逐次データ同化入門

樋口知之　（情報・システム研究機構　統計数理研究所）

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

大学共同利用機関法人と大学共同利用機関

文部科学省の国立研究所

全国に17設置

Universities/College

国立大学
83

私立大学
661

私立短大
468

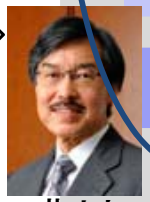
Inter-University Research Institutes

自然科学研究機構 （国立天文台、…）

高エネルギー加速器研究機構

人間文化研究機構

情報・システム研究機構



北川

統計数理研究所 (ISM) *1944年設立*

国立情報学研究所 (NII)

国立遺伝学研究所

国立極地研究所

樋口

喜連川

大学セクター

Bottom Up

国立研究開発法人 Top Down

理研、JAMSTEC, NIMS

アウトライン

1. ベイズ統計の基礎
2. 状態空間モデルと逐次ベイズ
3. 逐次データ同化
4. 統計的推測
5. 実験計画

専門分野
質問1： 地球物理 or それ以外
質問2： 情報・数理 or それ以外

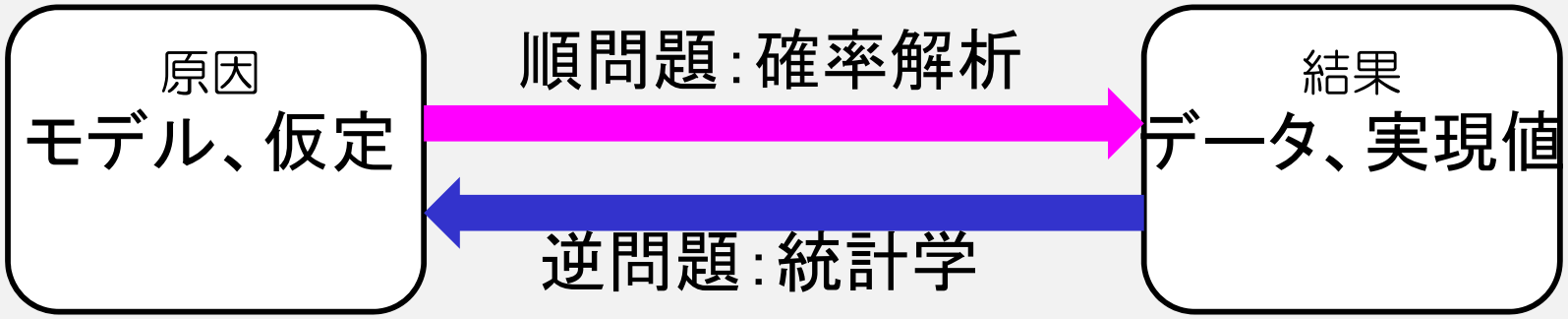
アウトライン

1. ベイズ統計の基礎
2. 状態空間モデルと逐次ベイズ
3. 逐次データ同化
4. 統計的推測
5. 実験計画

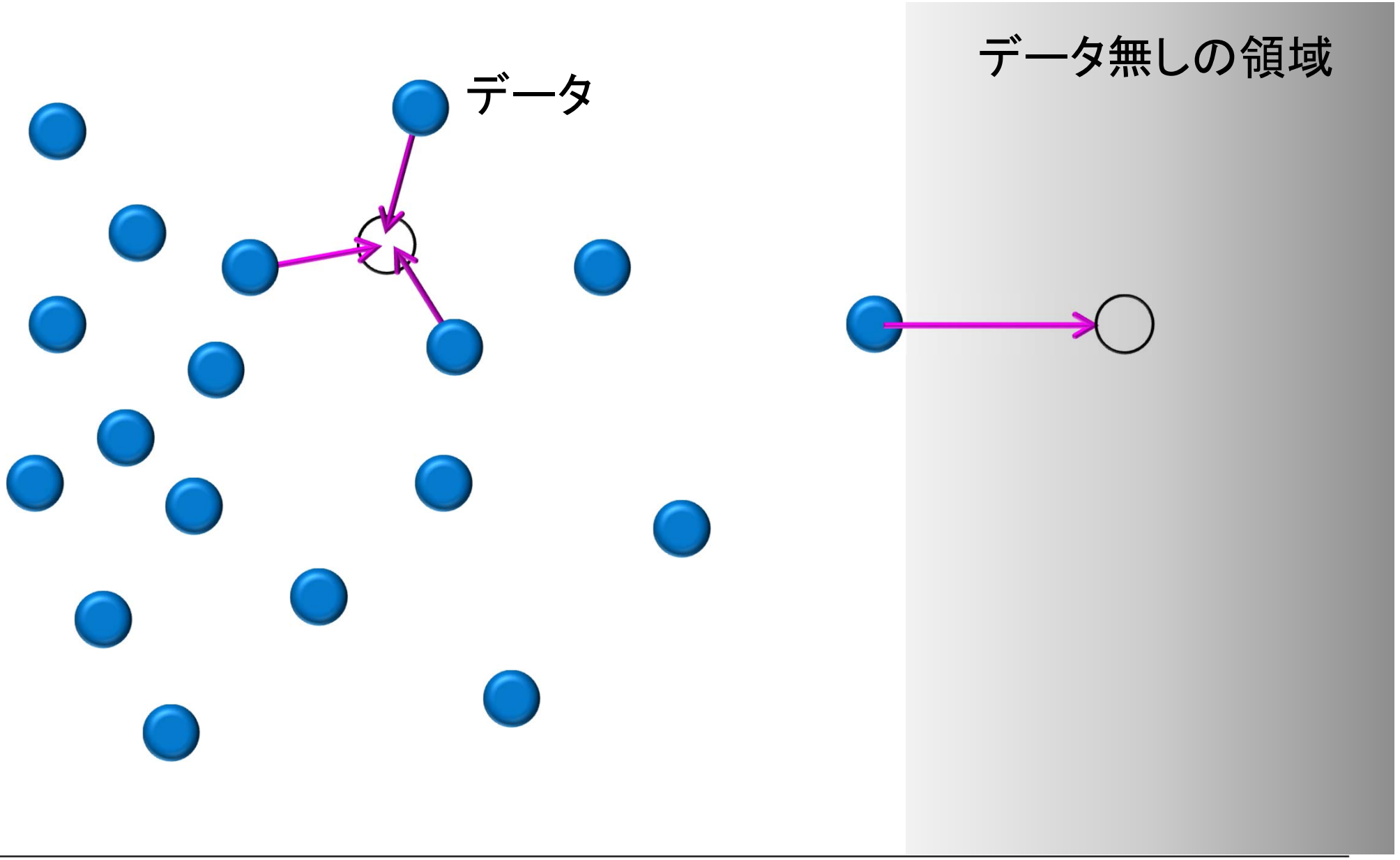
さいころ：確率と統計学

$$p(X=1)=\frac{1}{6}$$

5, 4, 1, 3, 3,…



内挿と外挿問題



同時分布と条件付き分布

$$p(\mathrm{A}) \equiv \text{Probability of } \mathbf{A}$$

$$p(\mathrm{A} \mid \mathrm{B}) \equiv \text{Probability of } \mathbf{A} \text{ given } \mathbf{B} \leftarrow \text{Conditional Probability}$$

$$p(\mathrm{A}, \mathrm{B}) \equiv \text{Probability of } \mathbf{A} \text{ and } \mathbf{B} \leftarrow \text{Joint Probability}$$

$$p(\mathrm{A}=1) = \frac{30}{100}, \quad p(\mathrm{A}=1, \mathrm{B}=1) = \frac{10}{100}, \quad p(\mathrm{B}=1 \mid \mathrm{A}=1) = \frac{10}{30} = \frac{10}{(20+10)}$$

Total: 100

# of コーヒー豆を買った人 : 30

# of ミルクを買った人: 60

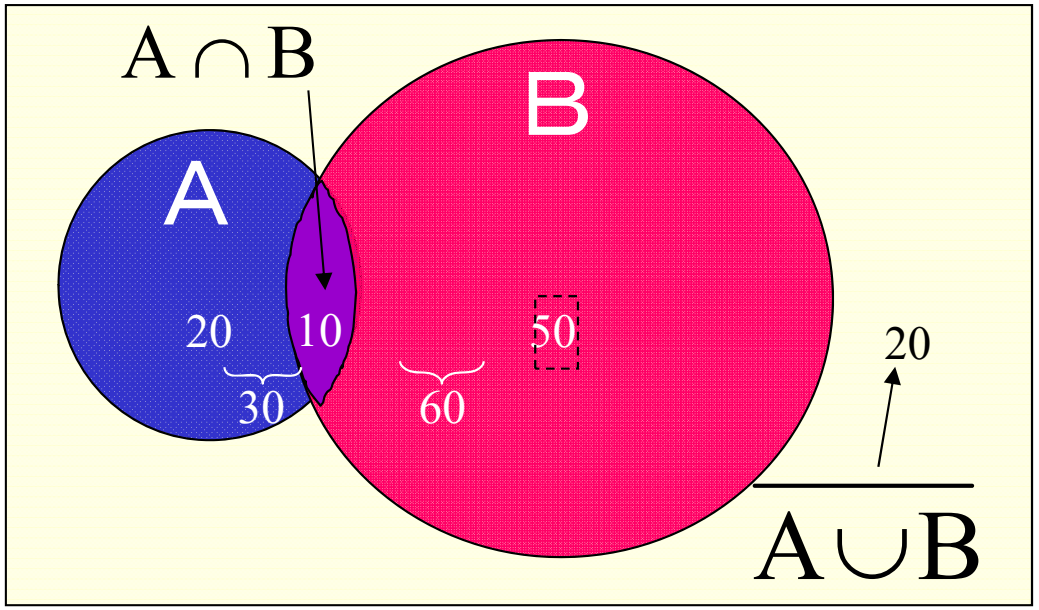
# of コーヒー豆とミルクをかった人: 10

A=1: コーヒー豆をかった

=0: 買わなかった

B=1: ミルクを買った

=0: 買わなかった



周辺化

$$p(\mathrm{A}=3)=\sum_{j=1}^{6} p(\mathrm{A}=3,\mathrm{B}=j)$$

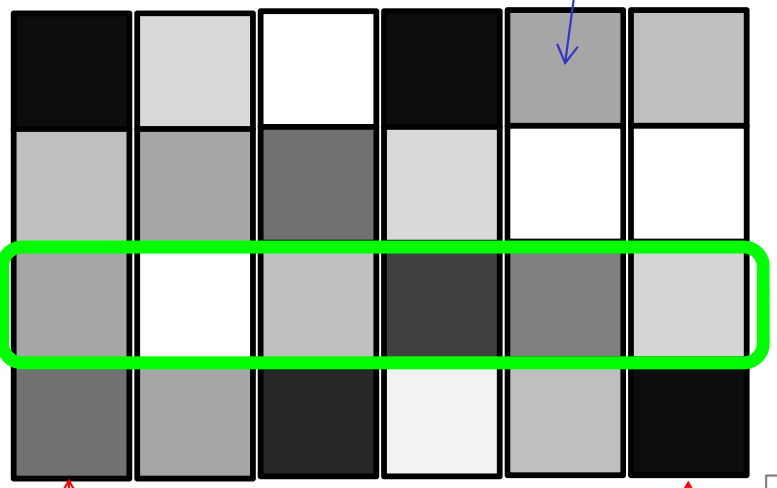
同時分布 $p(\mathrm{A}=i,\mathrm{B}=j)$

A=1: Yahoo
=2: Google
=3: NTTグループ
=4: Others

A

A=3

Web Search Engine



B=1 ..... B ..... B=6

携帯電話製造メーカ

B=1: Apple, =2: Kyocera, =3: Sharp
=4: SONY, =5: Fujitsu , =6: Others

確率分布と最適化関数の違い

$$\sum_{i=1}^{4}\sum_{j=1}^{6} p(\mathrm{A}=i,\mathrm{B}=j)=1$$

条件つき確率の意味

$$\sum_{j=1}^{6} p(\mathrm{A}=i\,|\,\mathrm{B}=j)\neq 1$$

離散	連続
$p(\mathrm{A}=\mathrm{A}_i)=\sum_{j\in \text{possible } \mathrm{B}_{j'}} p(\mathrm{A}_i,\mathrm{B}=\mathrm{B}_j)$	$\Rightarrow \quad p(\mathrm{A})=\int p(\mathrm{A},\mathrm{B})d\mathrm{B}$

