実践法としてのデータ科学

CDS 早稲田大学
データ科学センター

内容

- データ分析の課題に対して，どのように意思決定写像を定めるのかを例で説明

例　構造推定としての回帰問題

目的:	説明変数$x_1, \ldots, x_n$が与えられたもとでの$\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n$の確率的データ生成観測メカニズムを明らかにしたい
設定:	$\underset{\sim}{y}_i = \beta_0 + \beta_1 x_i + \underset{\sim}{\varepsilon}_i, \quad i = 1, \ldots, n$ $\underset{\sim}{\varepsilon}_i$はi.i.d.で正規分布$\mathcal{N}(0, \sigma_\varepsilon^2)$に従う
評価基準:	尤度最大化

2変数データ
（量的データと量的データ）
$(x_1, \underset{\sim}{y}_1), \ldots, (x_n, \underset{\sim}{y}_n)$

→ **意思決定写像** →

母回帰係数と誤差項の分散の推定量
$\hat{\underset{\sim}{\beta}}_0, \hat{\underset{\sim}{\beta}}_1, \widehat{\underset{\sim}{\sigma_\varepsilon^2}}$

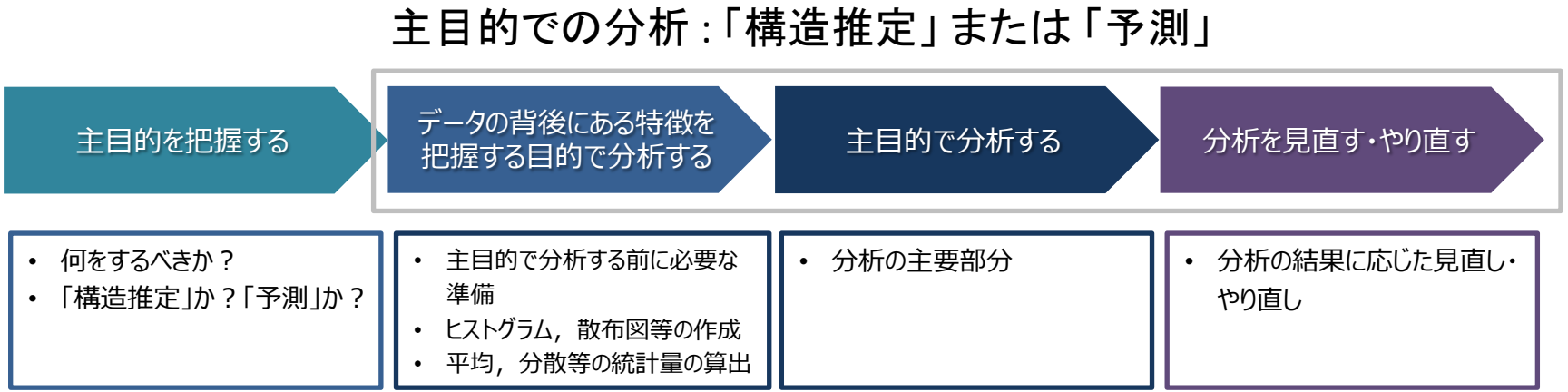
講義科目「データ科学実践」のご紹介

対象者：「データ科学入門α・β・γ・δ」を学んだ学生

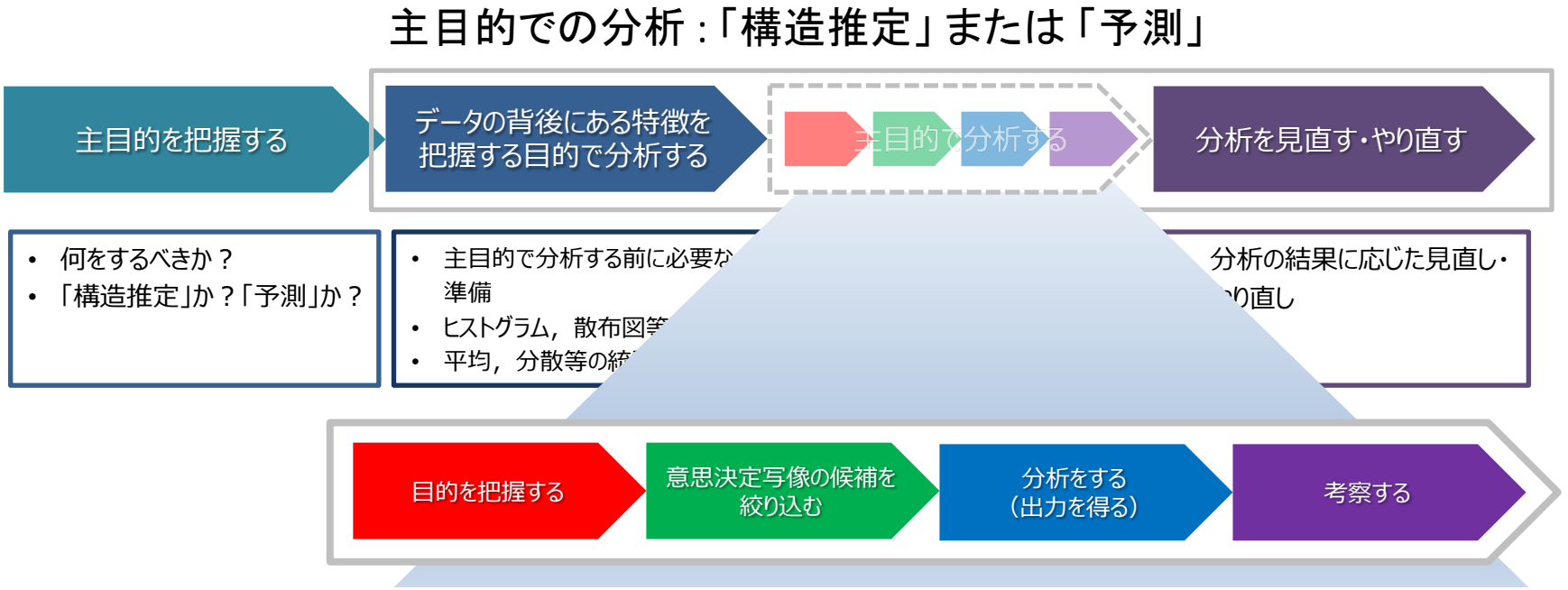
講義内容：いくつかのデータ分析の課題について，分析の筋道を立て、実際に分析を行って結果を考察するまでの一連の流れを演習する

