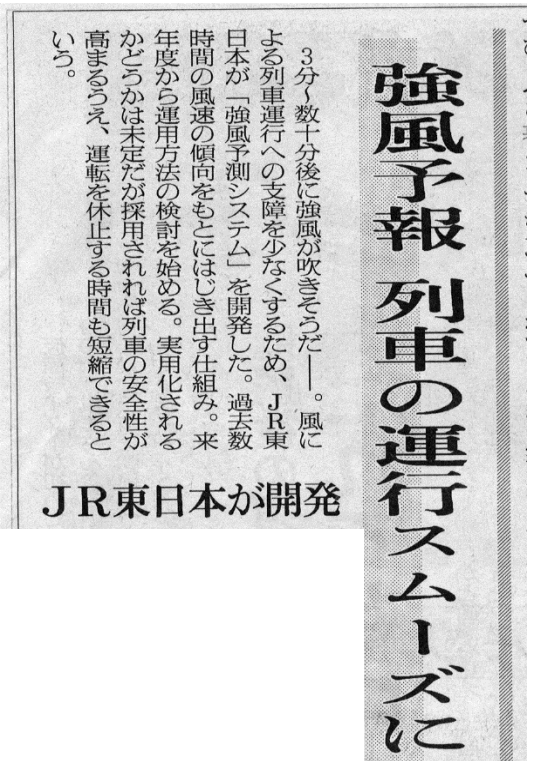
アンサンブルベース逐次データ同化とHPC

樋口知之　（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

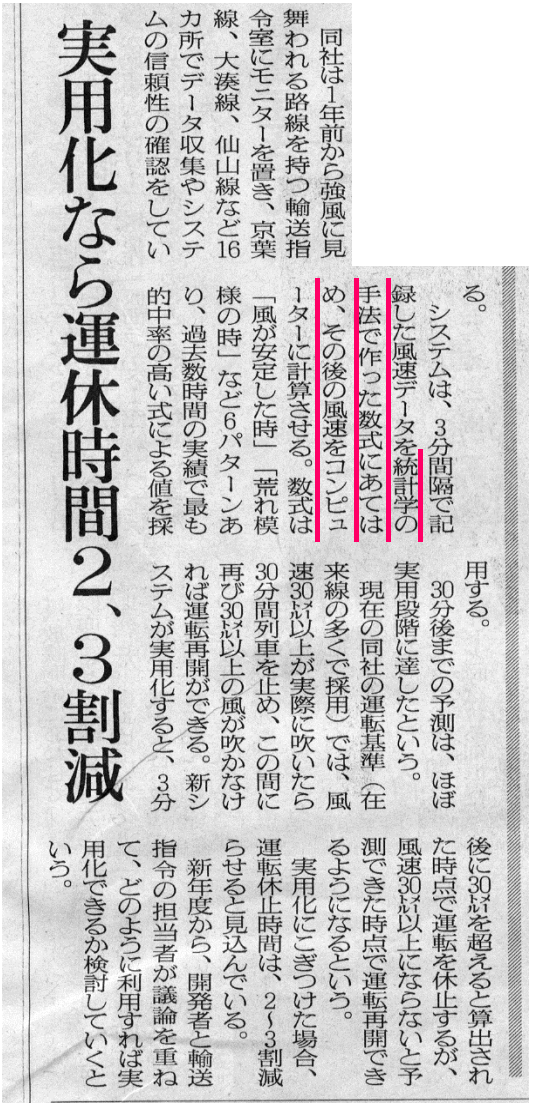
朝日新聞平成14年2月17日(日)　14版第２社会面



強風予報 列車の運行スムーズに

3分～数十分後に強風が吹きそうだ――。風による列車運行への支障を少なくするため、JR東日本が「強風予測システム」を開発した。過去数時間の風速の傾向をもとにはじき出す仕組み。来年度から運用方法の検討を始める。実用化されるかどうかは未定だが採用されれば列車の安全性が高まるうえ、運転を休止する時間も短縮できるという。

JR東日本が開発



実用化なら運休時間2、3割減

同社は1年前から強風に見舞われる路線を持つ輸送指令室にモニターを置き、京葉線、大湊線、仙山線など16カ所でデータ収集やシステムの信頼性の確認をしている。

システムは、3分間隔で記録した風速データを統計学の手法で作った数式にあてはめ、その後の風速をコンピューターに計算させる。数式は「風が安定した時」「荒れ模様の時」など6パターンあり、過去数時間の実績で最も的中率の高い式による値を採用する。