頻出する式変形

（1）ベイズの定理

$$p(\mathrm{B}\,|\,\mathrm{A},*)=\frac{p(\mathrm{A},\mathrm{B}\,|\,*)}{p(\mathrm{A}\,|\,*)}=\frac{p(\mathrm{A}\,|\,\mathrm{B},*)\cdot p(\mathrm{B}\,|\,*)}{p(\mathrm{A}\,|\,*)}$$

（2）周辺化

$$p(\mathrm{A}\,|\,*)=\int p(\mathrm{A},\mathrm{B}\,|\,*)d\mathrm{B}$$

同時分布の分解

$$p(y_1,\ldots,y_T,x_1,\ldots,x_T)$$

$$p(y_1,\ldots,y_T,x_1,\ldots,x_T)$$

$$= p(y_{1:T-1},y_T,x_{1:T})$$

$$= p(y_T|y_{1:T-1},x_{1:T})\cdot p(y_{1:T-1},x_{1:T})$$

$$= p(y_T|y_{1:T-1},x_{1:T})\cdot p(x_T|y_{1:T-1},x_{1:T-1})\,p(y_{1:T-1},x_{1:T-1})$$

$$= \prod_{t=1}^{T} p(y_t|y_{1:t-1},x_{1:t})\cdot p(x_t|y_{1:t-1},x_{1:t-1})$$

逐次ベイズでの表記法

$$\boldsymbol{x}_{1:T} \equiv \{\boldsymbol{x}_1,\ldots \boldsymbol{x}_T\}$$

$$\boldsymbol{y}_{1:T} \equiv \{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T\}$$

ただし、$y_{1:0}=\phi,\quad x_{1:0}=\phi$ とする

ベイズの定理がなぜ今役立つのか？４つの理由

イギリスの牧師・数学者（1702 – 1761年）
1763年に発見

x：興味のある対象

y：データ

2．対象の特徴をとらえるセンサー性能の向上
高精度センサーのコモディティ（日用品）化

4．高速（無線）インターネット網の整備

ベイズの反転公式

$$p(x \mid y) = \frac{p(y \mid x)\,p(x)}{\sum p(y \mid x)\,p(x)}$$

1．膨大な数の積分（和）操作には高速な計算機が必要
コンピュータの性能向上

3．対象の細かい情報を不確実性を含めて数値化。個人の情報を網羅的に収集
ストレージの廉価化

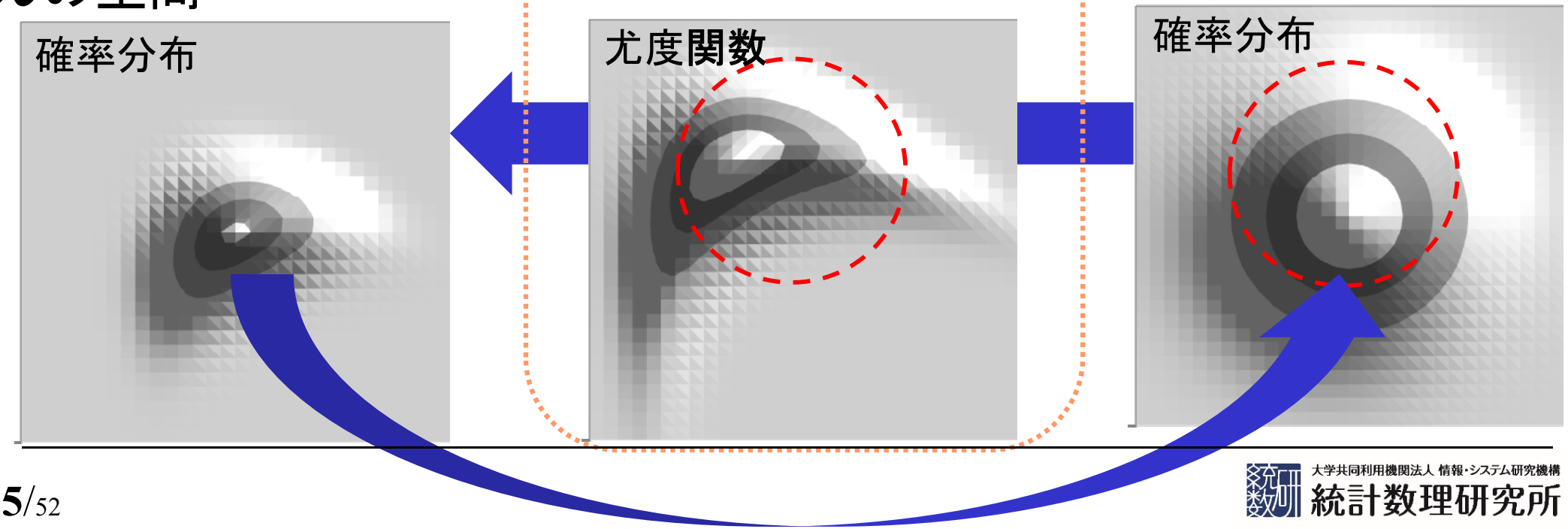
逆推論の実験

①ベイズの情報循環

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x}) \cdot p(\boldsymbol{x})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \boldsymbol{x}) \cdot p(\boldsymbol{x})$$

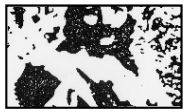
Posterior	Likelihood	Prior
Improved knowledge about values of $\boldsymbol{x}$	Feasibility of realization of $\boldsymbol{y}$ for given $\boldsymbol{x}$	Belief about values of $\boldsymbol{x}$

$\boldsymbol{x}$の空間



統計学が支える逆推論：②ベイズの反転公式

x：興味のある対象



y：データ

順解析：シミュレーション等

ベイズの反転公式

$$\underbrace{p(x \mid y)}_{\text{逆解析}} = \frac{\overbrace{p(y \mid x)}^{\text{結果}\ \ \text{原因}}\, p(x)}{\sum p(y \mid x)\, p(x)}$$

ベイズの定理。等号の右側と左側で，赤と青で示した変数部分の縦棒との相対関係が反転していることがわかる。この事実により，ベイズの反転公式と呼ばれる。

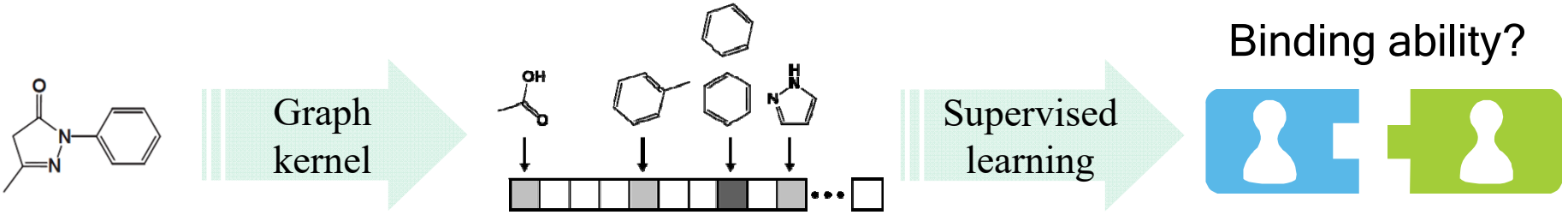
（Thanks to Prof. Yoshida; 吉田@統数研から借用、改変）

逆問題はむずかしい

■ Virtual screening, QSAR modeling: P(Y|G)　順問題

Quantitative Structure-Activity(Affinity) Relationship

Develop statistical models to predict biochemical or physiochemical activities Y of an input chemical structure G



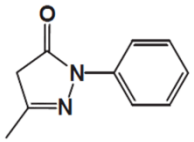
■ Chemical design, Inverse-QSAR: P(G|Y=y)　逆問題

Generate novel chemical structures G achieving desired activities Y=y

Preimage reconstruction of the graph kernel using a MCMC algorithm

非効率（現実には機能しない）探索法

構造　G



$f(G)$　物理あるいは化学的指数や係数、特性の計算値

機能発現　y　その（実験値あるいは経験値）データ

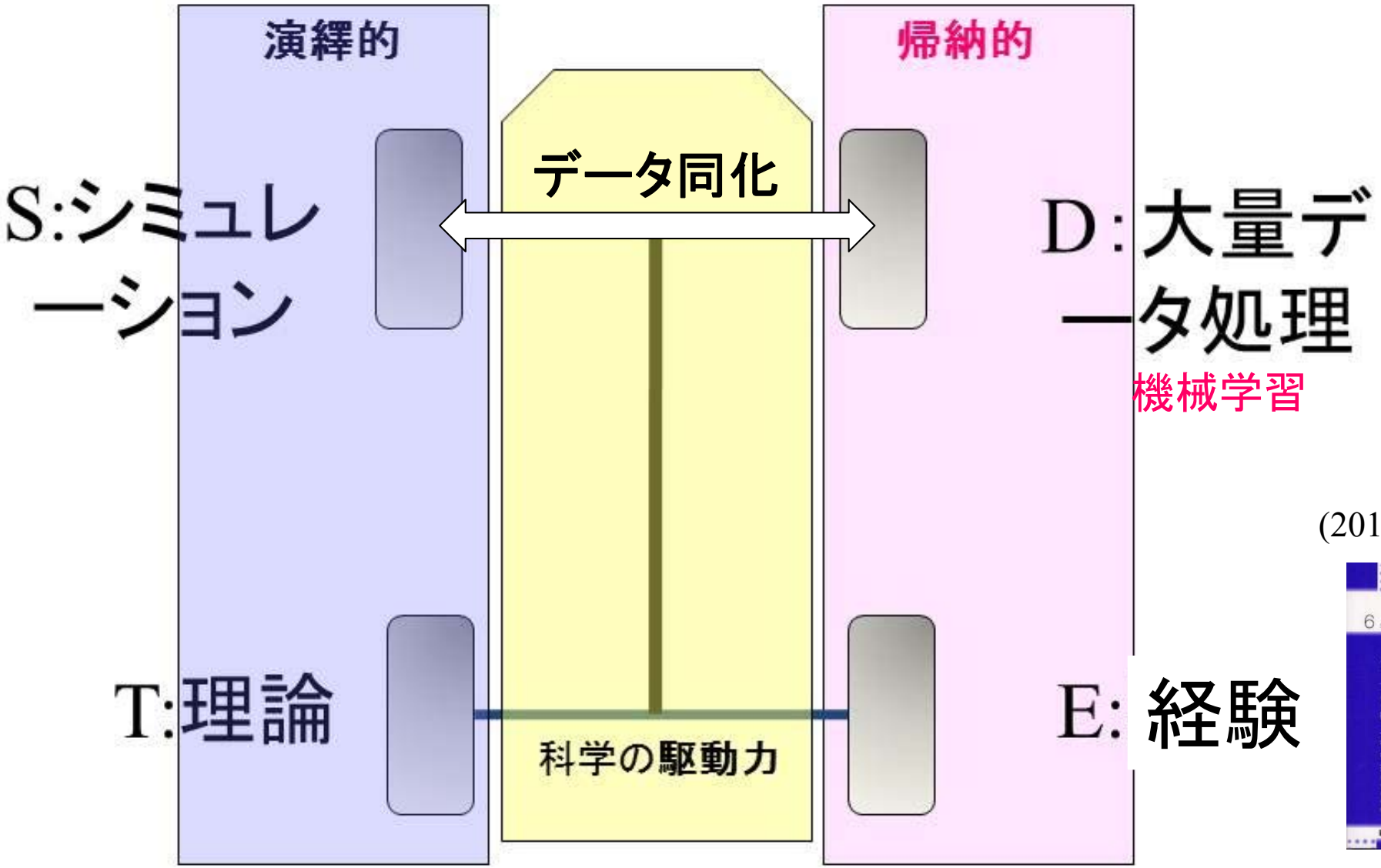
膨大な数のモンテカルロ計算

Gの候補を恣意的に多数発生

一つ一つの G に対して順解析（MDなどの第一原理計算）を実施。データとの整合性 $p(y \mid f(G))$ を計算

エキスパートが目と勘で判断

つなぐ：データ同化

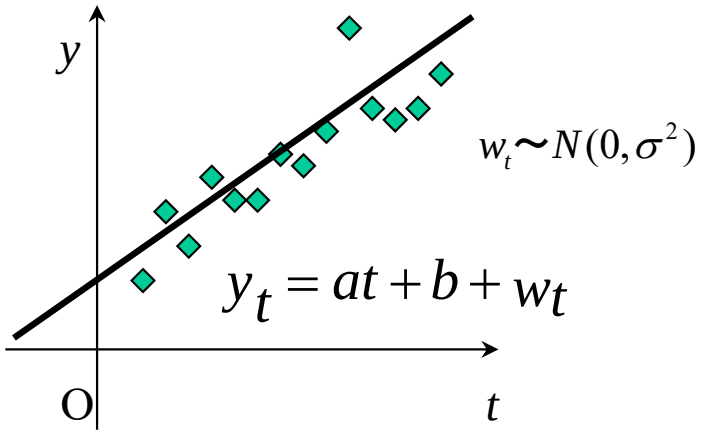


(2011年9月刊行)