講義形式：フルオンデマンド

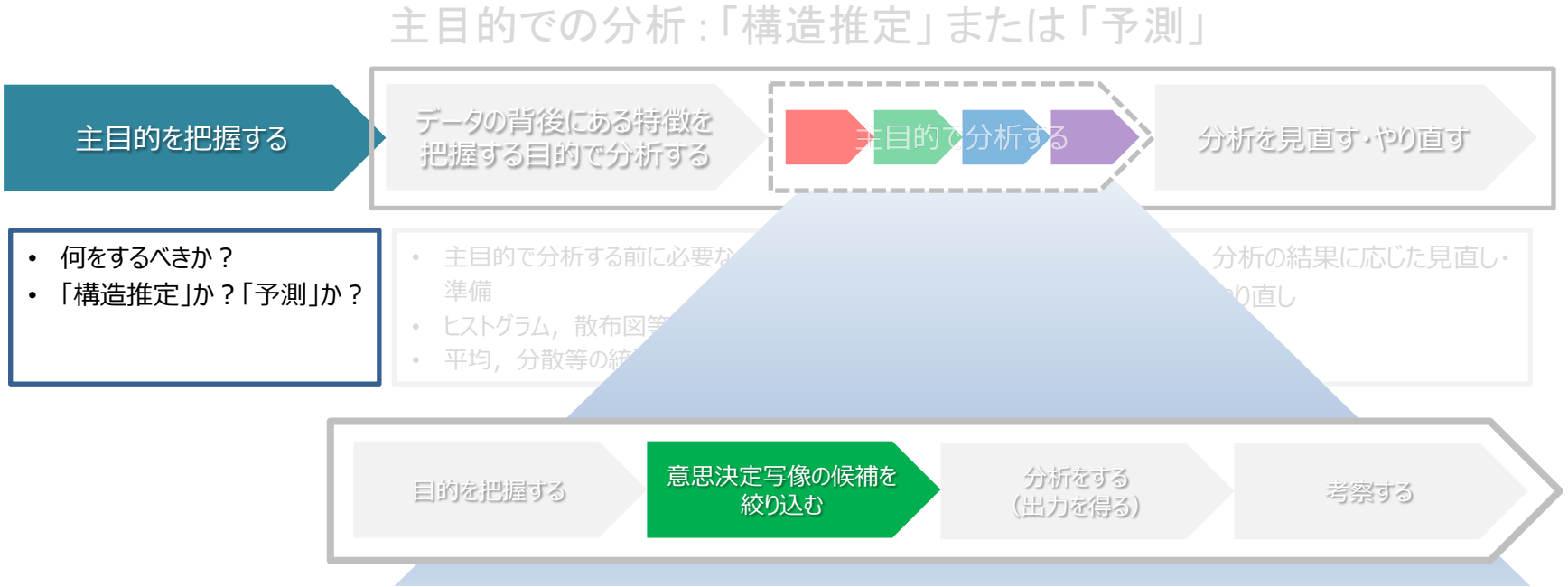
データ分析の全体の流れ



データ分析の全体の流れ



データ分析の全体の流れ



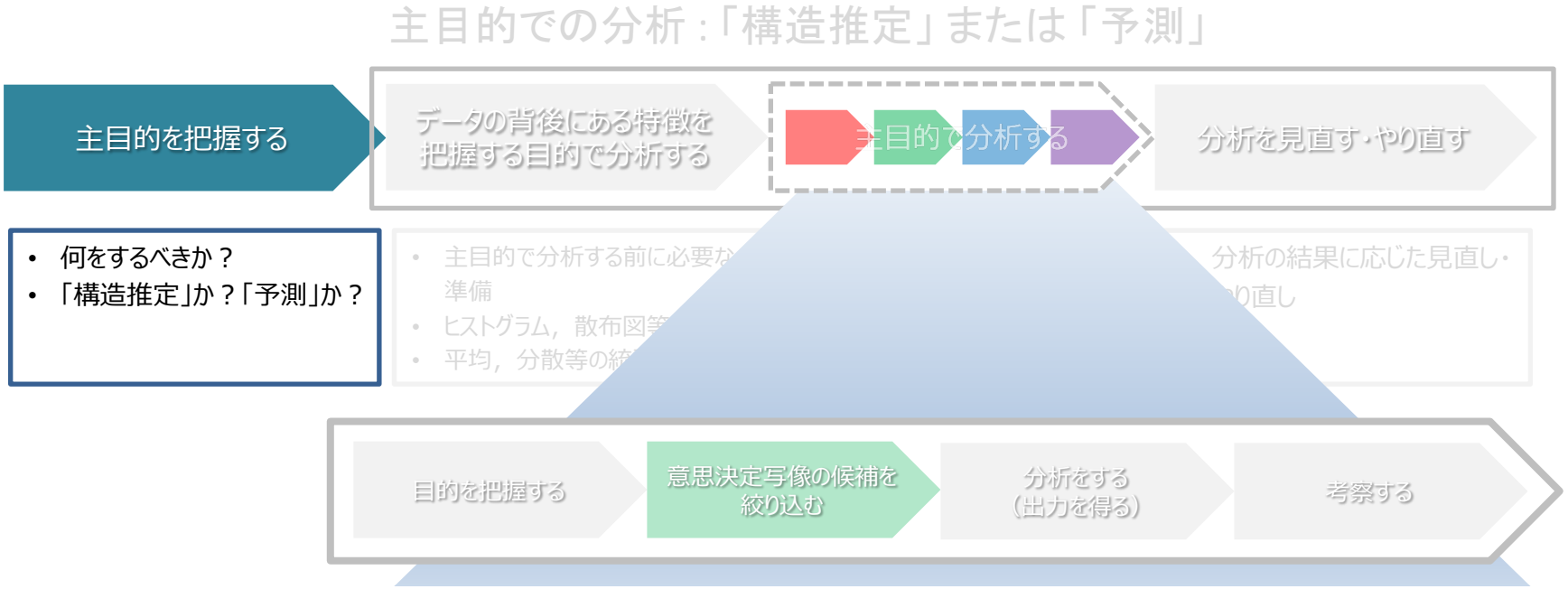
データ分析の課題

ある不動産会社では,

- 過去に販売した中古住宅に関するデータが得られている
- 物件情報から価格がどのように定まるのかを知りたいと考えている

No.	面積（m^2）	築年数（年）	主要駅からの電車時間（分）	最寄り駅までの徒歩時間（分）	建物種別	駐車施設	価格（百万円）
1	125	26	25	21	戸建て	有	26.9
2	85	18	19	11	戸建て	有	36.8
3	100	25	30	9	戸建て	有	7.5
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

