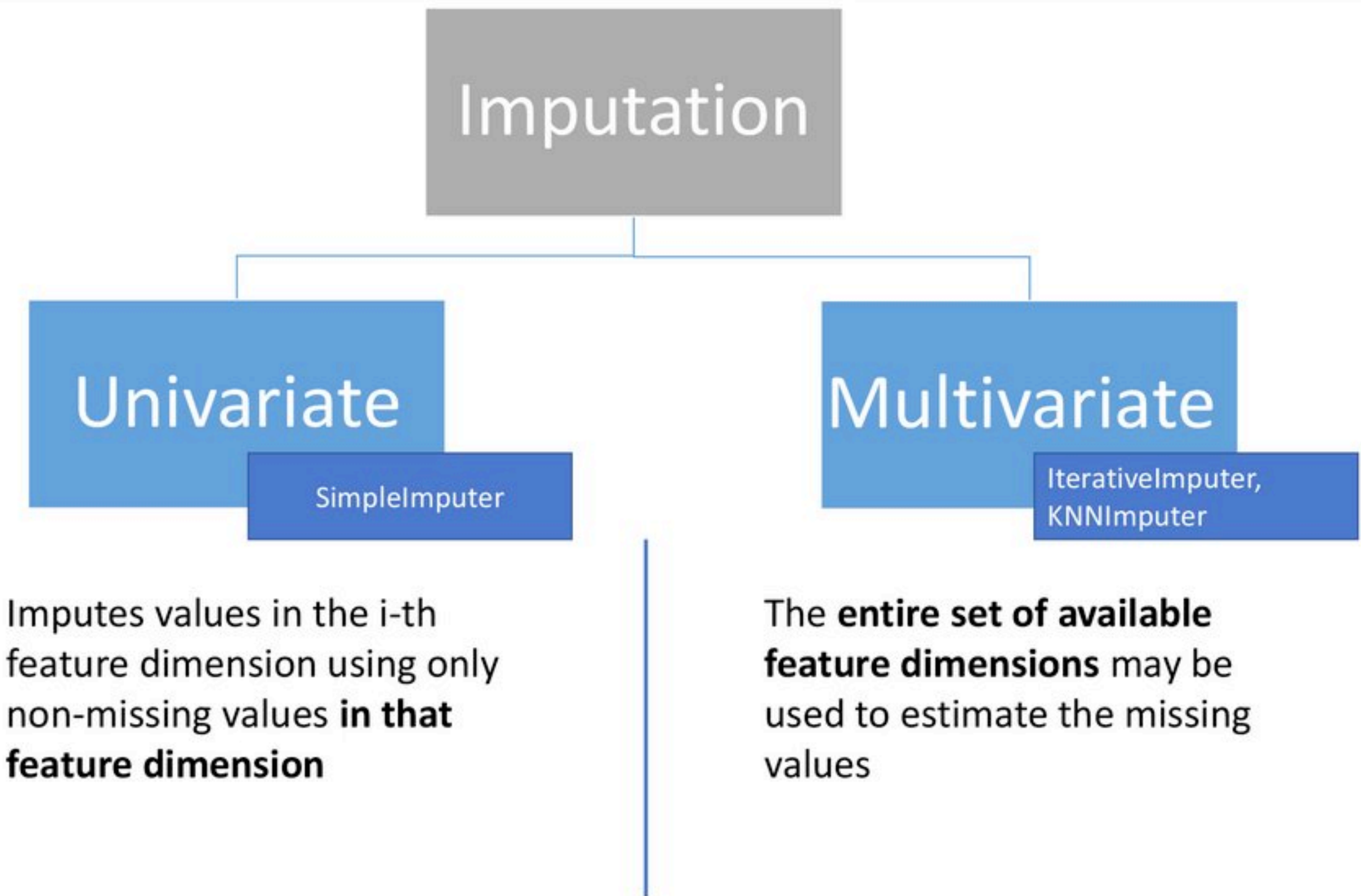