データ同化とは？

計算手続き的に見れば、
計測データにモデルをあてはめること

$$p(y_t \mid \theta = [a, b, \sigma^2])$$



データに直線をあてはめる

データにシミュレーションモデルをあてはめる

＝「シミュレーションモデルにデータを同化する」

データ同化

高次元の複雑なモデルを扱うことになるのが特徴

アウトライン

1. ベイズ統計の基礎
2. 状態空間モデルと逐次ベイズ
3. 逐次データ同化
4. 統計的推測
5. 実験計画

シミュレーションモデルの構成（1）

（Tsunami simulation model in Japan Sea）

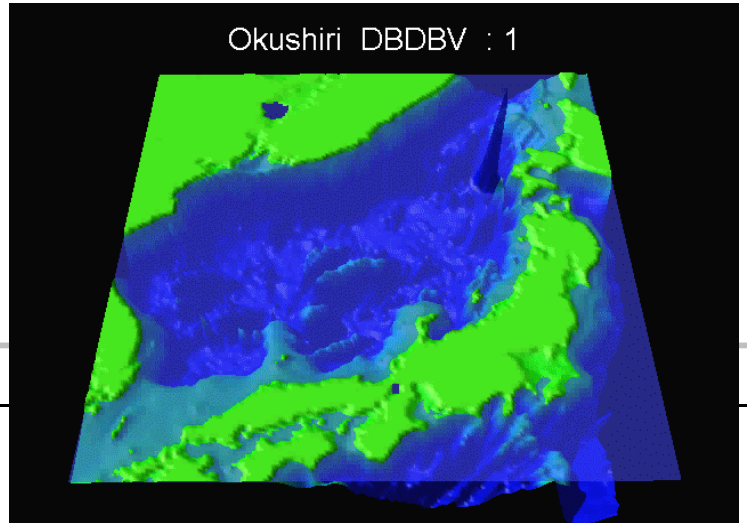
PDE to approximate real physical system
(continuous time/space)

$$\frac{\partial x}{\partial t} = cx^2 + \cdots$$

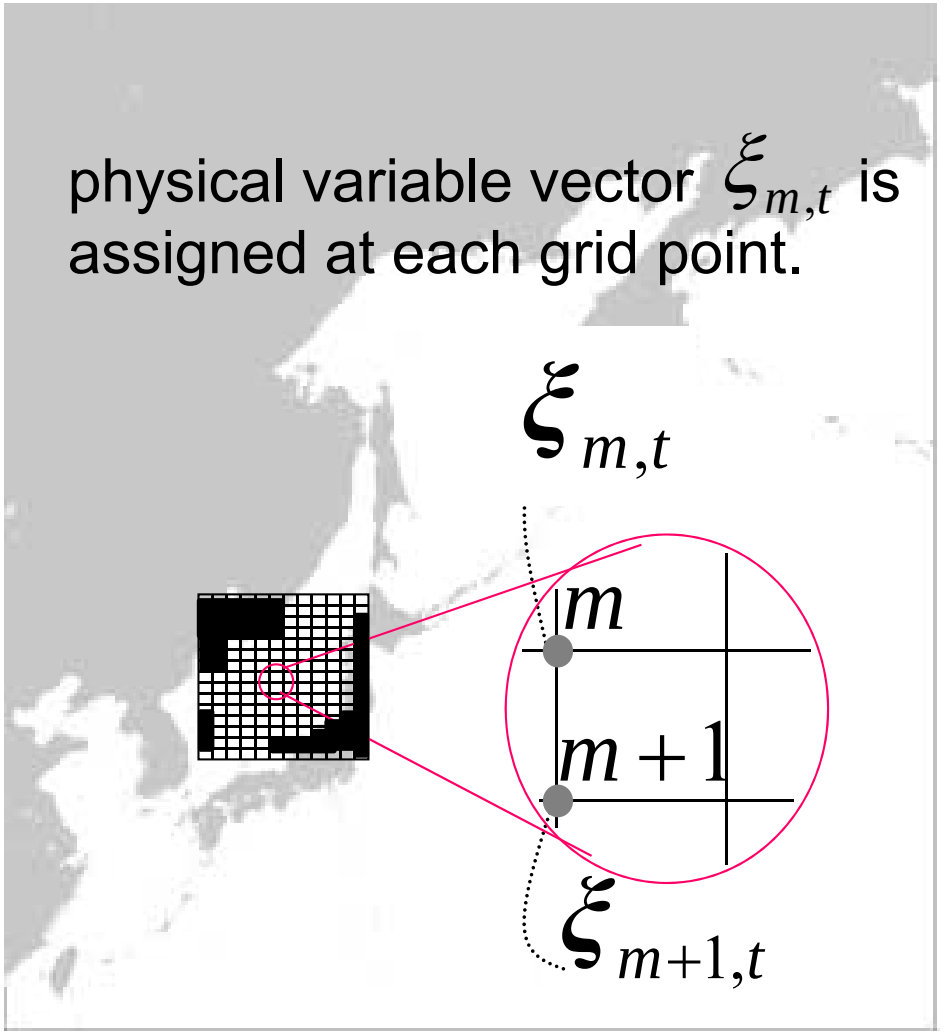
PDE : Partial differential equation

Discrete simulation model
(discrete time/space, FDE)

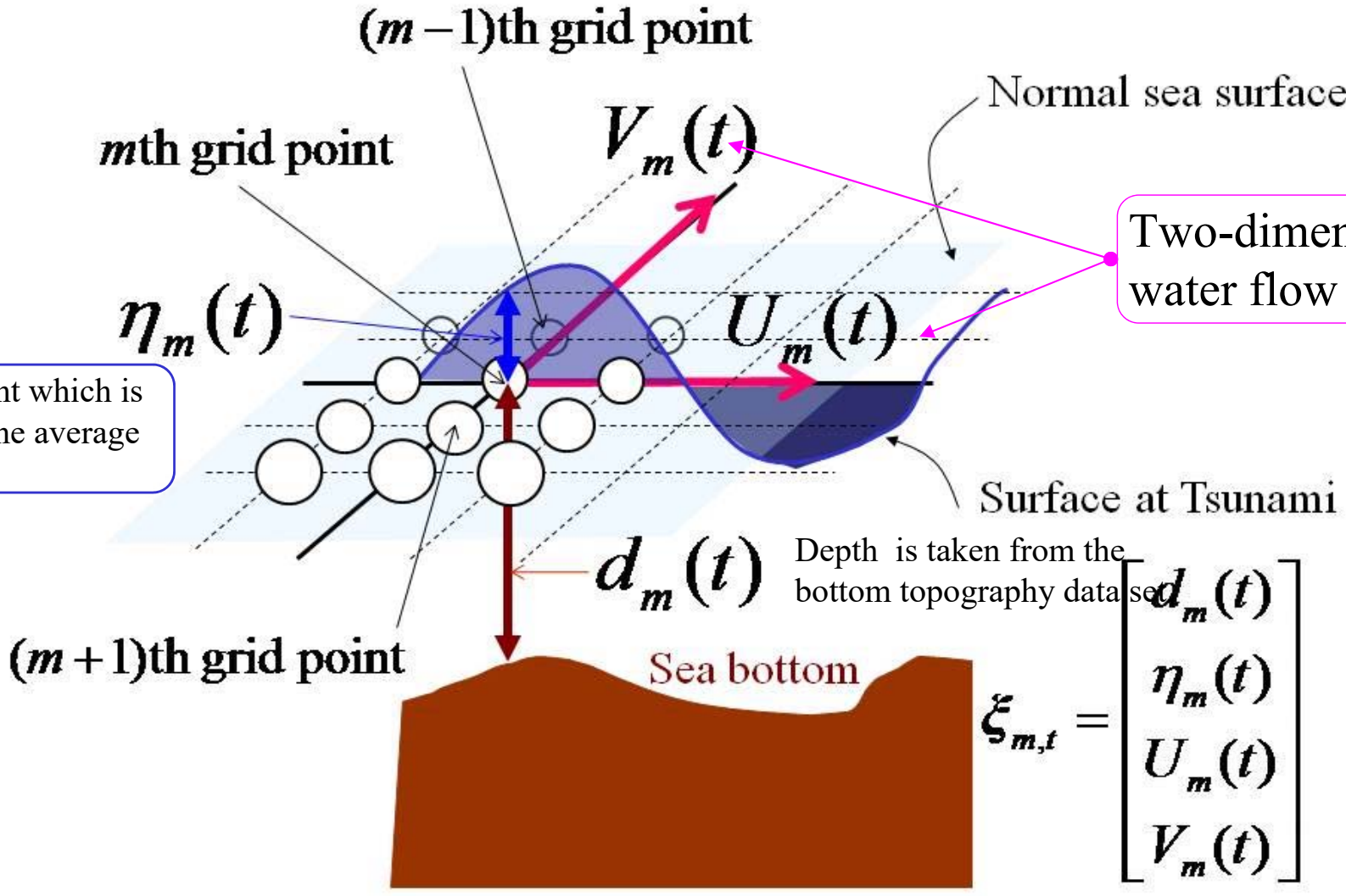
Suppose a case where we conduct a two-dimensional simulation experiment for understanding the flow of shallow water such as the tsunami.



physical variable vector $\xi_{m,t}$ is assigned at each grid point.



シミュレーションモデルの構成 (2)



In fact, this data set is known to be erroneous.