データ分析の全体の流れ



主目的の把握

目的：構造推定［物件情報から価格が定まる構造を表す式を推定する］

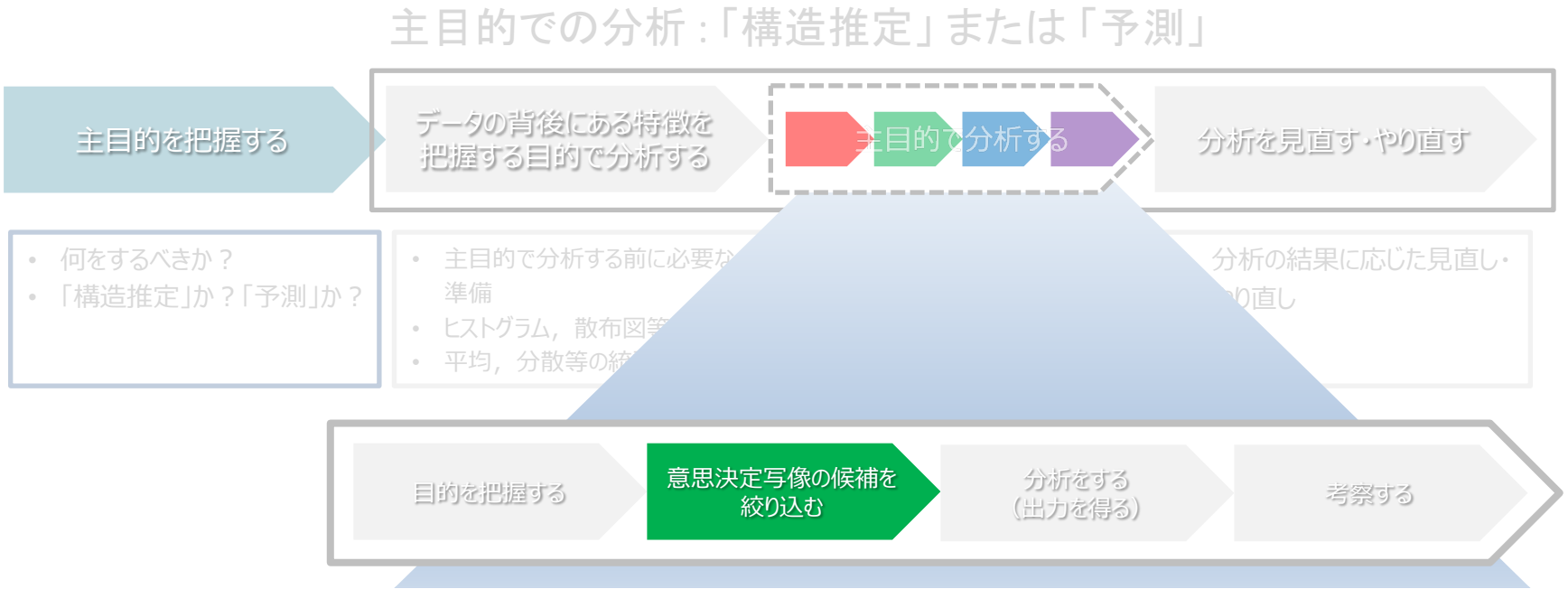
説明変数：面積，築年数，主要駅からの電車時間，徒歩時間，
建物種別，駐車施設

目的変数：価格（量的変数，観測可能）

問題の種類：回帰

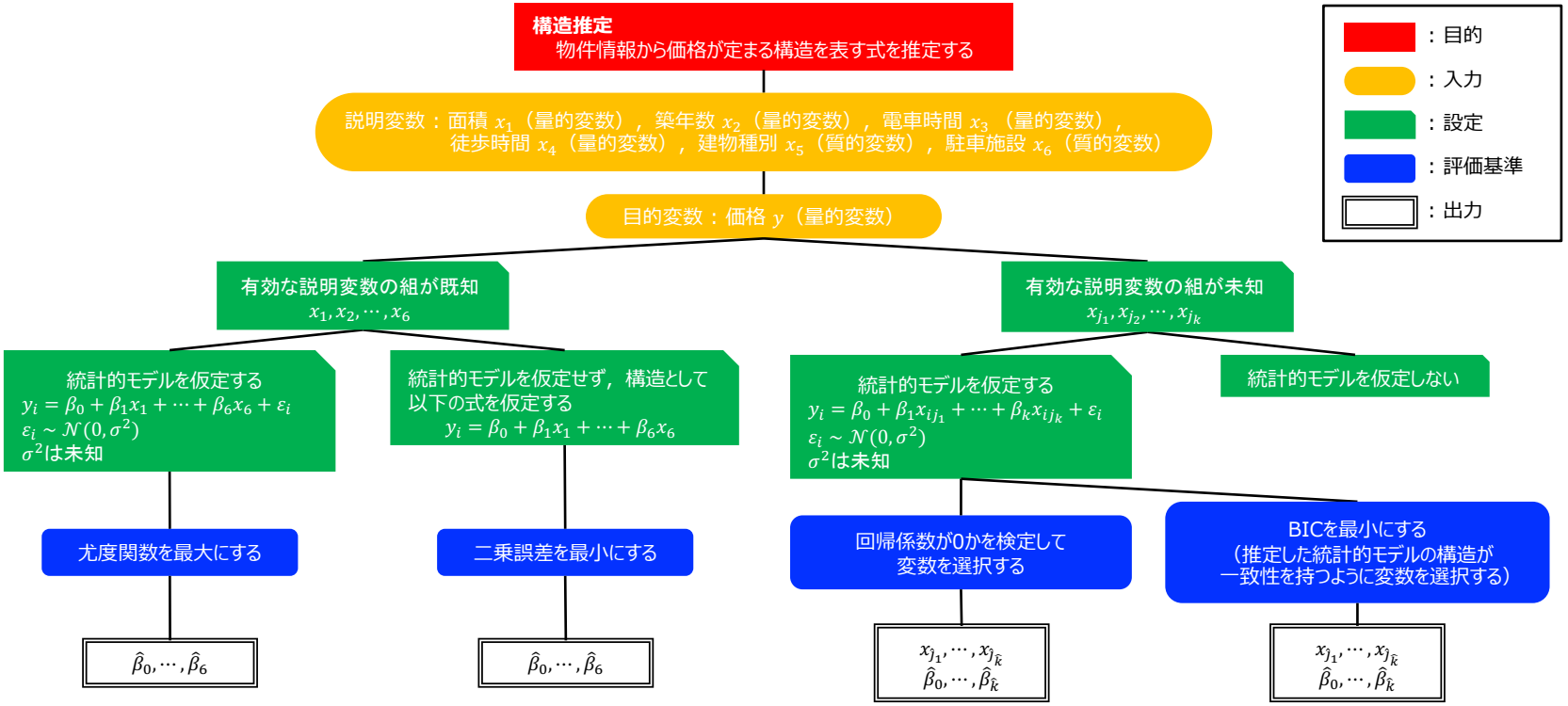
目的変数	量的	質的
観測可能	**回帰**	**分類**
観測不可能	**縮約**	**クラスタリング**