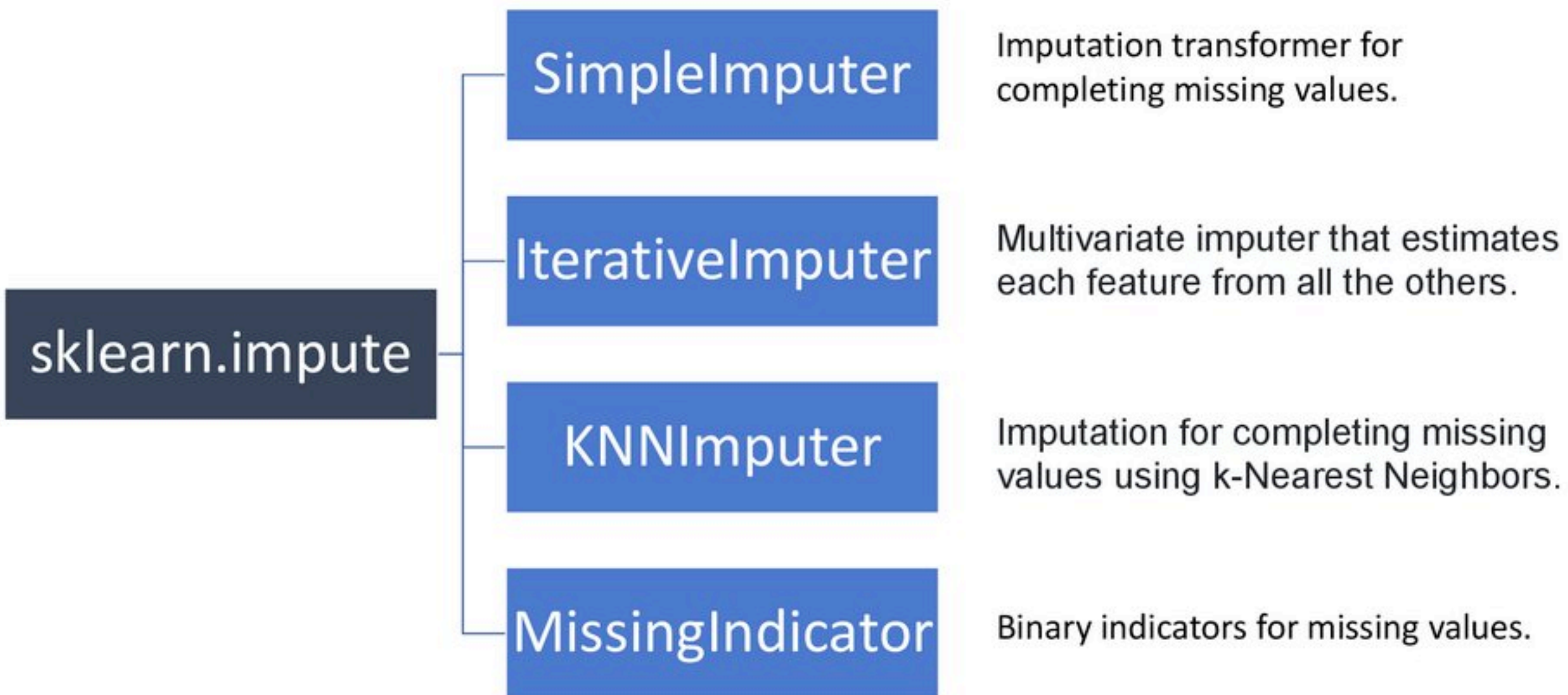