システムモデルとしてのシミュレーションモデル

（simplified meteorological model around Japan）

PDE to approximate real physical system
(continuous time/space)

$\frac{\partial x}{\partial t} = cx^2 + \cdots$

PDE : Partial differential equation

Discrete simulation model
(discrete time/space, FDE)

$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1})$

代入計算

(time varying)

Boundary conditions $\boldsymbol{v}_t$

Nonlinear state space model
(discrete time/space, stochastic(SDE))

$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t)$

State Vector

$$\boldsymbol{x}_t = \begin{bmatrix} \xi_{1,t} \\ \vdots \\ \xi_{m,t} \\ \xi_{m+1,t} \\ \vdots \\ \xi_{M,t} \\ \boldsymbol{\theta} \end{bmatrix}$$

データ同化と一般状態空間モデル

状態ベクトル（Simulation variables）

システムモデル
Stochastic simulation model

Markov性（1）

$$\begin{cases} \boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t), \quad \boldsymbol{v}_t \sim p(\boldsymbol{v} \mid \boldsymbol{\theta}_{\mathrm{sys}}) \\ \boldsymbol{y}_t = h_t(\boldsymbol{x}_t, \boldsymbol{w}_t), \quad \boldsymbol{w}_t \sim p(\boldsymbol{w} \mid \boldsymbol{\theta}_{\mathrm{obs}}) \end{cases}$$

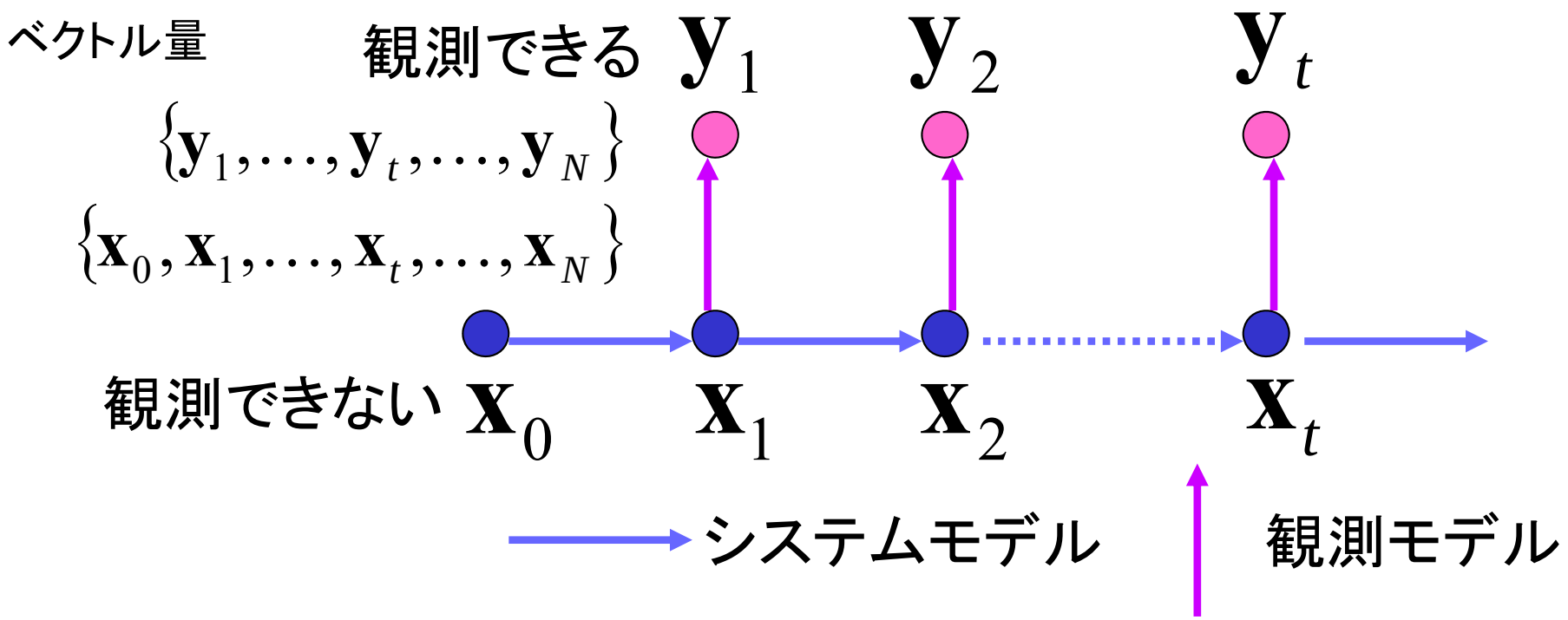
Markov性（2）

Observation model
Measurement model
観測モデル

気象・海洋のデータ同化の枠組み

$$\boldsymbol{y}_t = H_t \boldsymbol{x}_t + \boldsymbol{w}_t, \boldsymbol{w}_t \sim N(0, R_{\mathrm{obs}})$$

Chain Structure Graphical Model



$$p(\boldsymbol{y}_{1:T-1}, \boldsymbol{y}_T, \boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t | \boldsymbol{y}_{1:t-1}, \boldsymbol{x}_{1:t}) \cdot p(\boldsymbol{x}_t | \boldsymbol{y}_{1:t-1}, \boldsymbol{x}_{1:t-1})$$

$$\Rightarrow \prod_{t=1}^{T} p(\boldsymbol{y}_t | \boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t | \boldsymbol{x}_{t-1})$$

逐次ベイズ計算 --- 日次株価データを考えると ---

条件付き分布

予測分布 predictive density: $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})$ きのうまでのデータに基づく今日の状態

フィルタ分布 filter density: $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t})$ 今日までのデータに基づく今日の状態

平滑化分布 smoother density: $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:T})$ 数年後，データをすべて得たもとで振り返った今日の状態

$p(x_j \mid y_{1:i})$

j

i

時刻Tまでのデータをまとめたベクトル系列

$\boldsymbol{y}_{1:T} \equiv \{\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T\}$

prediction

$p(x_{t-1} \mid y_{1:t-1}) \rightarrow p(x_t \mid y_{1:t-1})$

filtering

$p(x_t \mid y_{1:t}) \rightarrow p(x_{t+1} \mid y_{1:t})$

$p(x_{t+1} \mid y_{1:t+1})$

３つの条件付分布と３つの操作

smoothing

$p(x_t \mid y_{1:T}) \leftarrow p(x_{t+1} \mid y_{1:T}) \leftarrow p(x_T \mid y_{1:T})$

Prediction

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})$$

$$= \int p(\boldsymbol{x}_t, \boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}) d\boldsymbol{x}_{t-1}$$

$$= \int p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:t-1}) p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}) d\boldsymbol{x}_{t-1}$$

$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{y}_{1:t-1}) = p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ Markov性（1）

$$= \int p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) p(\boldsymbol{x}_{t-1} \mid \boldsymbol{y}_{1:t-1}) d\boldsymbol{x}_{t-1}$$

時刻 *t-1* でのフィルタ分布

filtering

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t})$$

Posterior, Belief

$$= p(\boldsymbol{x}_t \mid \boldsymbol{y}_t, \boldsymbol{y}_{1:t-1})$$

$$= \frac{p(\boldsymbol{x}_t, \boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1})}$$

$$= \frac{p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{y}_{1:t-1}) \cdot p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1})}$$

$$= \frac{p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})}{p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1})}$$

$$= \frac{p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})}{\int p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) \cdot p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}) d\boldsymbol{x}_t}$$

Markov性（2）

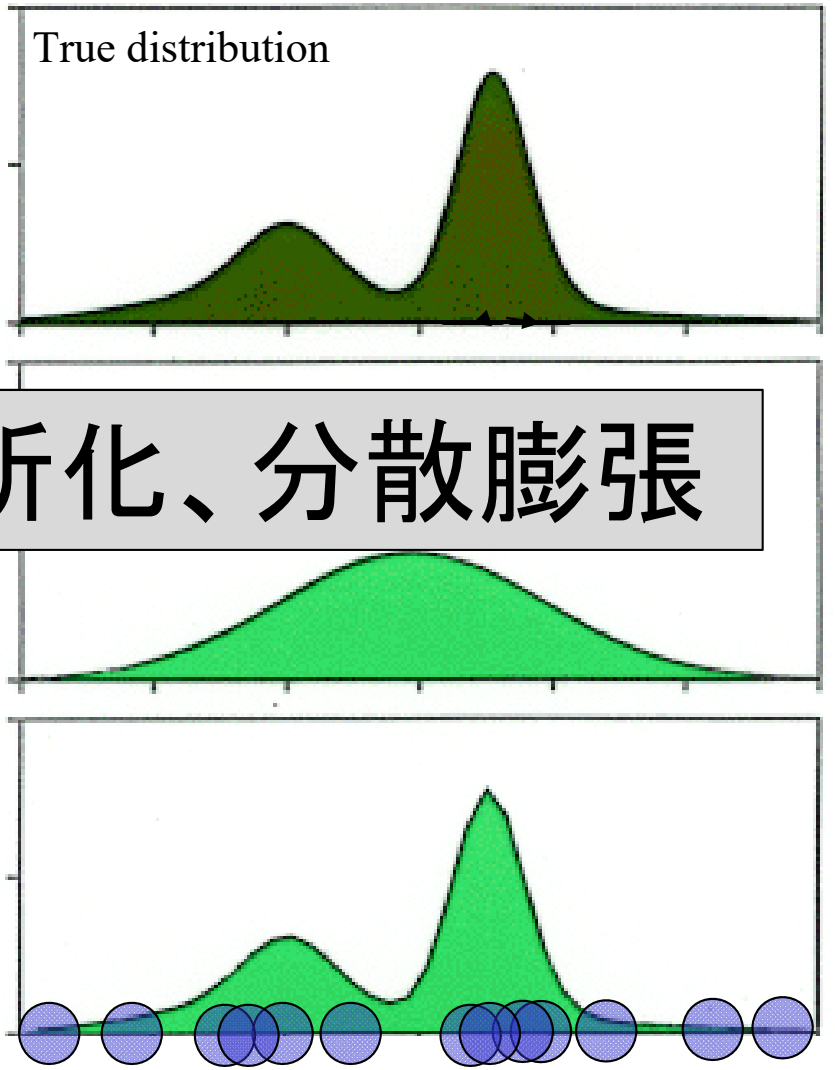
$$p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{y}_{1:t-1}) = p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$$

分布の表現：モンテカルロ近似（表現）

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}),\ p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t}),\ p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:T})$$

モンテカルロ近似

実現値の集合で分布を表現する



局所化、分散膨張

データ同化の場合、分布は本質的に非ガウス。だが、……

N: 粒子数

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}) \cong X_{t|t-1} \equiv \left[\boldsymbol{x}_{t|t-1}^{(1)}, \boldsymbol{x}_{t|t-1}^{(2)}, \ldots, \boldsymbol{x}_{t|t-1}^{(N)}\right]$$

$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t}) \cong X_{t|t} \equiv \left[\boldsymbol{x}_{t|t}^{(1)}, \boldsymbol{x}_{t|t}^{(2)}, \ldots, \boldsymbol{x}_{t|t}^{(N)}\right]$$

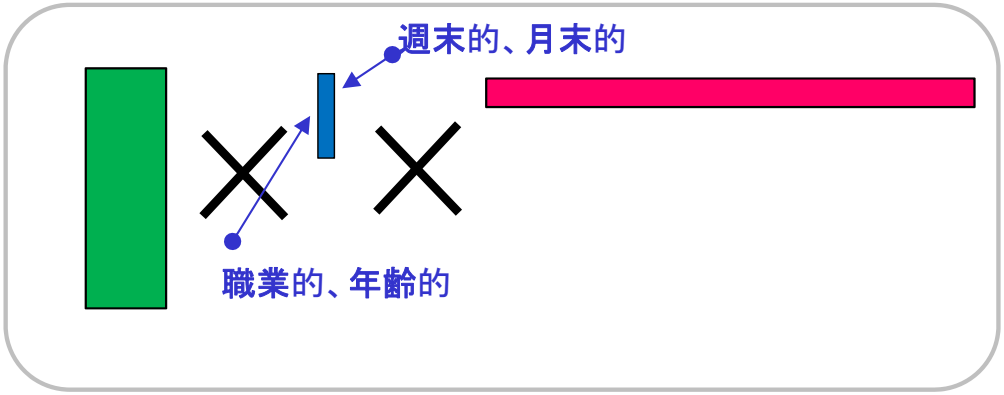
必要なメモリー量の比較

$$\dim(V_{t|t-1} = E\{\boldsymbol{x}_{t|t-1} \cdot \boldsymbol{x}'_{t|t-1}\}) = L \times L \qquad \dim(X_{t|t-1}) = L \times N$$