30分後までの予測は、ほぼ実用段階に達したという。

現在の同社の運転基準（在来線の多くで採用）では、風速30㍍以上が実際に吹いたら30分間列車を止め、この間に再び30㍍以上の風が吹かなければ運転再開ができる。新システムが実用化すると、3分後に30㍍を超えると算出された時点で運転を休止するが、風速30㍍以上にならないと予測できた時点で運転再開できるようになるという。

実用化にこぎつけた場合、運転休止時間は、2～3割減らせると見込んでいる。

新年度から、開発者と輸送指令の担当者が議論を重ねて、どのように利用すれば実用化できるか検討していくという。

ＪＲ東　安全研との共同研究

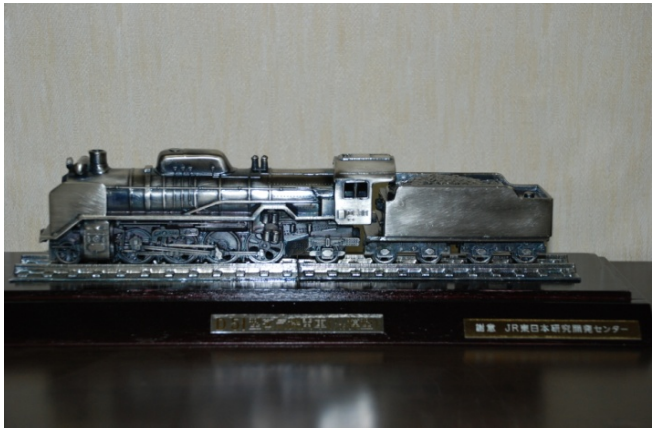
感謝状

大学共同利用機関法人
情報・システム研究機構
統計数理研究所殿

貴研究所は時系列解析手法を利用した「強風警報システム」について弊社との共同利用研究「鉄道の強風予測に関する時系列解析」に尽力され予測手法の確立とその導入によって弊社の安全安定輸送の確保に大きく貢献されましたよって深甚なる感謝の意を表します

平成二十年十月二日

東日本旅客鉄道株式会社
執行役員　ＪＲ東日本研究開発センター所長
荒井　稔



帰納の論理

●論理学者 vs. 数学者 『数学者の厳密性の甘さかげんに我慢がならない』

●物理学者:『60という数はすべての数で割りきれる』

1,2,3,4,5,6 OK. “勝手にとったところの”10,12,15,20,30,...

●技術者:『すべての奇数は素数』

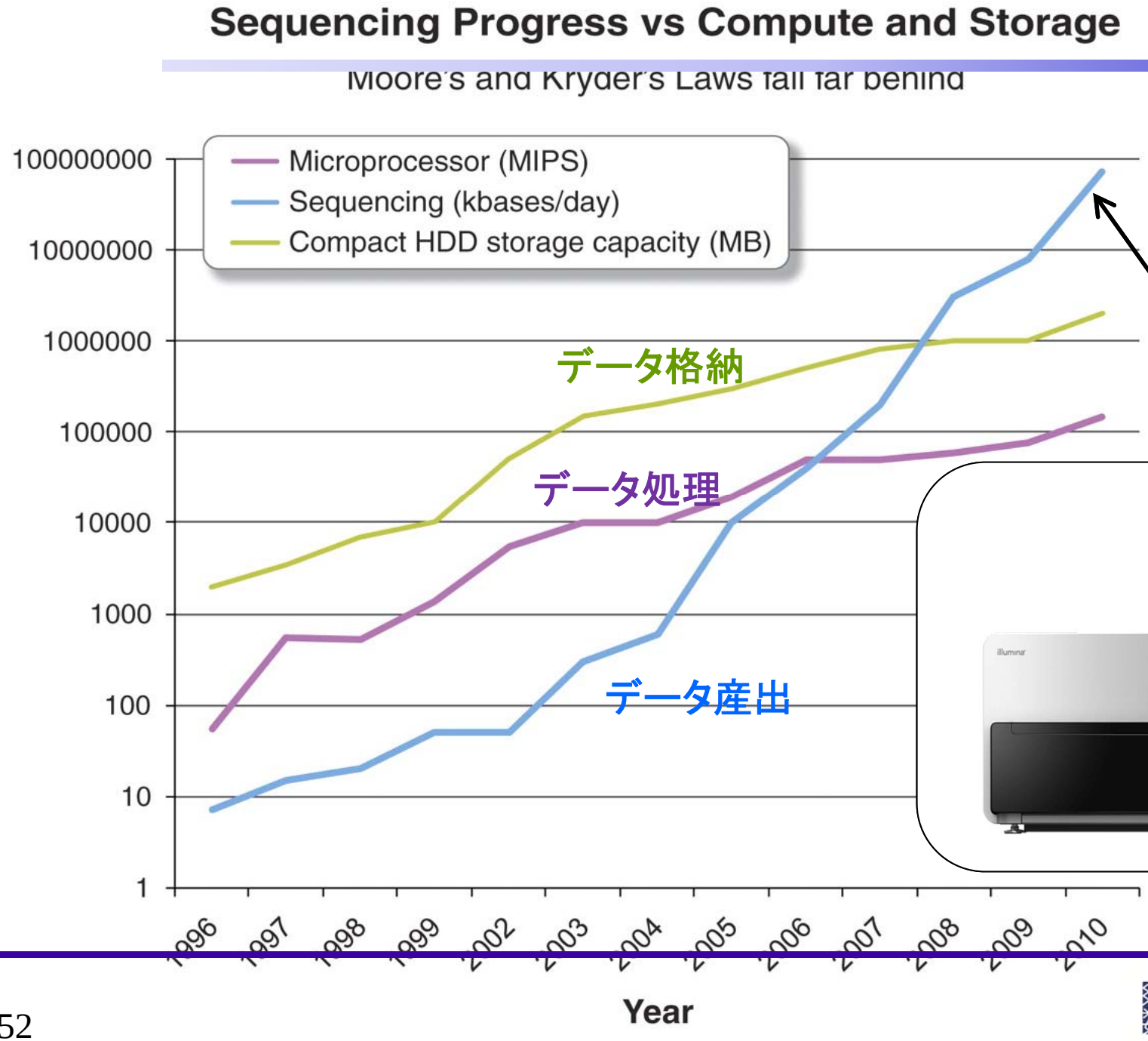
例を使って考えるなといっているのではない.
むしろそうしろと言っているのだ.