주어진 데이터와 L2,k 를 사용해 라그랑주 다항식을 세웁니다.

$$P(x) = \frac{2}{6}(x^2 - 7.5x + 12.5) - \frac{28}{15}(x^2 - 6x + 5) + \frac{6}{10}(x^2 - 3.5x + 2.5)$$
$$\Rightarrow P(x) = -\frac{14}{15}x^2 + \frac{33}{5}x - \frac{11}{3}$$

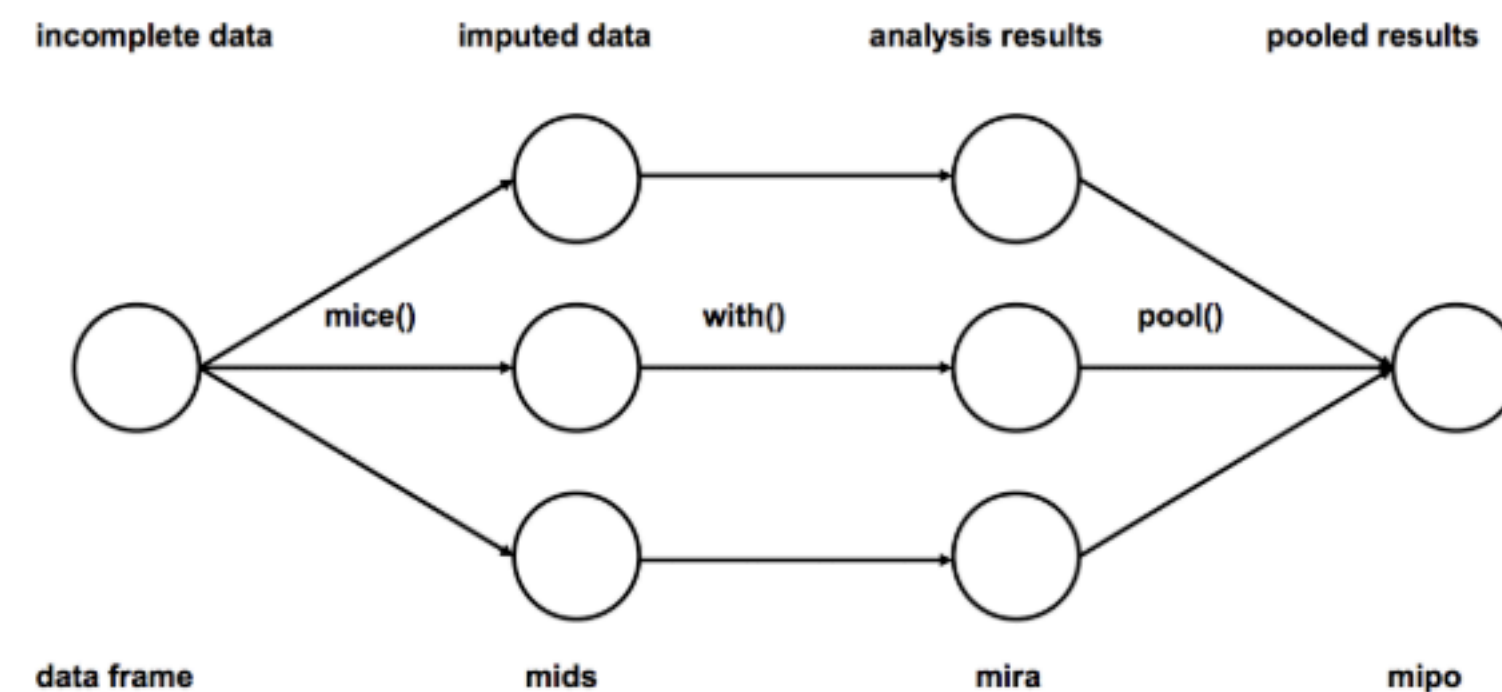
Sklearn impute

What's inside this module?



Iterative Imputer

1. **Imputation:** Impute the missing entries of the incomplete data sets m times ($m=3$ in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn't include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carlo (MCMC) simulation. This step results in m complete data sets.
2. **Analysis:** Analyze each of the m completed data sets.
3. **Pooling:** Integrate the m analysis results into a final result



단순 선형 회귀 대체법

확률적 회귀 대체법

다중 대체법

MICE 설명 :<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/>

몬테카를로 MCMC

A Markov Chain is a mathematical process that undergoes transitions from one state to another. Key properties of a Markov process are that it is random and that each step in the process is “memoryless;” in other words, the future state depends only on the current state of the process and not the past

<https://www.publichealth.columbia.edu/research/population-health-methods/markov-chain-monte-carlo>