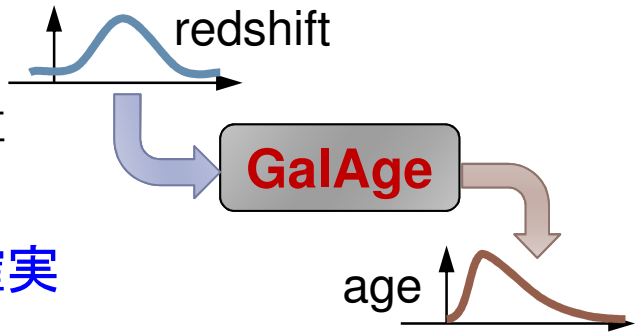
【VLDB2013勉強会】

Session 16: Concurrency and Query Processing

担当：董（名古屋大学）

Supporting User-Defined Functions on Uncertain Data

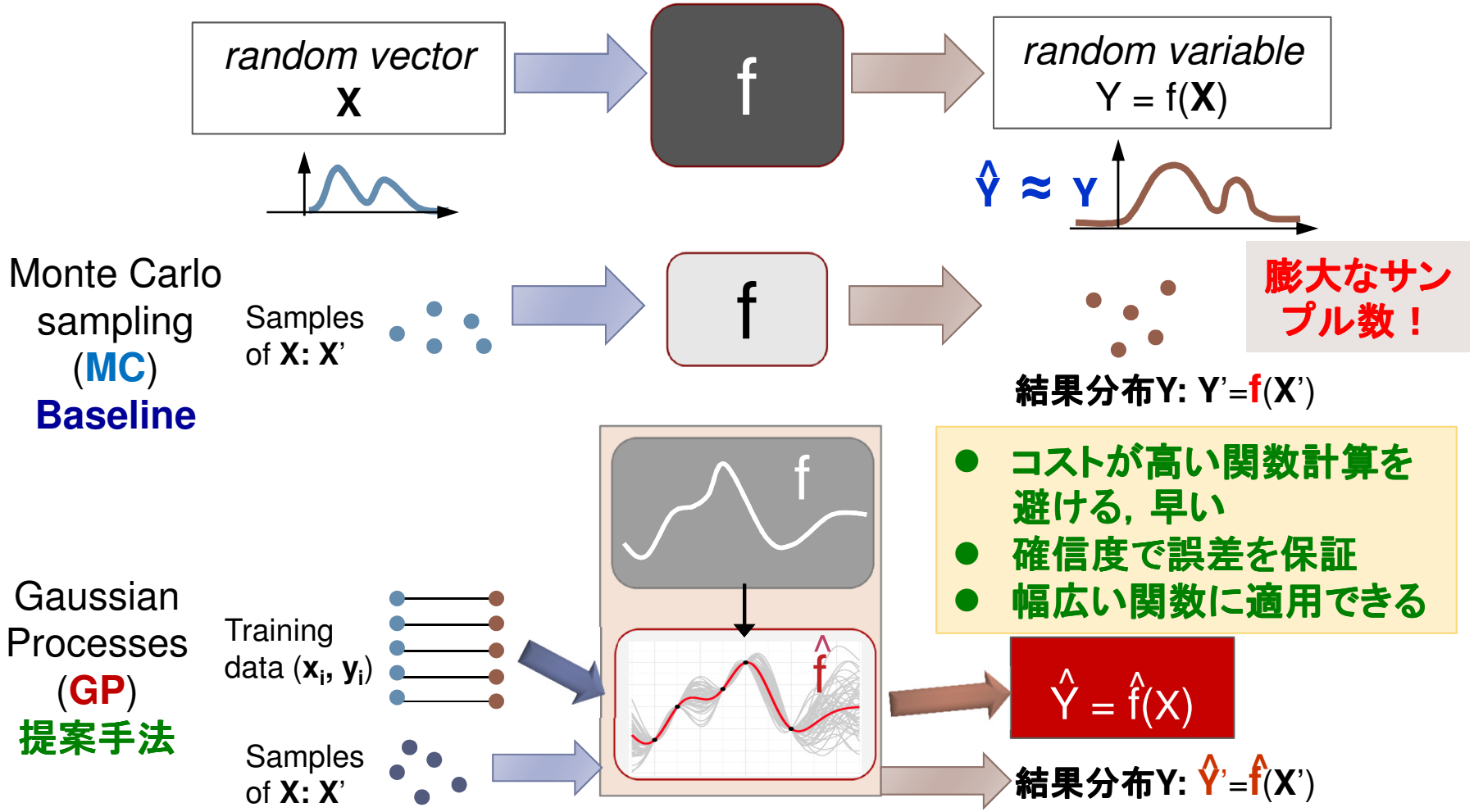
- Thanh Tran, Yanlei Diao (UMass Amherst),
  Charles Sutton (U. Edinburgh), Anna Liu (UMass Amherst)
- User-Defined Functions (UDFs)
  - 科学計算や金融分析によく使われている
    - 例：星雲の赤方偏移(redshift)から年齢(age)を計算
    - 特徴：外部コード，**black boxes**；計算コストが**高い**
  - 不確実データにおけるUDFs: 計算結果も不確実
  - 既存手法の問題点
    - 不確実性は定量化されない：**確信スコアがない**
    - 科学者にとって重荷となる：**人工的にコード**し，不確実さを計算