ようて，これは測定の誤差

個々の特殊な事例を調べると，そこに共通する性質や関係から，一般的な命題や法則を推論できることがある．これを帰納という．帰納の結果が誤りに導くこともいかに多いかをわれわれは知っている．しかし，注意深く，感性を研ぎ澄ましてかかれば，帰納がすばらしい結果に導くことがあるということもわれわれは数々の例からしっているのである．

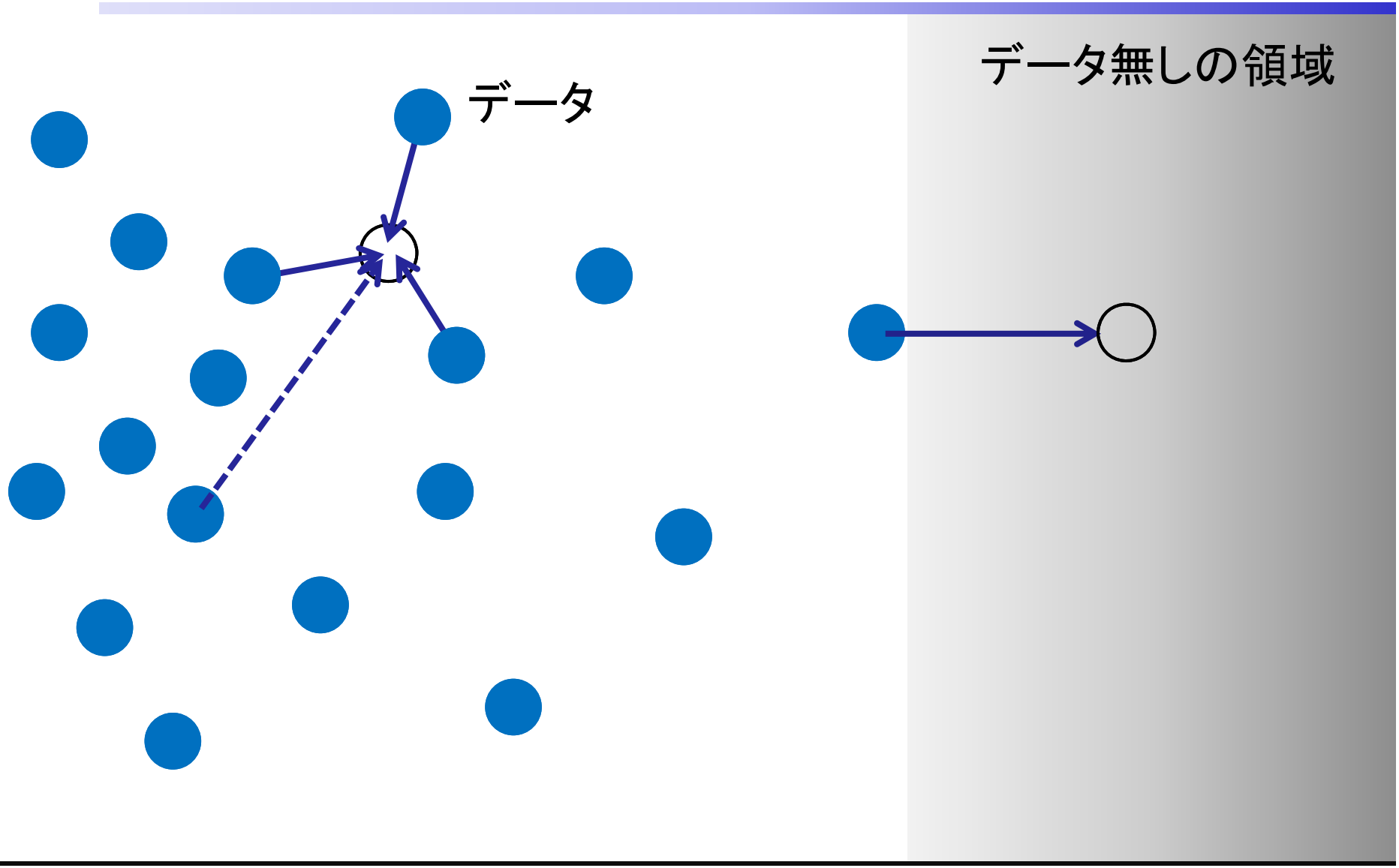
G．ポリア著，「帰納と類比」から
（実際は，金出武雄著「素人のように考え，玄人として実行する」から）