10^{5+5} 10^{5+2}

巨大次元の行列、テンソル分解計算技術の進化

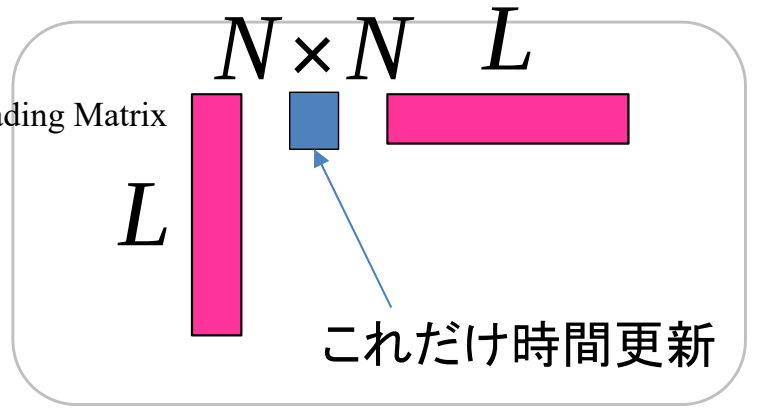
だいたい似かよるように
代数演算のみで分解



$$V_{t|t-1} = E\left\{ \boldsymbol{x}_{t|t-1} \cdot \boldsymbol{x}'_{t|t-1} \right\}$$

$L \times L$

近似



アウトライン

1. ベイズ統計の基礎
2. 状態空間モデルと逐次ベイズ
3. 逐次データ同化
4. 統計的推測
5. 実験計画

逐次データ同化のアルゴリズム

逐次データ同化では観測を得るたびに確率変数　　の分布または値の推定を行う

x_n

---

y_{n-1}

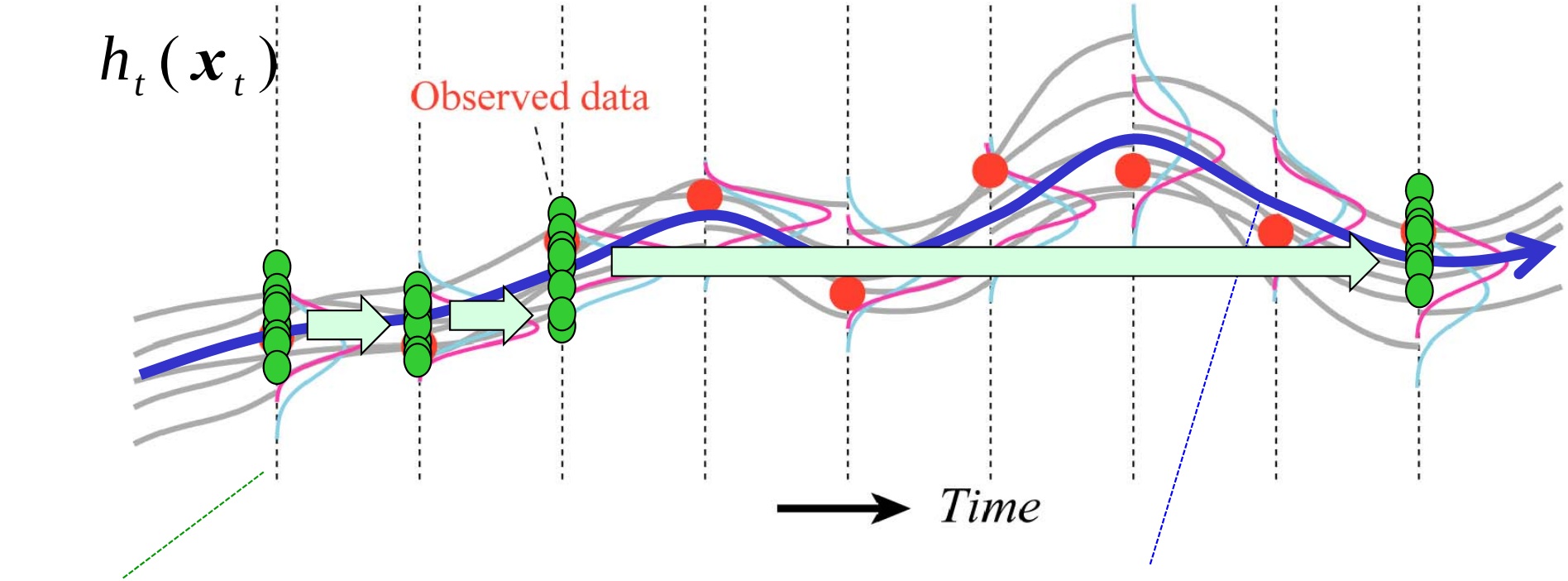
$p(x_{n-1} \mid y_{1:n-1})$

$\left(p(x_i \mid y_{1:k}) = p(x_i \mid y_1, y_2, \cdots, y_k) \right)$

逐次（アンサンブル）vs. 非逐次（最適パス）

データ同化のイメージ

データ空間にシミュレーション解を射影した時



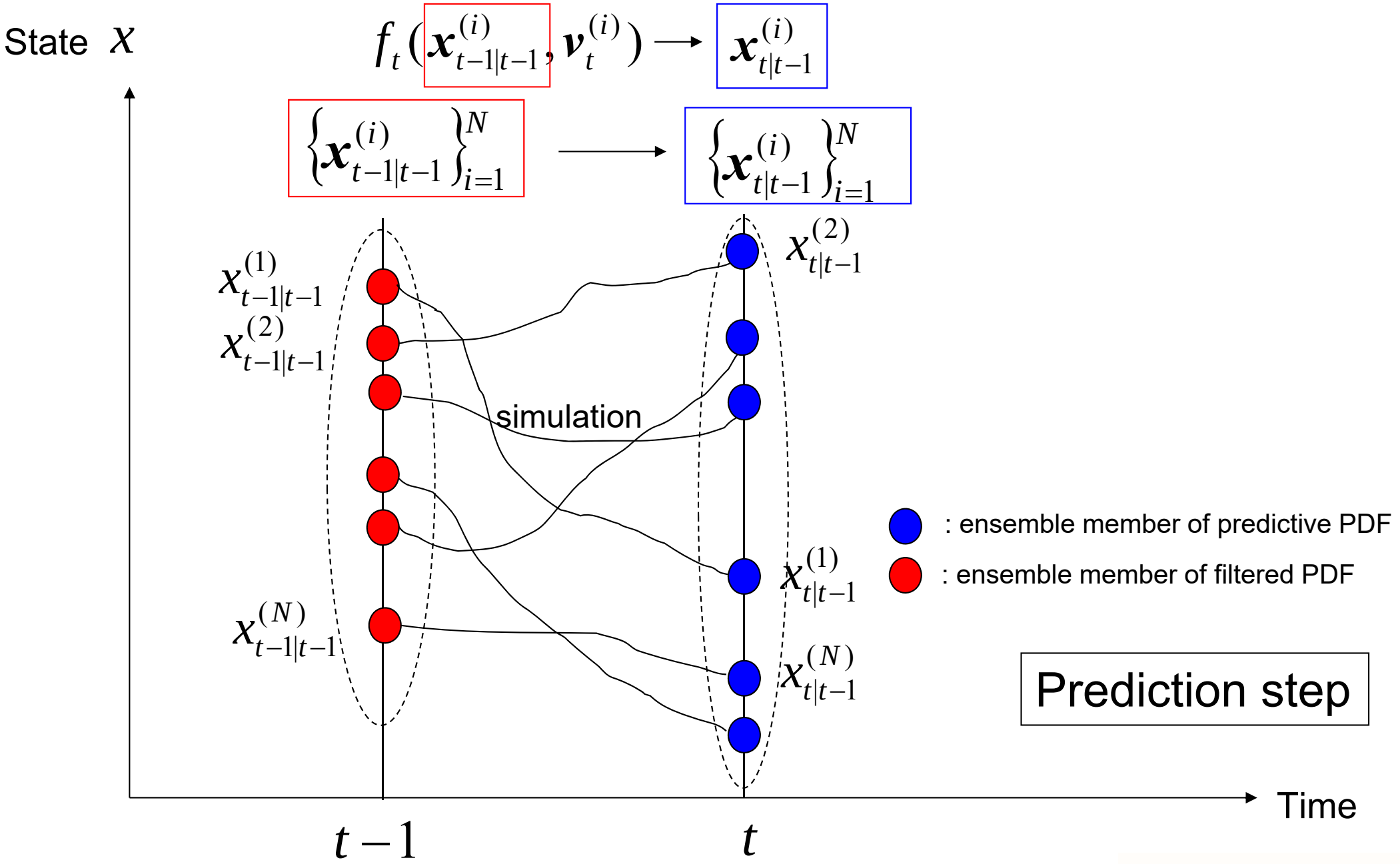
逐次（オンライン）型： 集団の時間発展を追う。つまり、**Swarm Filter**

代表例： EnKF (Ensemble Kalman Filter)