- Contribution
  - 結果分布の近似を計算し，エラー限度(error bound)を算出
    - ユーザ指定のエラー閾値に対応できる
  - 最適化されたオンラインアルゴリズム

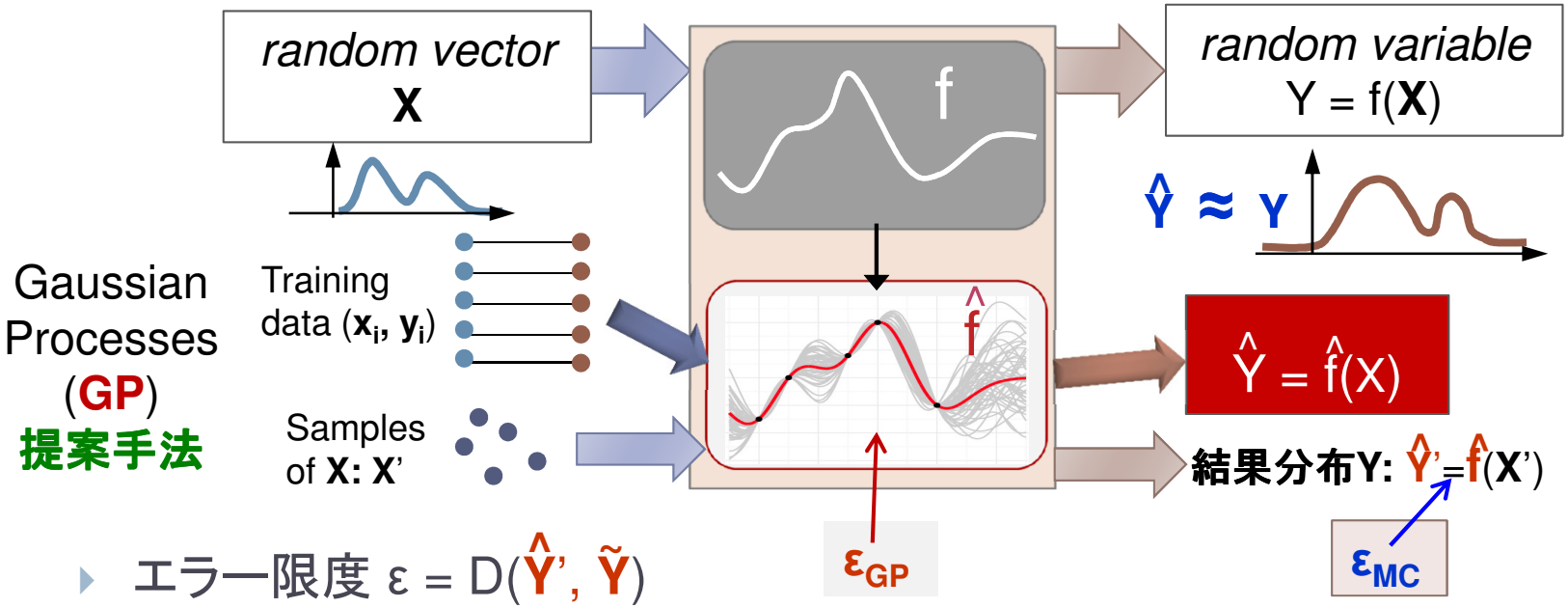
Supporting User-Defined Functions on Uncertain Data

- UDFsに対する計算方法: **MC** vs. **GP**



Supporting User-Defined Functions on Uncertain Data

- ガウス過程(**GP**)で結果分布を計算



- エラー限度 ε = D($\hat{Y}'$, $\tilde{Y}$)
  - $\tilde{Y}=\tilde{f}(X)$, $\tilde{f}$: GPから生成した関数群のランダムなサンプル関数
  - D: discrepancy measure, Kolmogorov-Smirnov(KS) measure
  - 2種類のエラーを結合 : D($\hat{Y}'$,$\tilde{Y}$) <= D($\hat{Y}'$,$\tilde{Y}'$) + D($\tilde{Y}'$,$\tilde{Y}$) = ε_{GP} + ε_{MC}
    - **GP** modeling error
    - **MC** sampling error

Supporting User-Defined Functions on Uncertain Data

- 最適化オンラインアルゴリズム（訓練しながら推定する）
  - ユーザ指定のエラー閾値に満たすまでに，インクリメント的にGPモデルの更新を反復
    - エラー閾値の条件判定に部分訓練データのみを使う
  - オンラインチューニング(online tuning)
    - エラー閾値に満たすための訓練データの数を計算し，インクリメント的に訓練データを追加
  - オンライン再訓練(online retraining)
    - 訓練データを追加する時に，変化が閾値以上のパラメータを学習し，GPモデルを更新する
  - ローカル推定(local inference)
    - 訓練データの部分集合を使うことで推定コストを減らす