データ分析の全体の流れ



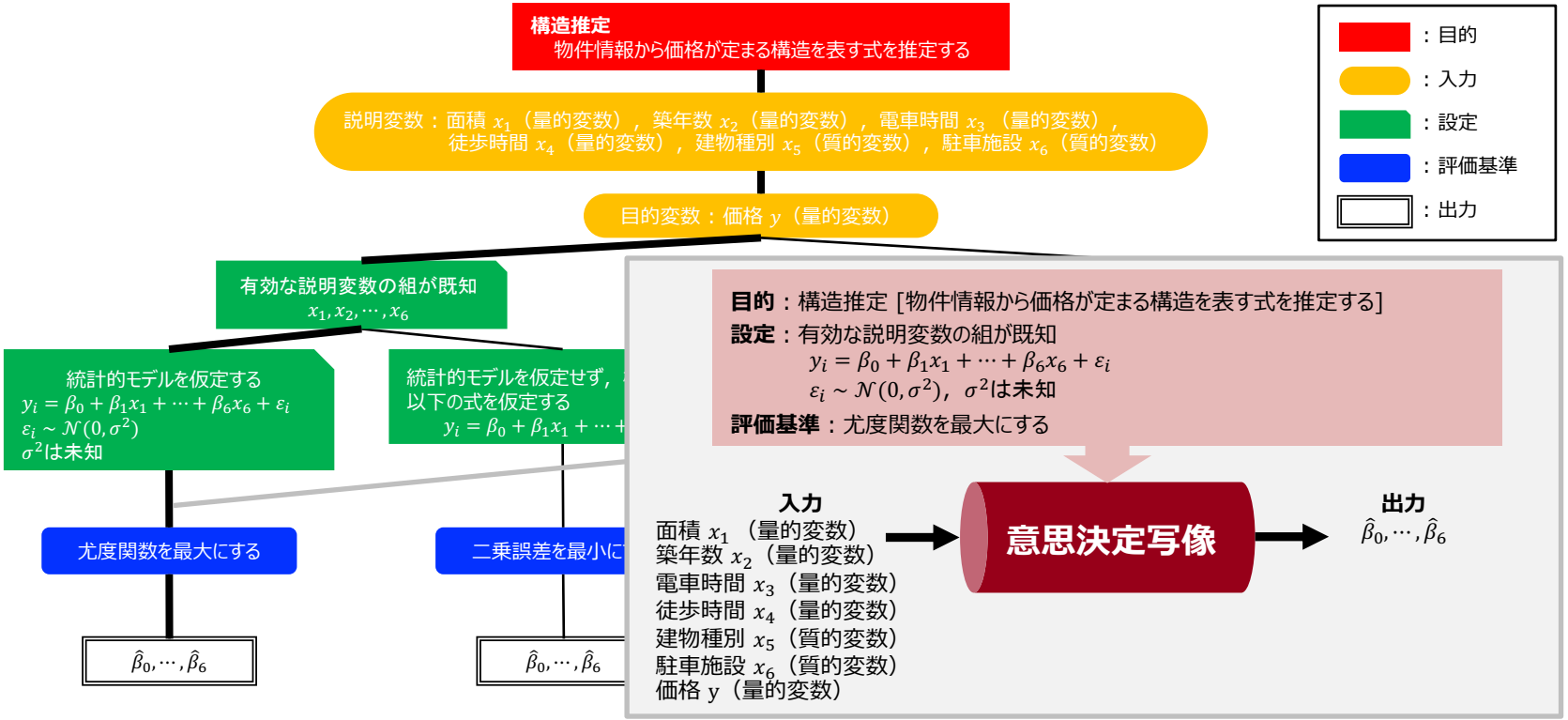
意思決定写像の候補の絞り込み

意思決定写像構成木



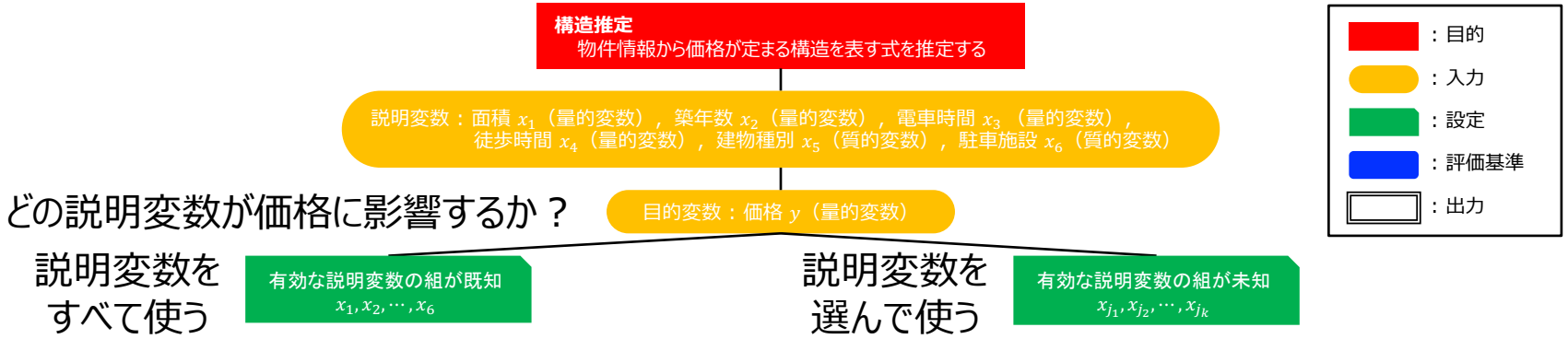
意思決定写像の候補の絞り込み

意思決定写像構成木



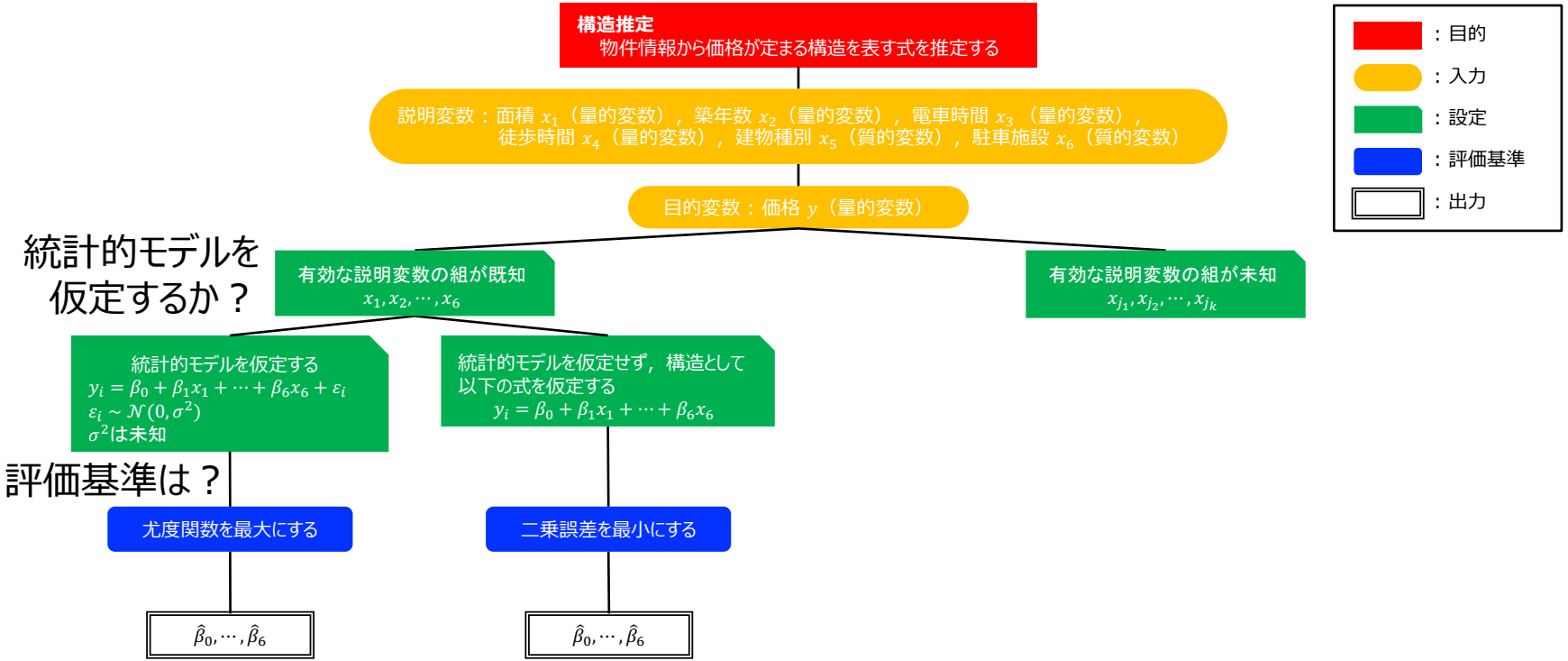
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



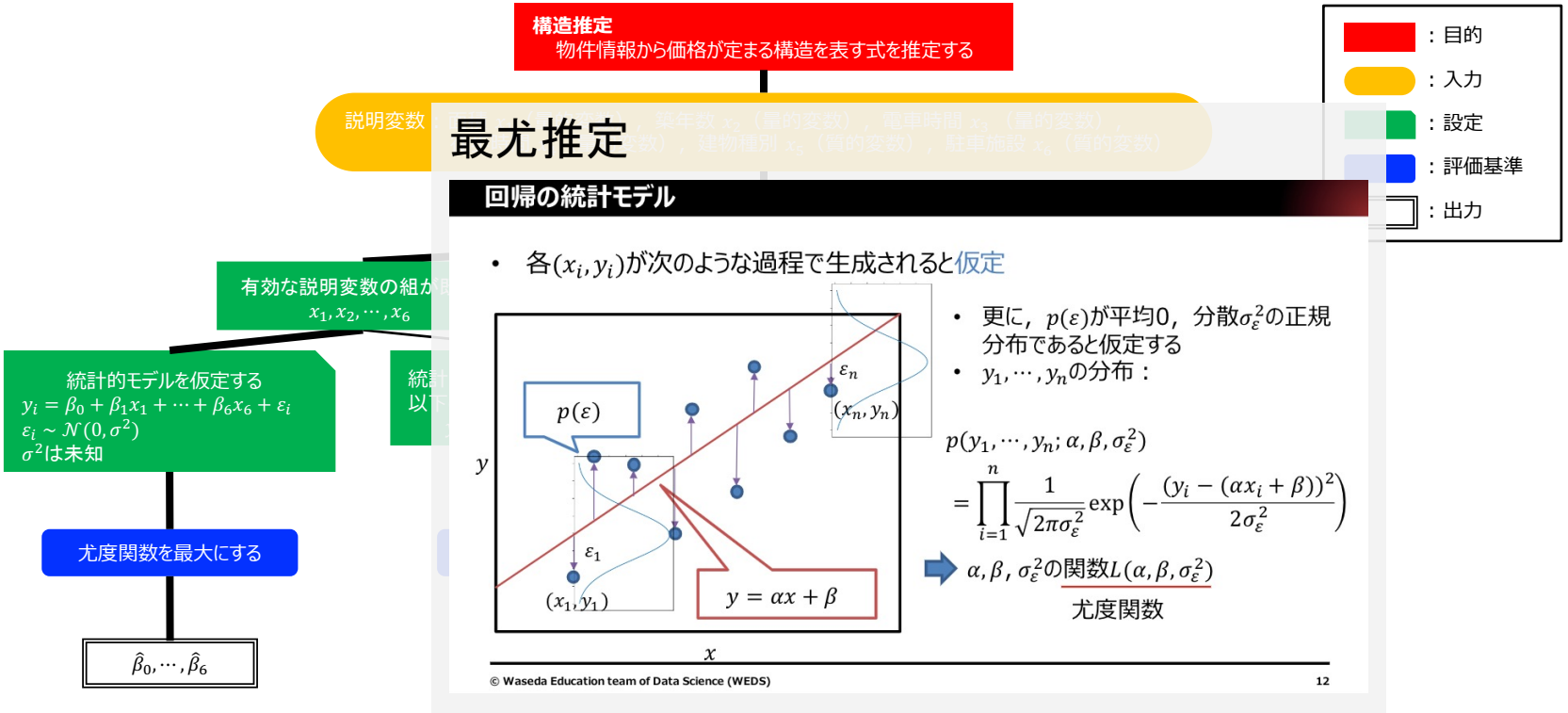
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



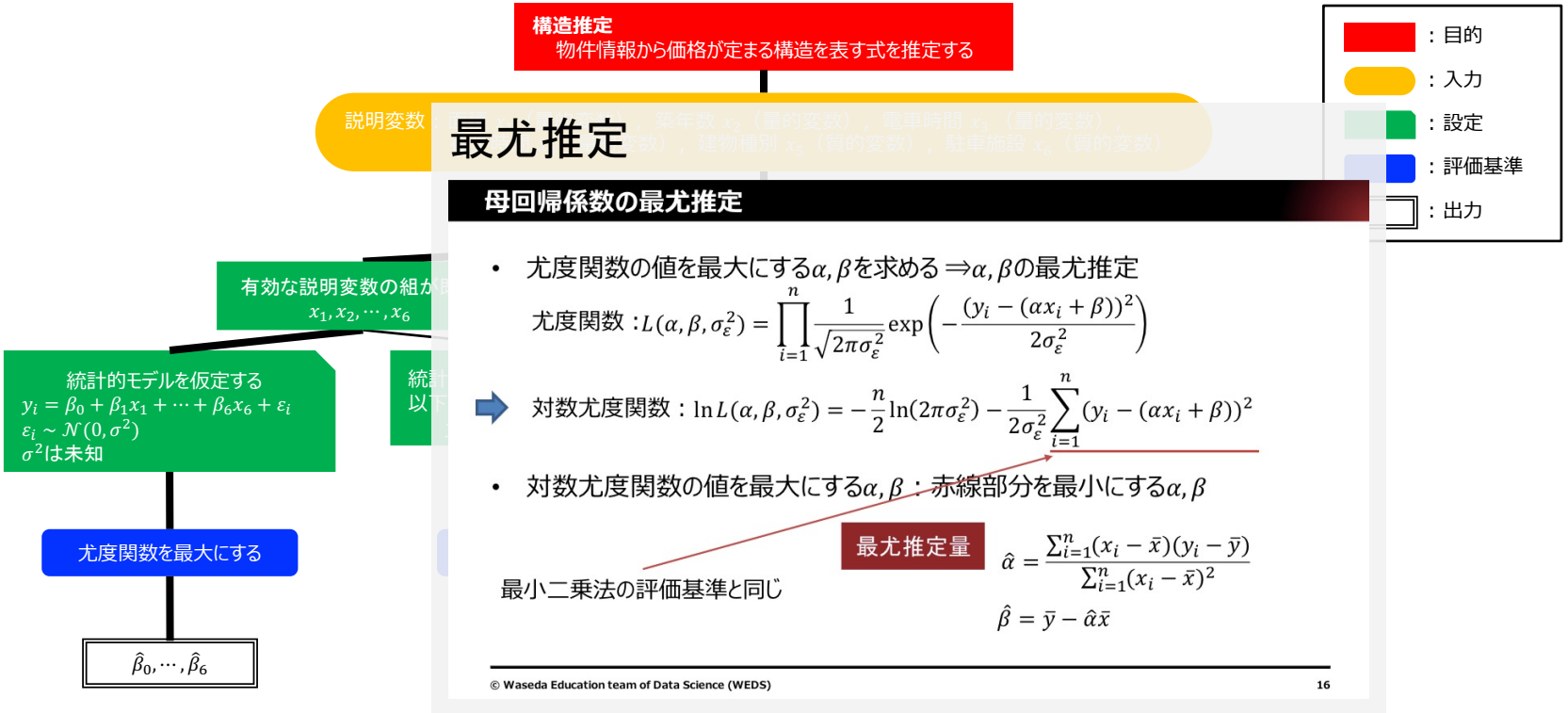
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