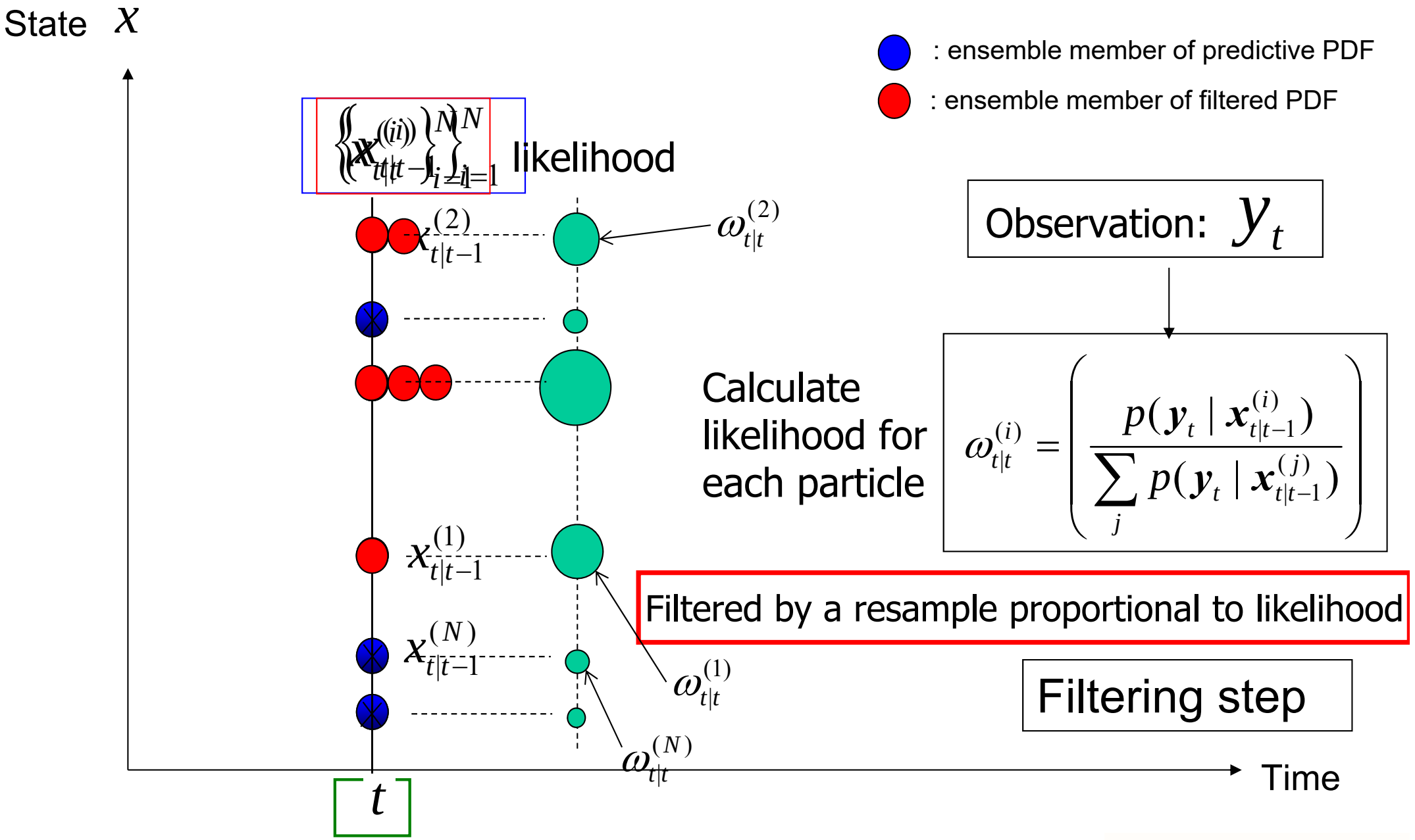
非逐次（オフライン）型：**ベスト初期値をもつパスを求める**

代表例： 4次元変分法（Adjoint法）

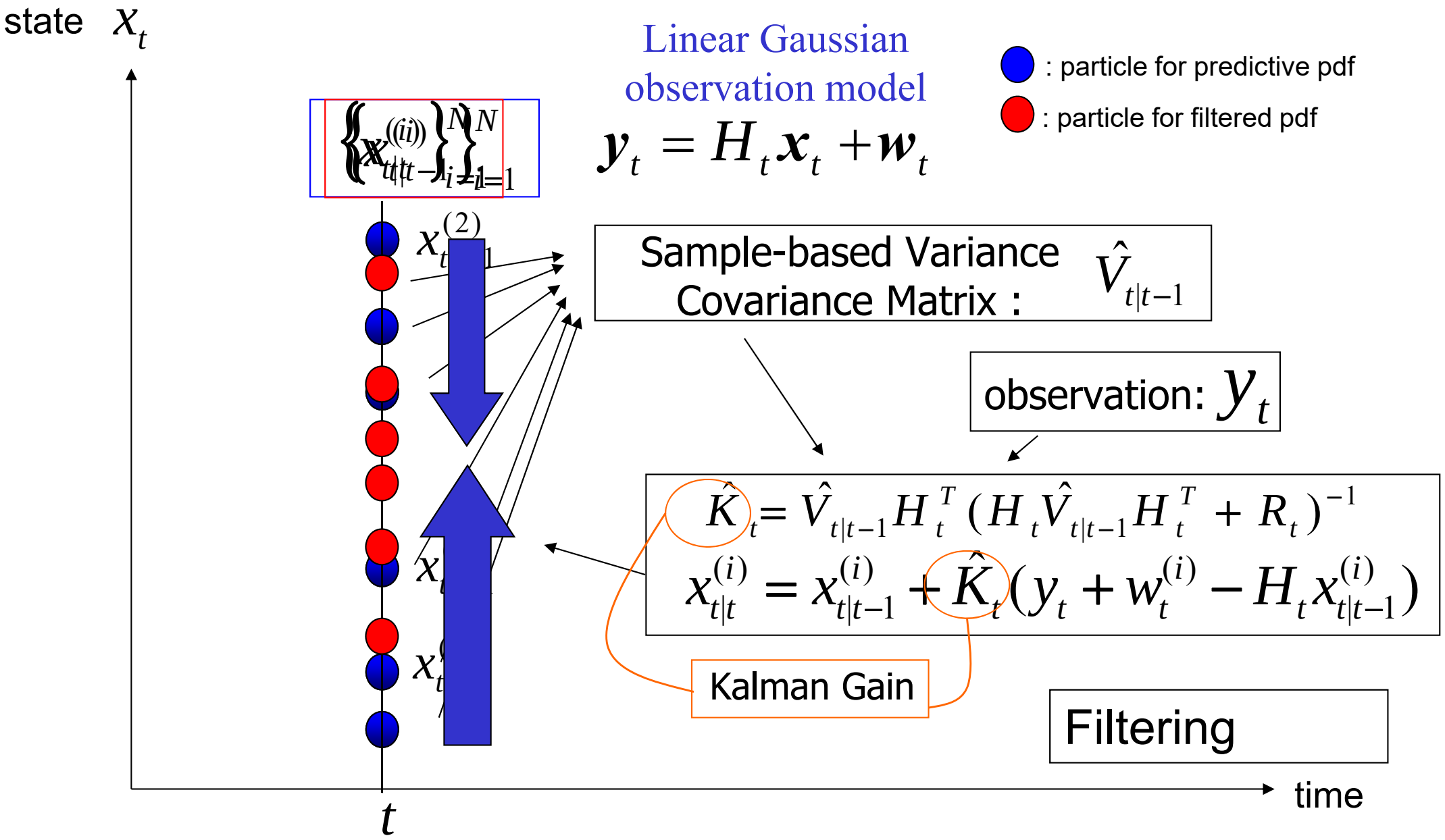
予測のステップ (EnKFとPFで共通)



PFでのフィルタリングのステップ

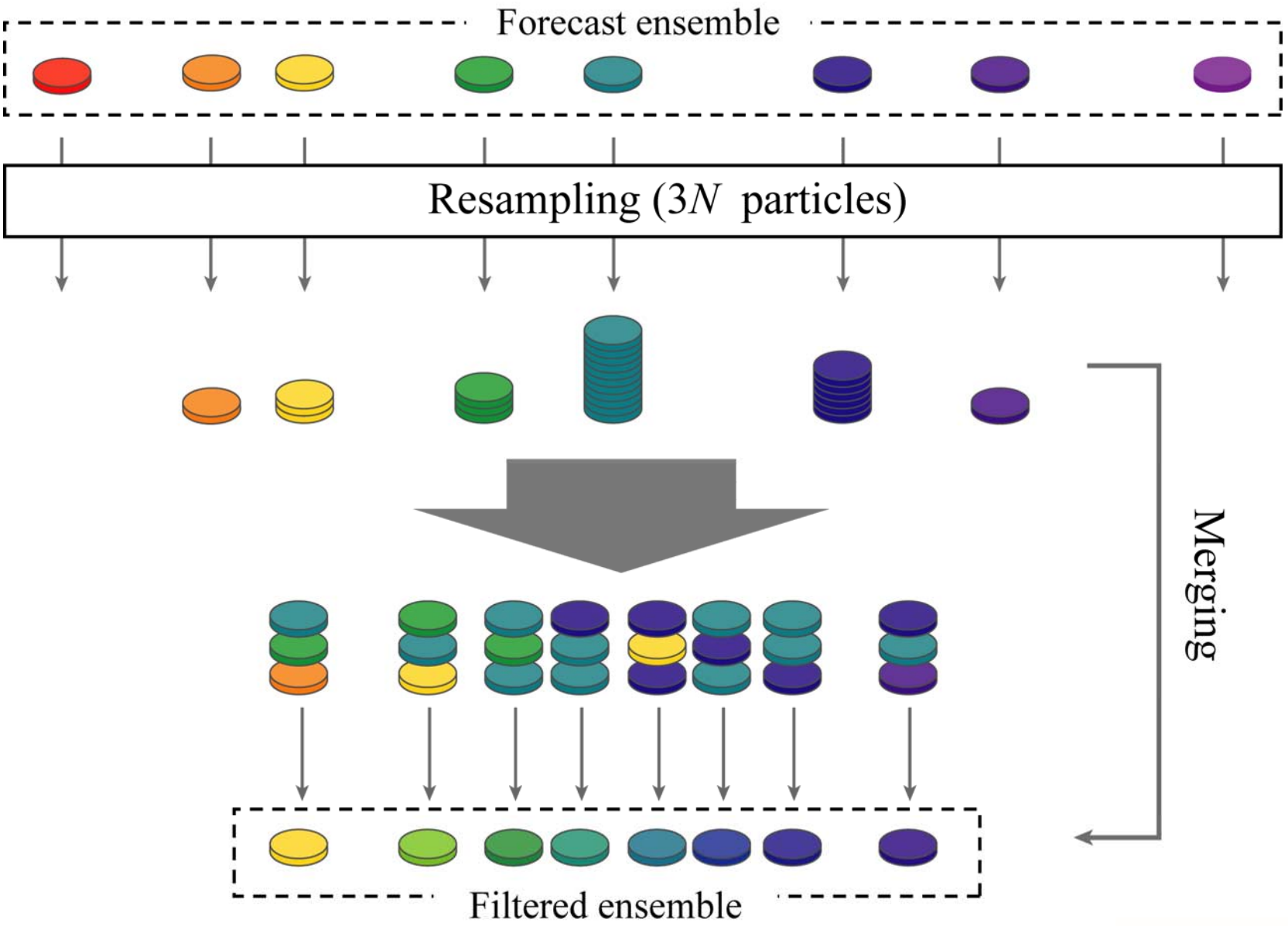


EnKFでのフィルタリングのステップ



Merging particle filter (MPF)
(Nakano et al., 2007)

For a case with n=3



アウトライン

1. ベイズ統計の基礎
2. 状態空間モデルと逐次ベイズ
3. 逐次データ同化
4. **統計的推測**
5. 実験計画

二つのタイプの事後分布

$\boldsymbol{x}_t$：時刻 tの興味のある対象　　$\boldsymbol{x}_{1:T} \equiv \{\boldsymbol{x}_1, \ldots \boldsymbol{x}_T\}$

$\boldsymbol{y}_t$：データ　　$\boldsymbol{y}_{1:T} \equiv \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T\}$

フィルタリング

フィルタ分布 $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:T}) = \dfrac{\overbrace{p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)}^{\text{観測モデル}} \overbrace{p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})}^{\text{予測分布}}}{\sum_{\boldsymbol{x}_t} p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1})}$

データの尤度 $p(\boldsymbol{y}_{1:T} \mid \theta) = \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}, \theta)$ 一期先尤度

パラメータの事後分布

事後分布 $p(\theta \mid \boldsymbol{y}_{1:T}) = \dfrac{p(\boldsymbol{y}_{1:T} \mid \theta) \overbrace{p(\theta)}^{\text{事前分布}}}{\sum_{\theta} p(\boldsymbol{y}_{1:T} \mid \theta) p(\theta)}$

システムノイズがないシミュレーション

予測

システムモデル

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1})$$

システムノイズを入れたくない
（保存則が破れることを避けたい）

フィルタリング

$$p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}) = \int p(\boldsymbol{y}_t \mid \boldsymbol{x}_t) p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t-1}) d\boldsymbol{x}_t = p(\boldsymbol{y}_t \mid \boldsymbol{x}_t)$$

$$p(\boldsymbol{y}_{1:T} \mid \theta) = \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}, \theta)$$

パラメータの事後分布

次に詳説

$$p(\theta \mid \boldsymbol{y}_{1:T}) \propto p(\boldsymbol{y}_{1:T} \mid \theta^{[k]}) p(\theta^{[k]})$$

$$\theta^{[k+1]} \leftarrow \theta^{[k]}$$

Importance Sampling
MCMC, pMCMC

パラメータ推定問題

統計学では、状態ベクトルの次元の低い問題に、$\boldsymbol{x}_t$ を滑らかな解析関数（スプライン関数　$\boldsymbol{x}_t = s(\boldsymbol{x}, t)$ ）で近似表現して解く方法も盛んに研究されている。

単純なモンテカルロ計算および４次元変分法

$\theta^{[k]}\ (k=1,\ldots)$を　$p(\theta)$　から多数サンプルする

$p(\theta)$ が無情報分布の場合がモンテカルロ計算

$for\ \ k=1,\ldots$

$\theta^{[k]} \Rightarrow$ シミュレーションの結果が一つ得られる $\Rightarrow p(\boldsymbol{y}_{1:T}\mid\theta^{[k]})=\prod_{t=1}^{T} p(\boldsymbol{y}_t\mid\boldsymbol{x}_t,\theta^{[k]})$

$p(\theta^{[k]}\mid\boldsymbol{y}_{1:T})\propto p(\boldsymbol{y}_{1:T}\mid\theta^{[k]})p(\theta^{[k]})$

事後確率の評価

$\theta^{[k+1]}\leftarrow\theta^{[k]}$

- 最初に効果的に発生：Importance Sampling
- 逐次的に更新：MCMC

初期値のみ最適化

$$p(\boldsymbol{x}_0^{[k]}\mid\boldsymbol{y}_{1:T})\propto p(\boldsymbol{y}_{1:T}\mid\boldsymbol{x}_0^{[k]})p(\boldsymbol{x}_0^{[k]}\mid\boldsymbol{y}_{*:0})$$