データ中心主義の時代へ



100Gb/day
/sequencer

内挿と外挿問題　－得意と不得意－



CPSとは

- Center for Planetary Science

予測能力の向上に集中！

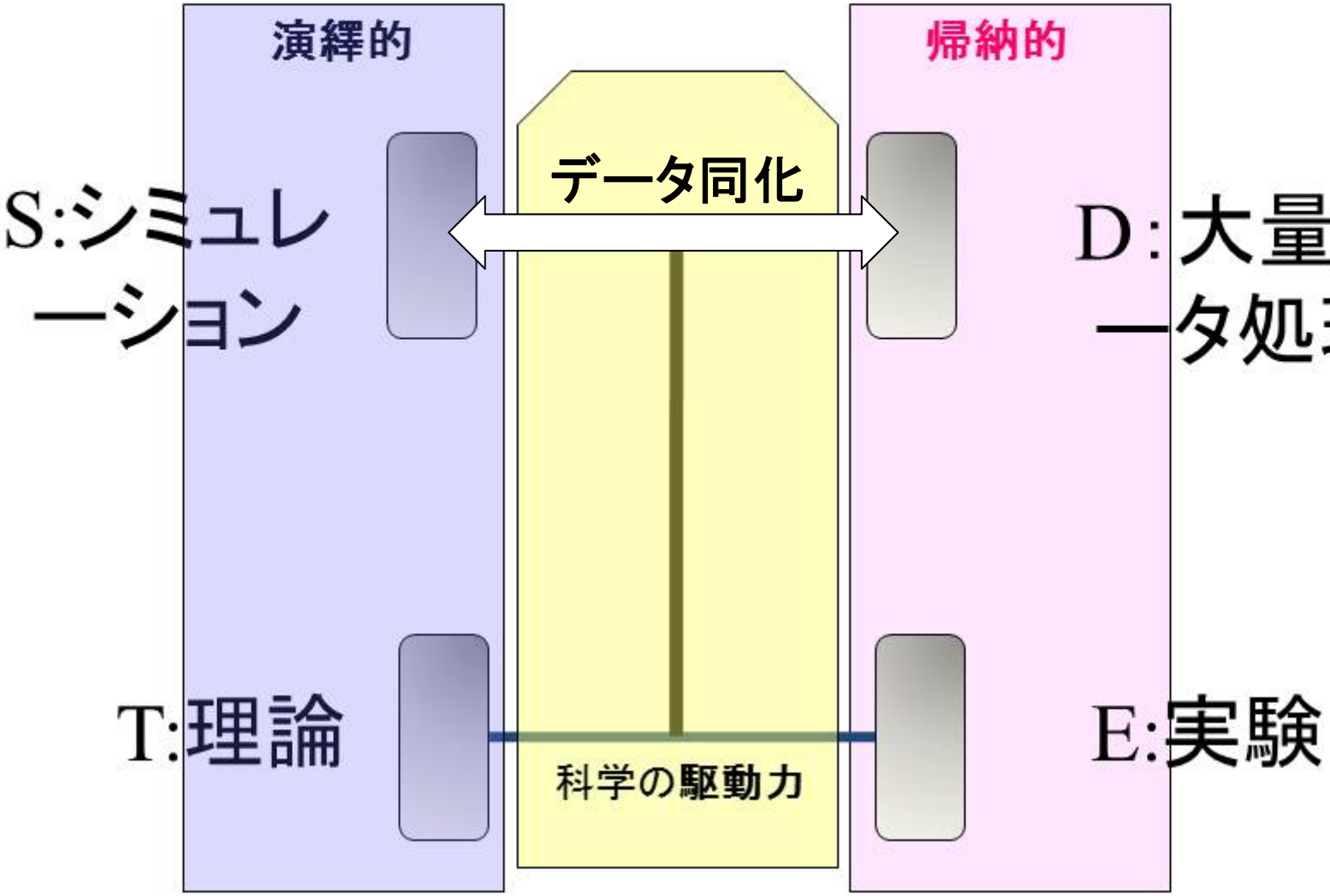
二つの機能の性能の合成

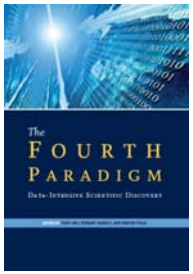
フォワード（前向き）計算モデルの記述力
＋
対象の現状態（現況）を捉える認識力
＝　予測能力

計測手法のイノベーションと直結

地球や生命体のような複雑なシステムの理解と制御においては、対象に関する知識は常に不完全であることを前提に、現象の予測能力でもって研究の進め方を評価し、修正する方策が有効

つなぐ：データ同化





The Fourth Paradigm: Data-Intensive Scientific Discovery
By *Tony Hey, Stewart Tansley, and Kristin Tolle, Eds*.
Microsoft Research, 2009.

データ同化の目的：気象・海洋学の観点から

[1] 予報を行うための最適な初期条件を求める。これは既に、現業の天気予報で実用化されていることである。