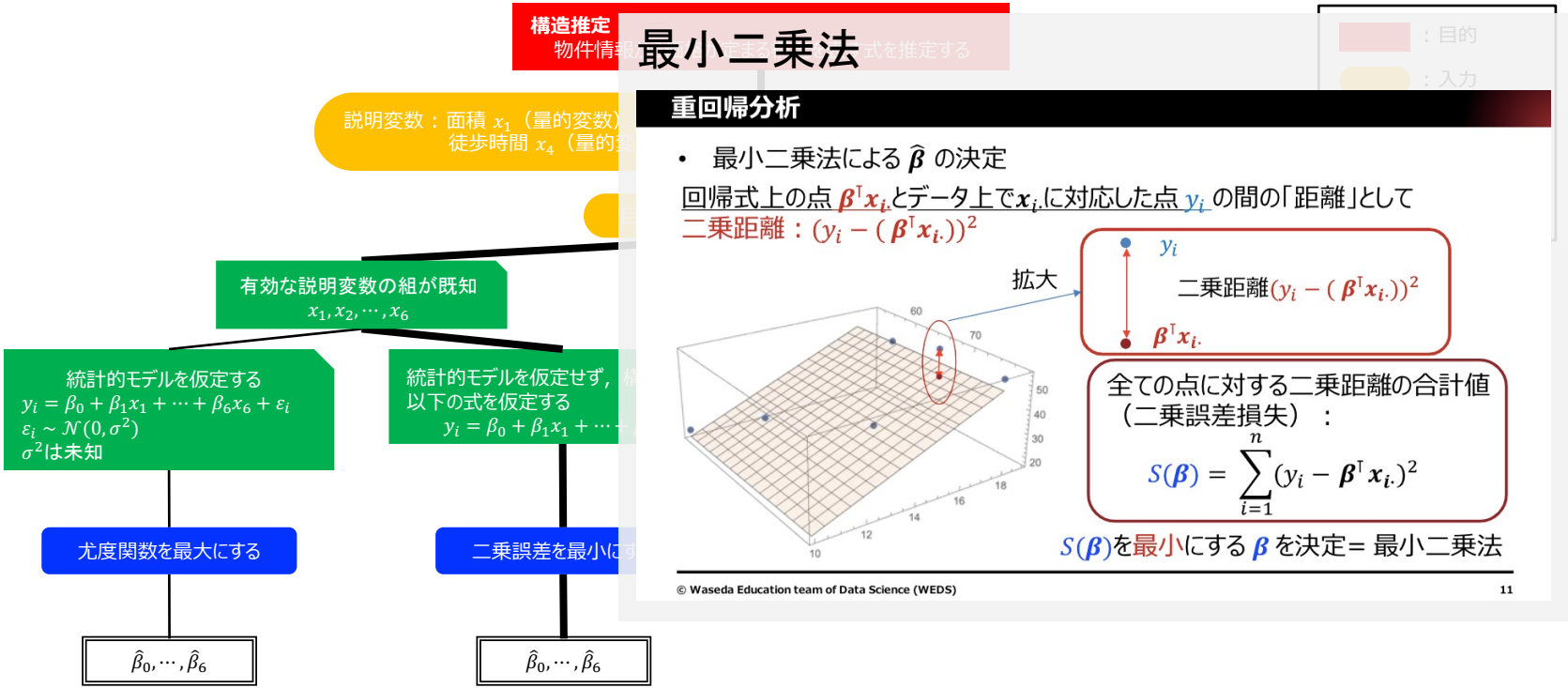
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



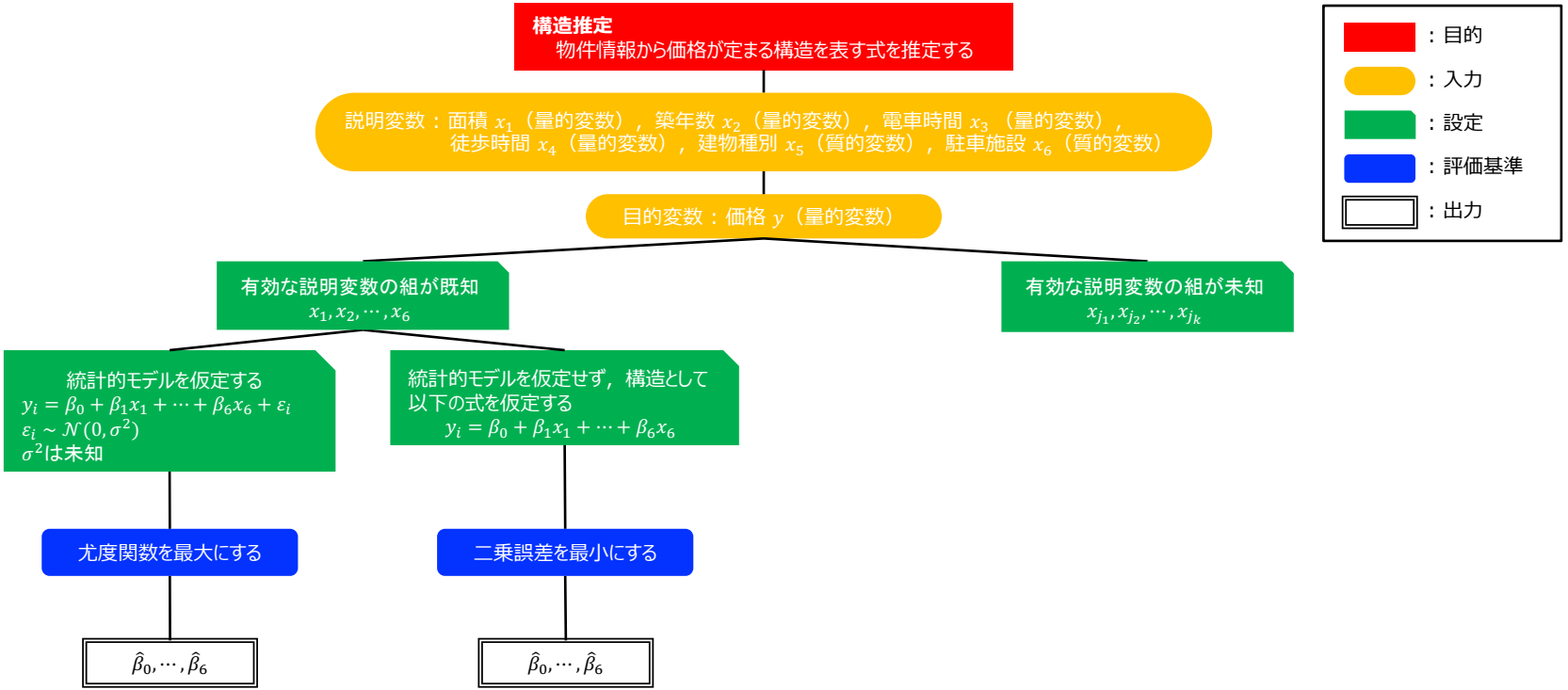
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