$\boldsymbol{x}_0^{[k+1]}\leftarrow\boldsymbol{x}_0^{[k]}$　次元が巨大なため、更新スキームの工夫が必要

pMCMC: Particle MCMC (Andrieu *et al.*, *JRSS*, 2010)

PFの最大のデメリットである退化問題を気にしなくて良い

1) Particle filter (Particle Smoother)

尤度

$$\hat{p}(\boldsymbol{y}_{1:T} \mid \theta^{[k]}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}, \theta^{[k]})$$

状態ベクトル列のサンプル

$$\left\{ \boldsymbol{x}_{1:T}^{(i)} \right\}_{i=1}^{m}$$

SIS を適用

繰り返す

$\theta^{[k+1]} \leftarrow \theta^{[k]}$

2) MCMC

Metropolis 法

Proposal function: $\theta^* \sim q(\theta^* \mid \theta^{[k]})$ ただし $q(\theta^* \mid \theta^{[k]}) = q(\theta^{[k]} \mid \theta^*)$

θ^* is accepted with the following probability: $\min\left(1, \dfrac{\hat{p}(\boldsymbol{y}_{1:T} \mid \theta^*) p(\theta^*)}{\hat{p}(\boldsymbol{y}_{1:T} \mid \theta^{[k]}) p(\theta^{[k]})}\right)$

$p(\boldsymbol{x}_{1:T}, \theta \mid \boldsymbol{y}_{1:T})$ からのサンプルを同時に得られる。

4DVar法

初期値のみ最適化

$$p(\boldsymbol{x}_0^{[k]} \mid \boldsymbol{y}_{1:T}) \propto p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_0^{[k]}) p(\boldsymbol{x}_0^{[k]} \mid \boldsymbol{y}_{*:0})$$

$$\boldsymbol{x}_0^{[k+1]} \leftarrow \boldsymbol{x}_0^{[k]}$$

次元が巨大なため、更新スキームの工夫が必要

$$\left. \frac{\partial \log p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_0)}{\partial \boldsymbol{x}'_0} \right|_{\boldsymbol{x}_0 = \boldsymbol{x}_0^{[k]}}$$

を効率よく求める計算法

$\boldsymbol{x}'_0$ は $\boldsymbol{x}_0$ の転置

微分の連鎖率

$$\frac{\partial \log p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_0)}{\partial \boldsymbol{x}'_0} = \frac{\partial \log p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_0)}{\partial \boldsymbol{x}'_T} \cdot \frac{\partial \boldsymbol{x}_T}{\partial \boldsymbol{x}'_{T-1}} \cdot \frac{\partial \boldsymbol{x}_{T-1}}{\partial \boldsymbol{x}'_{T-2}} \cdots \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{x}'_0}$$

システムモデル

$$\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1})$$

尤度関数の分解公式

$$\log p(\boldsymbol{y}_{1:T} \mid \boldsymbol{x}_0) = \sum_{t=1}^{T} \log p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}, \boldsymbol{x}_0) = \sum_{t=1}^{T} J_t = J_{1:T}$$

$$J_t = \log p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}, \boldsymbol{x}_0)$$

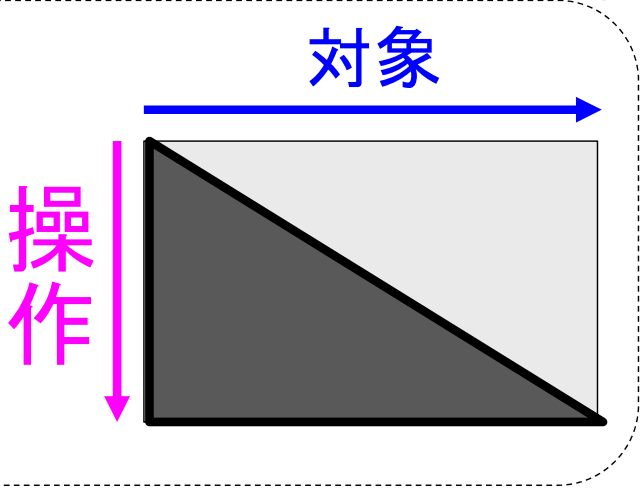
$$J_1 = \log p(\boldsymbol{y}_1 \mid \boldsymbol{x}_0)$$

アジョイント法

$$\frac{\partial J_{1:T}}{\partial \boldsymbol{x}'_0} = \frac{\partial J_0}{\partial \boldsymbol{x}'_0} + \left(\frac{\partial J_1}{\partial \boldsymbol{x}'_1} + \left(\cdots + \left(\frac{\partial J_{T-2}}{\partial \boldsymbol{x}'_{T-2}} + \left(\frac{\partial J_{T-1}}{\partial \boldsymbol{x}'_{T-1}} + \frac{\partial J_T}{\partial \boldsymbol{x}'_T} \frac{\partial \boldsymbol{x}_T}{\partial \boldsymbol{x}'_{T-1}} \right) \frac{\partial \boldsymbol{x}_{T-1}}{\partial \boldsymbol{x}'_{T-2}} \right) \cdots \frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{x}'_1} \right) \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{x}'_0}$$

$$\boldsymbol{d}_T = \frac{\partial J_T}{\partial \boldsymbol{x}'_T}, \quad \boldsymbol{d}_{T-1} = \frac{\partial J_{T-1}}{\partial \boldsymbol{x}'_{T-1}} + \boldsymbol{d}_T \frac{\partial \boldsymbol{x}_T}{\partial \boldsymbol{x}'_{T-1}}$$

平行移動しても解法に影響はない



$$\boldsymbol{d}_T = \boldsymbol{0}, \quad \boldsymbol{d}_{t-1} = \frac{\partial J_{t-1}}{\partial \boldsymbol{x}'_{t-1}} + \boldsymbol{d}_t \frac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}'_{t-1}} \quad (t = T, \ldots, 1)$$

時間をさかのぼる漸化式

$$\boldsymbol{d}_0 = \frac{\partial J_0}{\partial \boldsymbol{x}'_0} + \boldsymbol{d}_1 \frac{\partial \boldsymbol{x}_1}{\partial \boldsymbol{x}'_0} = \frac{\partial J_{1:T}}{\partial \boldsymbol{x}'_0}$$

$J_0 = \log p(\boldsymbol{y}_0 \mid \boldsymbol{x}_0) = 0$ とする

降下法で更新 $\boldsymbol{x}_0^{[k+1]} = \boldsymbol{x}_0^{[k]} + s \cdot \boldsymbol{d}_0^{[k+1]} \quad (k = 0, 1, \ldots)$

S: スケーリング係数

アジョイントコード

$$\begin{cases} \boldsymbol{d}_{t-1} = \dfrac{\partial J_{t-1}}{\partial \boldsymbol{x}'_{t-1}} + \boldsymbol{d}_t \dfrac{\partial \boldsymbol{x}_t}{\partial \boldsymbol{x}'_{t-1}} \quad (t = T, \ldots, 1) \\ \boldsymbol{y}_t = H\boldsymbol{x}_t + \boldsymbol{w}_t,\ \boldsymbol{w}_t \sim N(0, R_{\mathrm{obs}}) \ \text{の時は} \end{cases}$$

$$\boldsymbol{d}_{t-1} = -\left(\boldsymbol{y}_{t-1} - H\boldsymbol{x}_{t-1}\right)' R_{\mathrm{obs}}^{-1} H + \boldsymbol{d}_t \frac{\partial f(\boldsymbol{x}_{t-1})}{\partial \boldsymbol{x}'_{t-1}}$$

転置（Ajoint）をとった形で計算

ヤコビ行列

$$\boldsymbol{d}'_{t-1} = -H' R_{\mathrm{obs}}^{-1}\left(\boldsymbol{y}_{t-1} - H\boldsymbol{x}_{t-1}\right) + \left[\frac{\partial f(\boldsymbol{x}_{t-1})}{\partial \boldsymbol{x}'_{t-1}}\right]' \cdot \boldsymbol{d}'_t$$

アジョイントコード

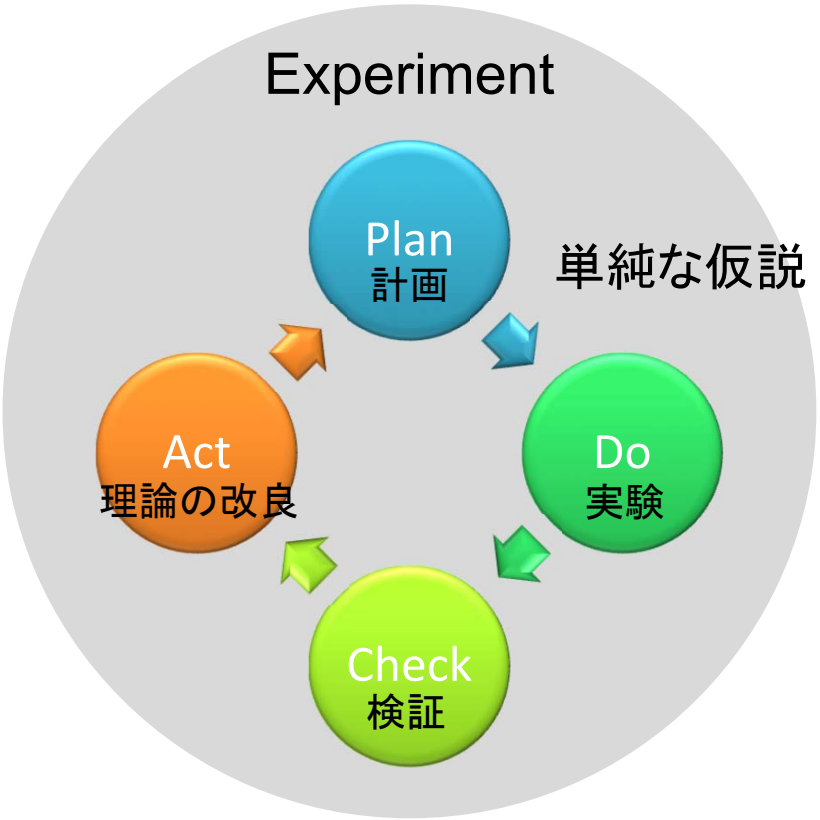
4DVar vs. アンサンブルベースデータ同化法

	非逐次型	逐次型	
	4DVar	EnKF	PF
	Ajoint	*EnTKF, LETKF*	*MPF, pMCMC*
	力学的バランスを重視（システムノイズ無し）	簡便＆安心	超簡便＆原理的には万能
Pros（○）	・状態ベクトルの次元が非常に大きい、最大規模のシミュレーションモデルが取り扱える。 ・感度解析が可能 ・ベクトル計算向き	・実装が容易 ・退化現象（分布表現能力の減少。）が原理的におきない。	・実装が著しく容易 ・観測モデルが非線形の場合にも自然に対応可能 ・並列計算向き
Cons（×）	・時間を遡るシミュレーションコードを書きおろす必要があるため、人的労力の負荷が高い。 ・統計モデルでないので、統一的な視点や基準でもってモデル解析ができない。	・共分散行列の更新ステップの計算コストが高い。 *TKFでは大幅に軽減* ・分布がガウスから大きく逸脱した時には誤った結果を導く。 ・超高次元状態ベクトルのシミュレーションモデルが取り扱えない。	・時不変パラメータ推定問題の場合は、**退化現象**がおきる。 ・厳密な平滑化アルゴリズムの実現が実質的に無理
	気象庁	全世界的にこちらにシフト 東大・大気海洋研 理研AICS アンサンブル数は数十から百のオーダー	非線形度が高い 小規模問題

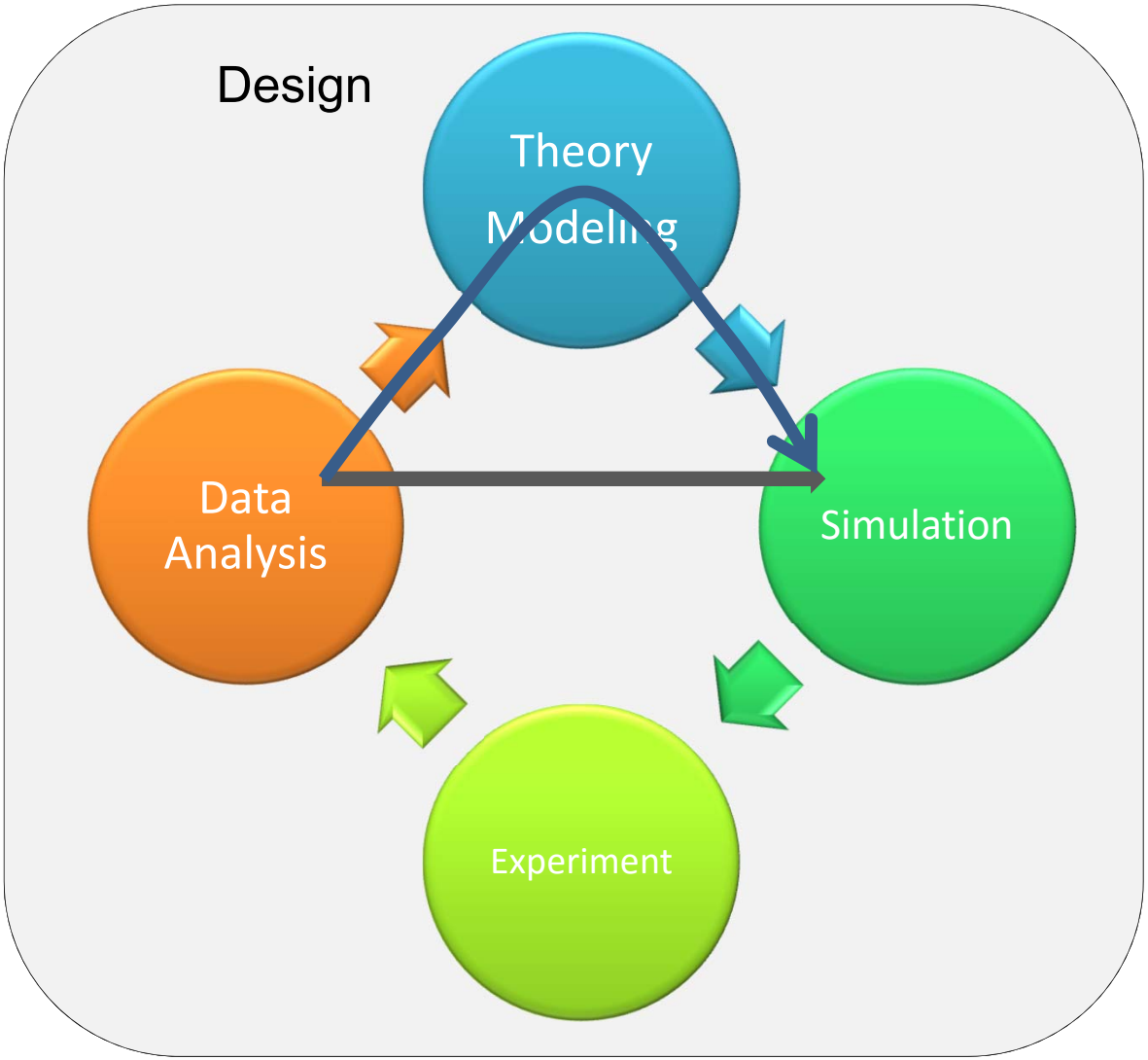
アウトライン

1. ベイズ統計の基礎
2. 状態空間モデルと逐次ベイズ
3. 逐次データ同化
4. 統計的推測
5. 実験計画

大規模実験システム



古典的研究開発の枠組み



現代的製品開発研究の枠組み

UQ: Uncertainty Quantification

- 欧米では、計算機シミュレーション結果の信頼性を具体的に確立するための方法論の研究が急速に熱を帯びてきており、ASME(The American Society of Mechanical Engineers)がVerification and Validation（通常V&Vと呼称）の標準化に大きな力を注いでいる。例えば、2006年には固体力学に対して、2009年には流体力学および熱解析に関する計算機シミュレーションのV&Vが公表されている。
- 欧州においては流体力学の分野で同種の研究活動が2012年から活発化しており、Uncertainty Quantification (UQ) in Industrial Analysis and Design の名のプロジェクト研究が現在進行中である。
- NASAでは、NASA UQ challenge 2014と題して、スパースな限定されたパラメータセットに関するシミュレーションの結果データから、UQをモデル化するコンペを開始した。
- 米国統計コミュニティは、2011-12年に、NSFのサポートを受ける機関SAMSI（Statistical and Applied Mathematical Sciences Institute）にてUQを集中的に研究するプログラムを立ち上げた。
- 米国統計学会はSIAM(Society for Industrial and Applied Mathematics)と共同でJournal on UQの刊行を2014年に開始した。その雑誌の取り扱う主たる分野としてsensitivity analysis, model validation, model calibration, data assimilationの4つがあげられている。最新号の論文（4本掲載）は、感度解析、ガウス過程回帰、モデル較正、ギプスサンプラーの解析のテーマとなっており、ほぼ統計学の範疇である。

重要な技術：

ガウス過程回帰や、その古典版とも言えるクリギング

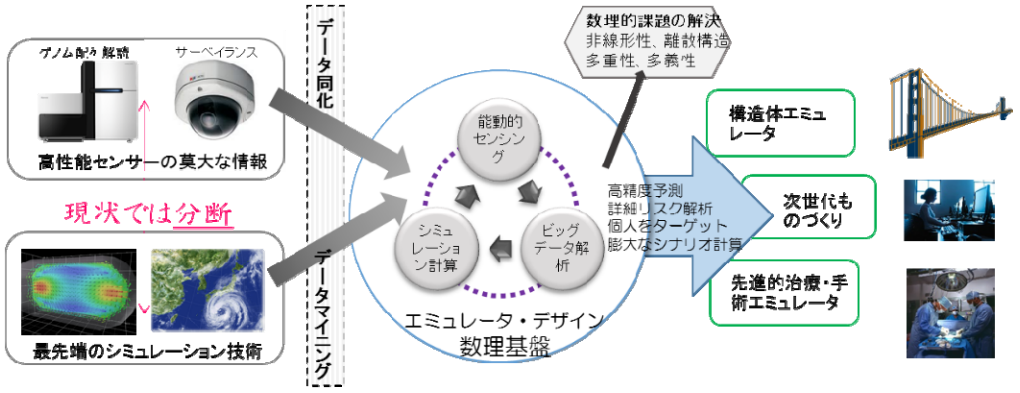
次元削減を目的としたスパース回帰

中野慎也、樋口知之、地球科学におけるシミュレーションとビッグデータ—データ同化とエミュレーション—、電子情報通信学会誌、Vol.97(10), pp.869-875, 2014.

樋口知之、中村和幸、データ同化によるオンラインセンシングの高度化, 計測自動制御学会誌, Vol.51(9), 2012.

長尾大道、佐藤光三、樋口知之, マルコフ連鎖モンテカルロ法を利用したトレーサー試験からフラクチャーの物理パラメータを推定する方法, 石油技術協会誌, Vol.78(2), pp.197-209, 2013.

Iba, Y. and Akaho, S., Gaussian process regression with measurement error, IEICE Trans. E93-D(10), 2010.



データ同化型シミュレーション技術開発によるものづくり・設計の革新
―データ同化技術の工学への応用研究開発―

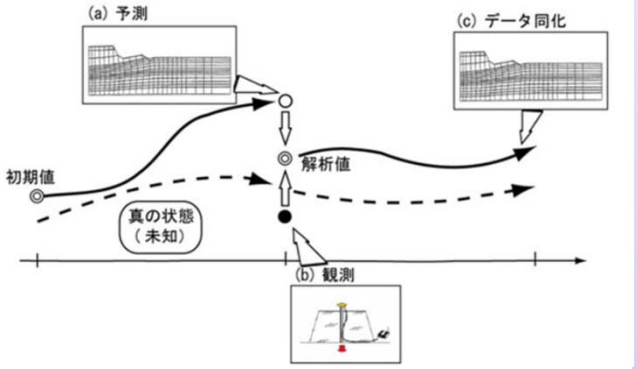
COCN
Council on Competitiveness-Nippon
HPC応用研究会提言
産業競争力懇談会
(2012年3月6日)
より抜粋・要約

【効果あるモデリング方法論】
(ア)自然現象を数学モデルに近似するモデルによる誤差 (イ)材料データなど入力データを持つばらつき(構成式，減衰率。熱伝達率など)に対応するシーズ技術としては，DA(データ同化)を提言したい。DAは複雑な現象を科学的に理解し，精度よく予測したいという要求にこたえる技術であり，モデルから複雑現象を再現する演繹アプローチと複雑現象の観測結果からモデルを推測する帰納アプローチが融合した次世代のシミュレーション技術である。これまで主に地球科学の分野において数値モデルの再現性を高めるためにモデルに観測データを埋め込み，馴染ませることを意図して研究された。モノづくり分野において，なじみが薄い名前ではあるが，たとえば，電子機器の設計分野では20年も前から実質的に同等の技術が活用されている。

―現代社会に強く求められるデータ同化技術の研究開発―

社会インフラ エミュレータ研究事例

地盤工学におけるデータ同化構造シミュレーション

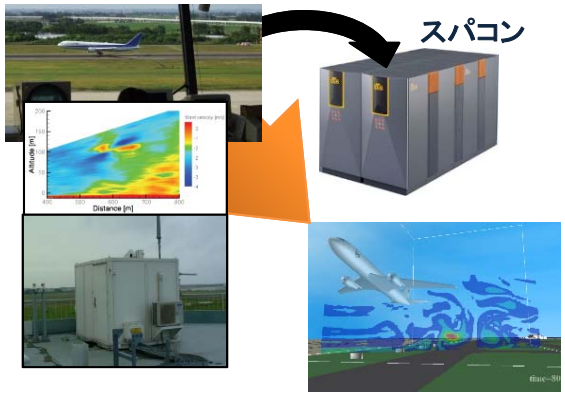


地盤挙動メカニズム解析と将来予測システムの開発
地盤の変位や応力、間隙水圧の確率分布を得てこれらの確率分布を基礎や土構造物のリスク評価に適用

地盤内応力状態把握 精度検証の実現

次世代ものづくり開発 研究事例

後方乱気流のライダ計測融合シミュレーション

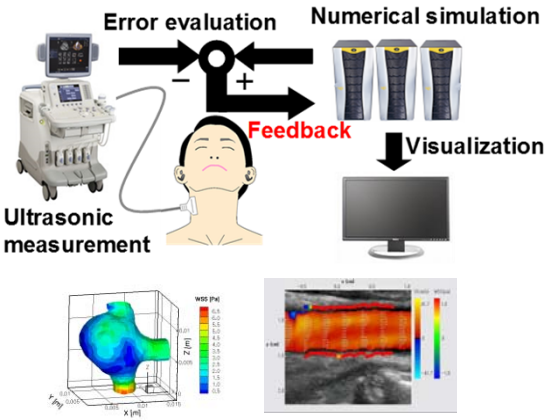


次世代の飛行機運航システムの開発
仙台空港における実運航機の後方乱気流をライダー計測し3次元の流体シミュレーションに融合

離発着間隔の制限解除 安全時短省エネ運航の実現

先進的治療・手術エミュレータ研究事例

データ同化型血流シミュレーション



超音波計測融合血流解析システムの開発
血流の超音波計測と数値解析を一体化し血流解析システムを開発し、疾患との関係や診断指標について研究

高度診断の実現

大学共同利用機関法人 情報・システム研究機構
統計数理研究所

メゾスコピック・モデリング