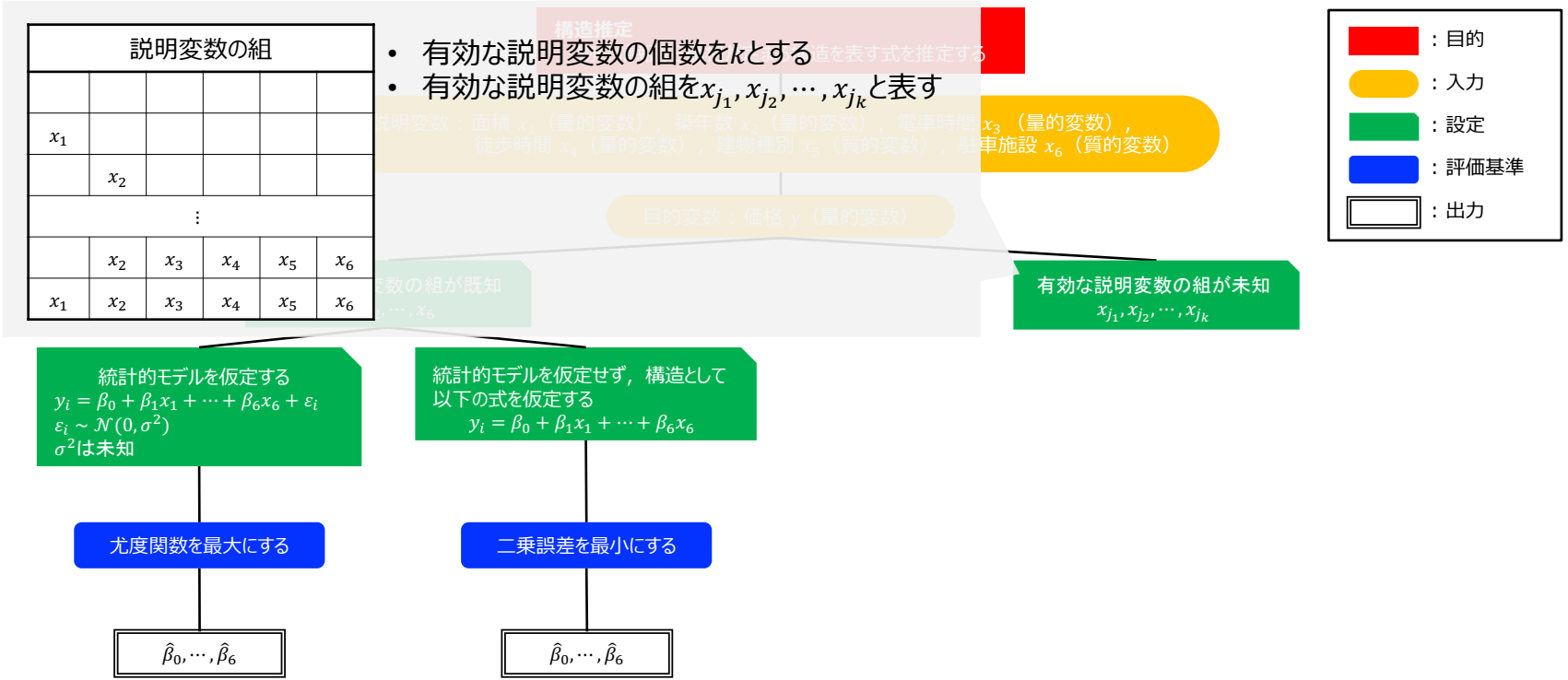
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



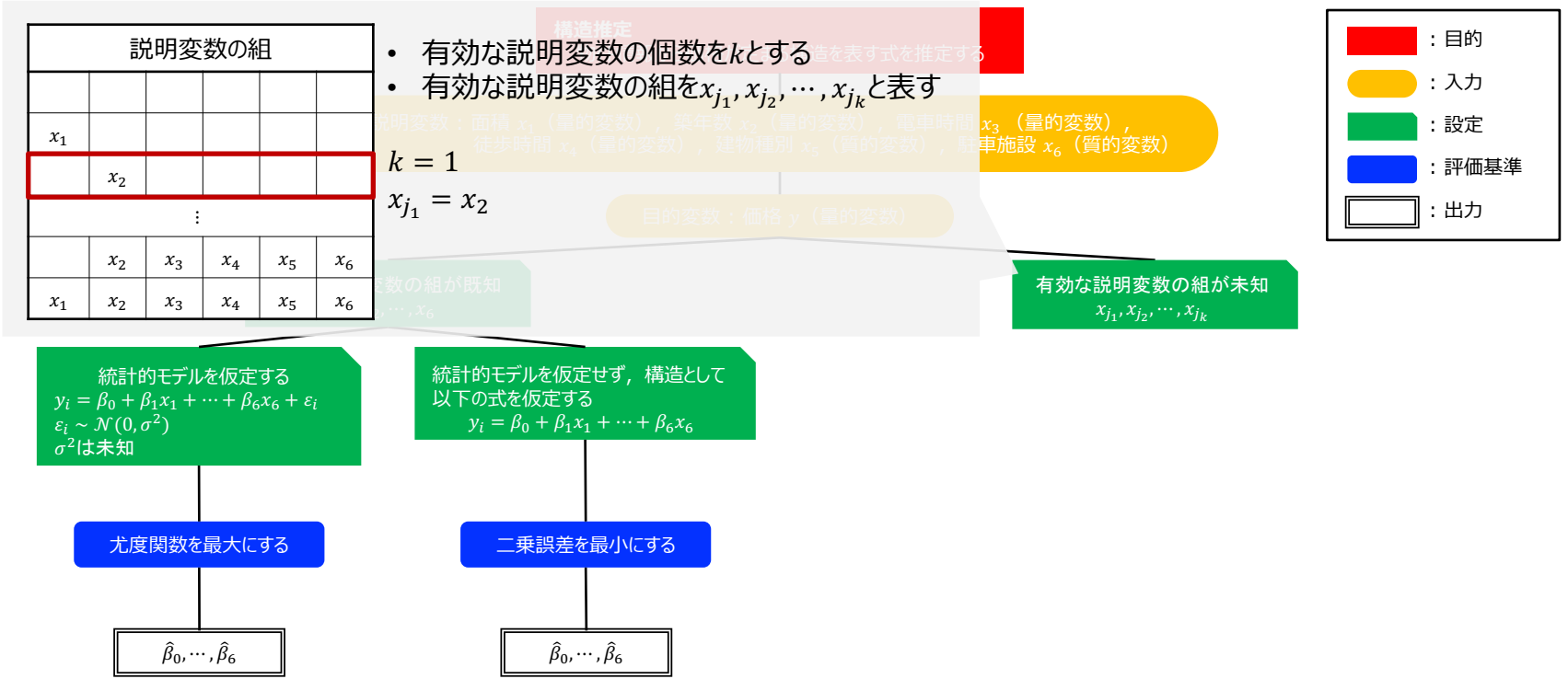
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



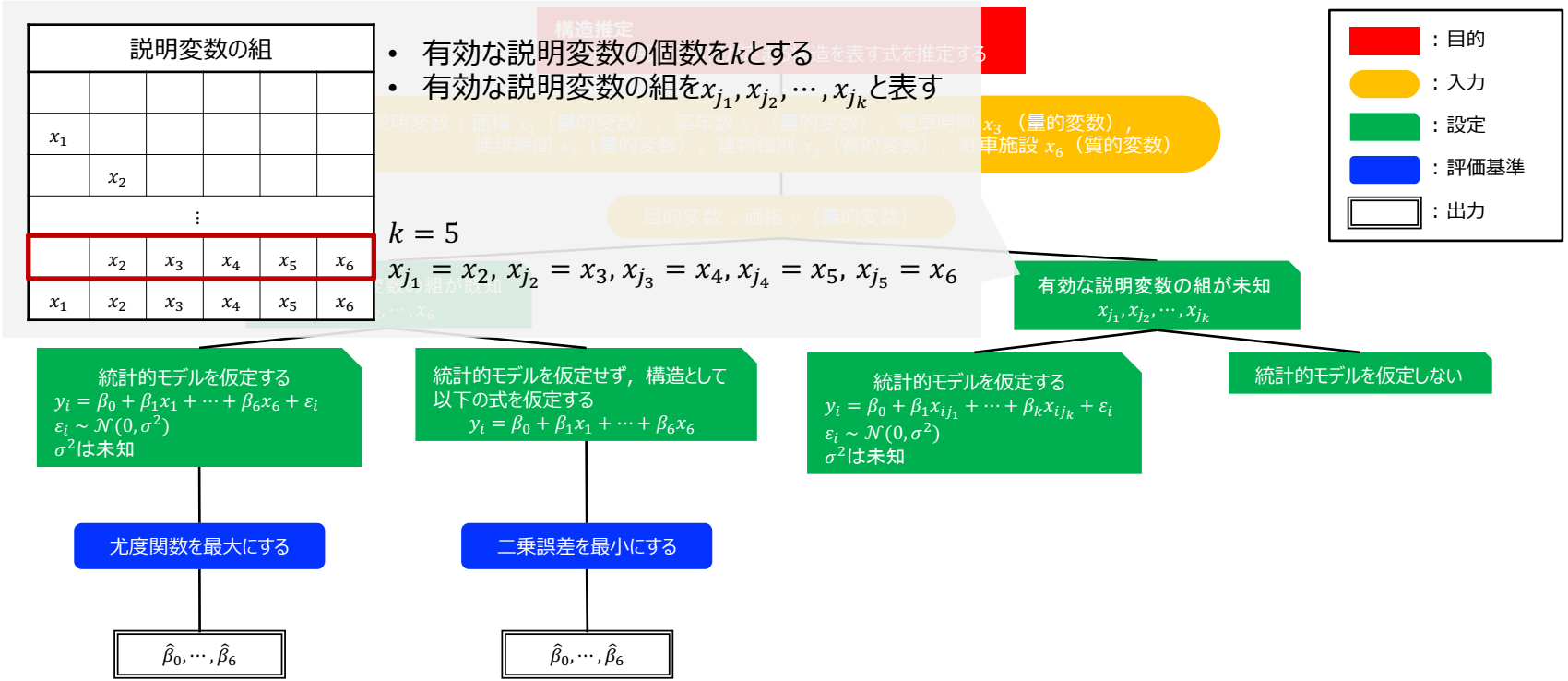
意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる



意思決定写像の候補の絞り込み

- 目的，設定，評価基準を具体化し，それらの候補を挙げる

説明変数の組						統計的モデル
						$y_i = \beta_0 + \varepsilon_i$
x_1						$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$
	x_2					$y_i = \beta_0 + \beta_2 x_{i2} + \varepsilon_i$
⋮						⋮
	x_2	x_3	x_4	x_5	x_6	$y_i = \beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$
x_1	x_2	x_3	x_4	x_5	x_6	$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i$

を推定する

電車時間 x_3（量的変数），
数），駐車施設 x_6（質的変数）

：目的
：入力
：設定
：評価基準
：出力

有効な説明変数の組が未知
$x_{j_1}, x_{j_2}, \cdots, x_{j_k}$

統計的モデルを仮定する
$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6 + \varepsilon_i$
$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
σ^2は未知

尤度関数を最大にする

$\hat{\beta}_0, \cdots, \hat{\beta}_6$

統計的モデルを仮定せず，構造として
以下の式を仮定する
$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6$

二乗誤差を最小にする

$\hat{\beta}_0, \cdots, \hat{\beta}_6$

統計的モデルを仮定する
$y_i = \beta_0 + \beta_1 x_{ij_1} + \cdots + \beta_k x_{ij_k} + \varepsilon_i$
$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$
σ^2は未知

回帰係数が0かを検定して
変数を選択する

$x_{j_1}, \cdots, x_{j_{\hat{k}}}$
$\hat{\beta}_0, \cdots, \hat{\beta}_{\hat{k}}$

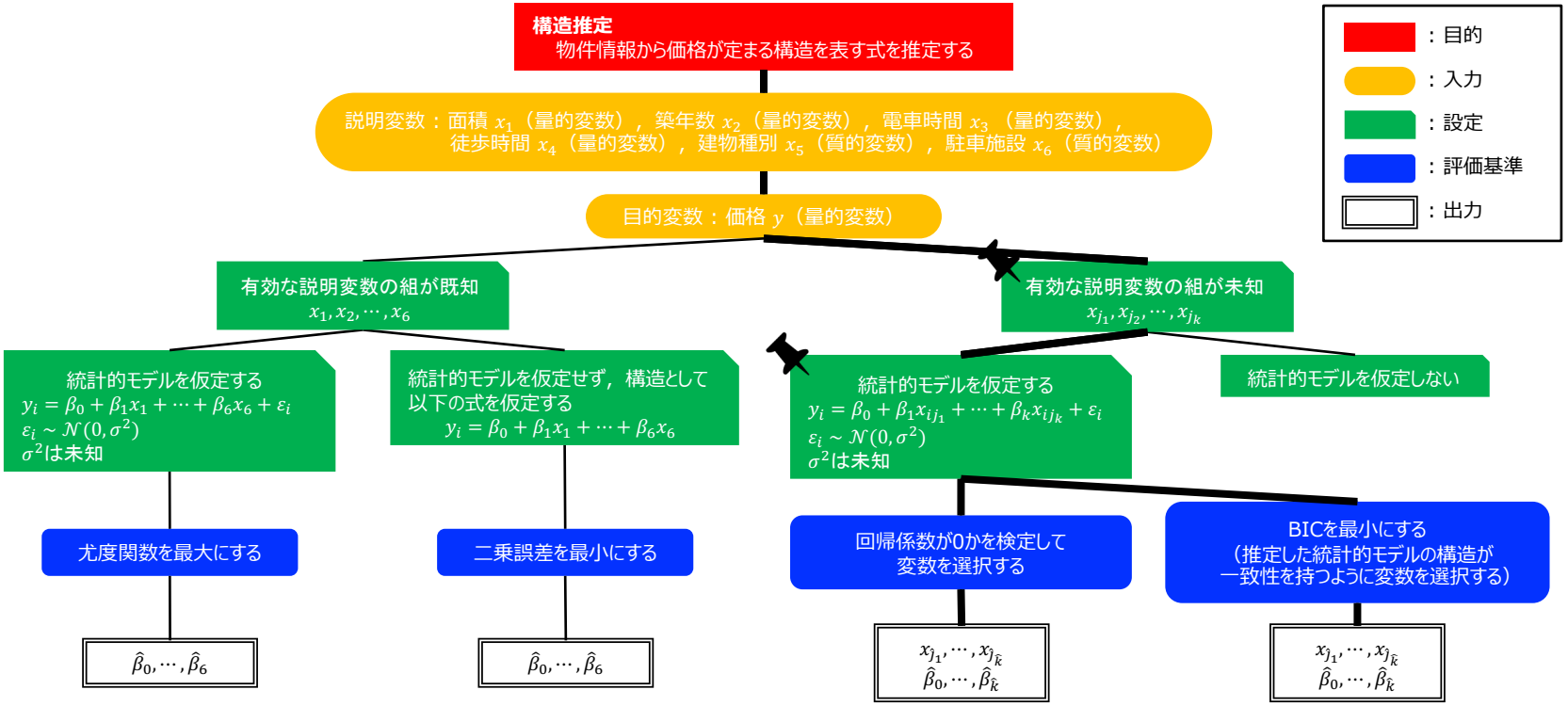
統計的モデルを仮定しない

BICを最小にする
（推定した統計的モデルの構造が
一致性を持つように変数を選択する）

$x_{j_1}, \cdots, x_{j_{\hat{k}}}$
$\hat{\beta}_0, \cdots, \hat{\beta}_{\hat{k}}$

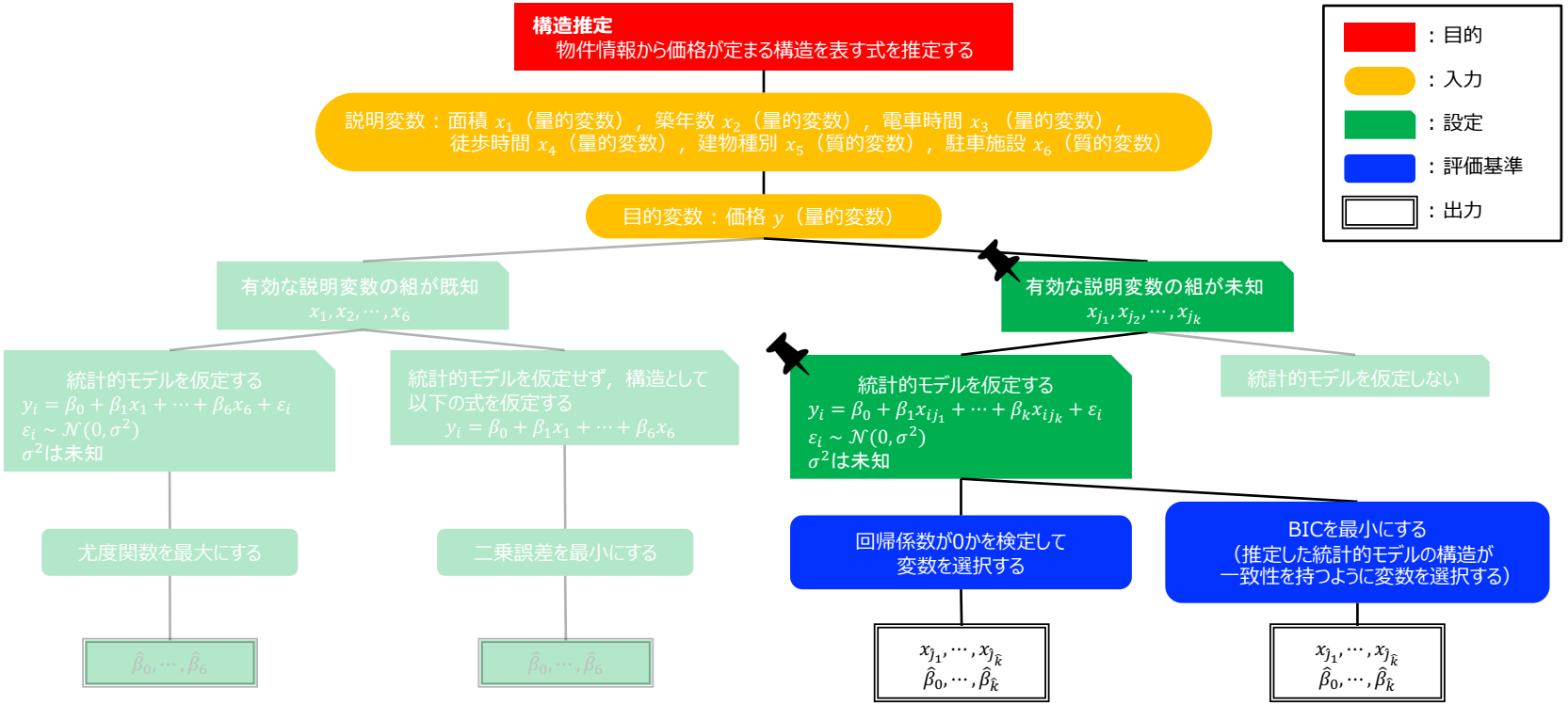
意思決定写像の候補の絞り込み

- 候補の中から，目的，設定，評価基準を定める



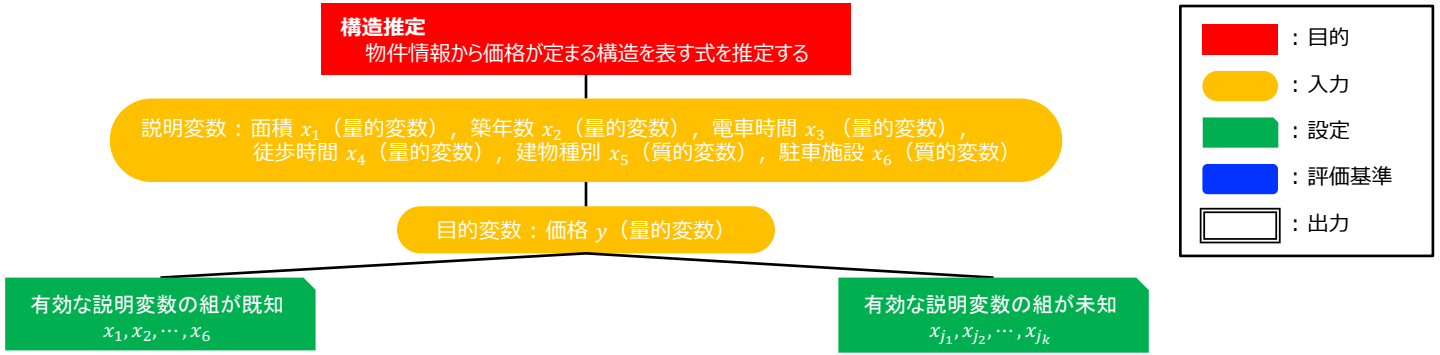
意思決定写像の候補の絞り込み

- 候補の中から，目的，設定，評価基準を定める



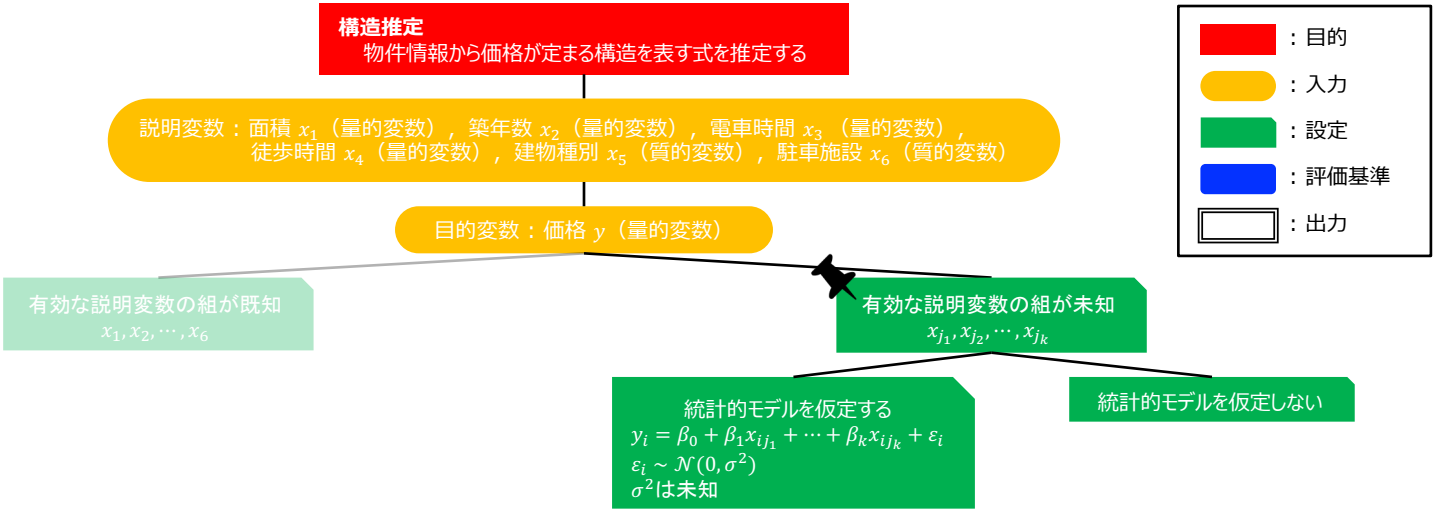
意思決定写像の候補の絞り込み

すべての分岐を考えなくてもよい



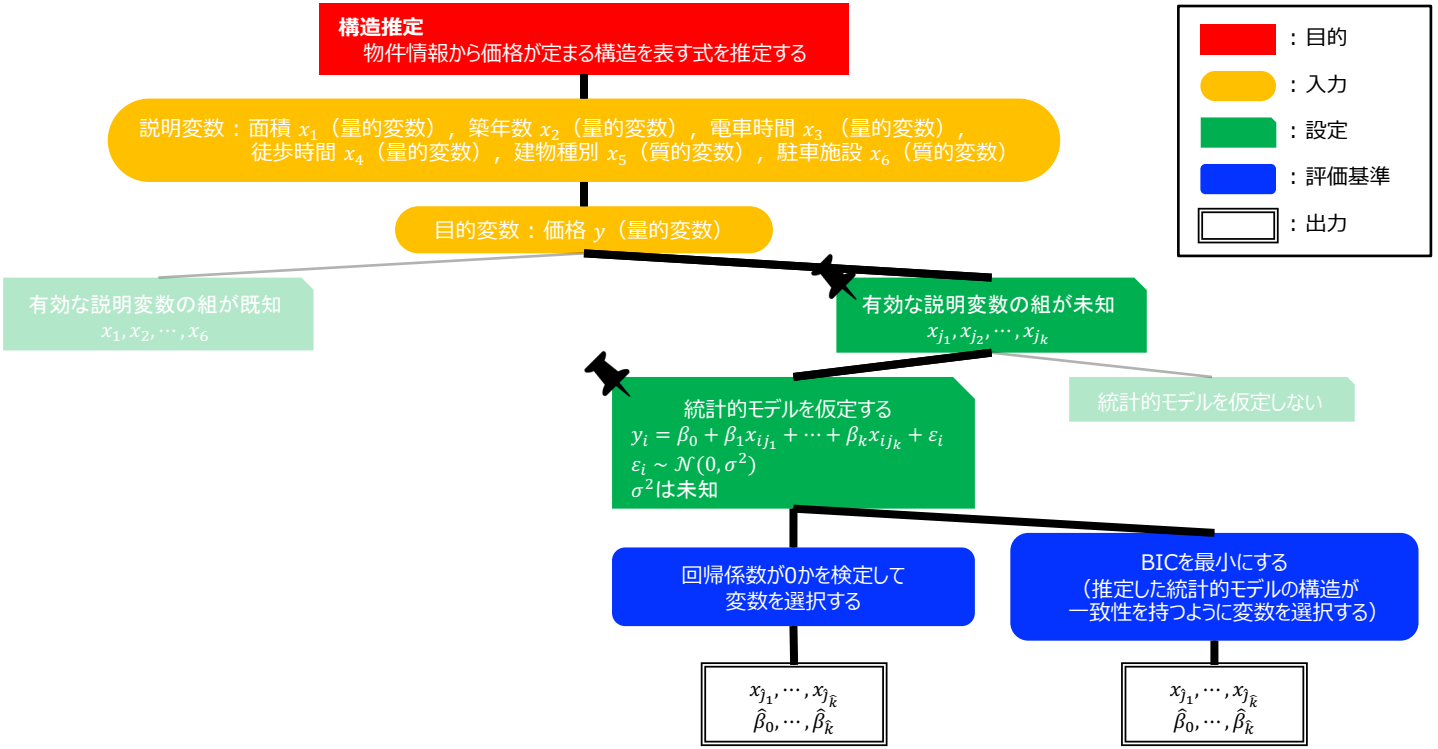
意思決定写像の候補の絞り込み

すべての分岐を考えなくてもよい



意思決定写像の候補の絞り込み

すべての分岐を考えなくてもよい



まとめ

- 「主目的の把握」と「意思決定写像の候補の絞り込み」のプロセスを通して，どのように意思決定写像を定めるかを説明
- 「意思決定写像構成木」のような考え方によって，意思決定写像を定めることができる