[2] シミュレーションモデルを構成する際の最適な境界条件を求める。連成現象を取り扱う際の適応的な境界条件設定もこの作業に含まれる。

[3] スケールが異なるシミュレーションモデル間の橋渡しを行うスキーム内に含まれる諸パラメータの最適な値を求める。経験的に与えられるモデル内のパラメータ値の検証も一つの具体例である。

[4] シミュレーション（物理）モデルにもとづいた、観測されていない時間・空間点における観測値の補間を行う。この作業は再解析データセットの生成とも呼ばれる。このデータセットから新しい科学的発見をもくろむ。

[5] 時間・経費を節約できる効率的な観測システムを構築するための仮想観測ネットワークシミュレーション実験や感度解析を行う。

（参考文献： 蒲地　他、「統計数理」、54(2), 223-245, 2006.)

シミュレーションモデルの構成（1）

（Tsunami simulation model in Japan Sea）

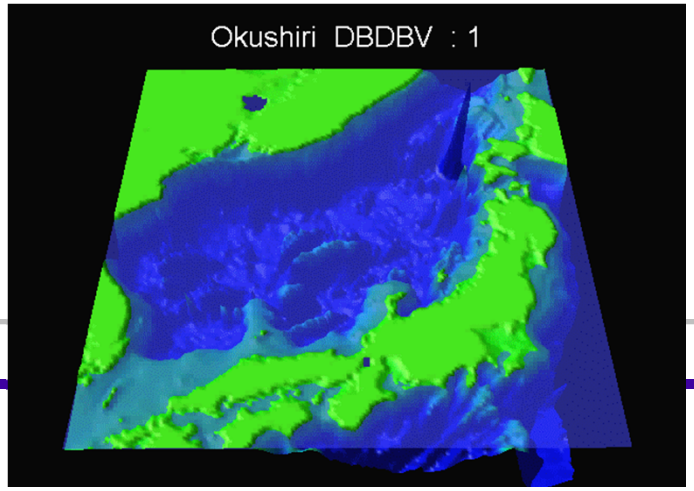
PDE to approximate real physical system
(continuous time/space)

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

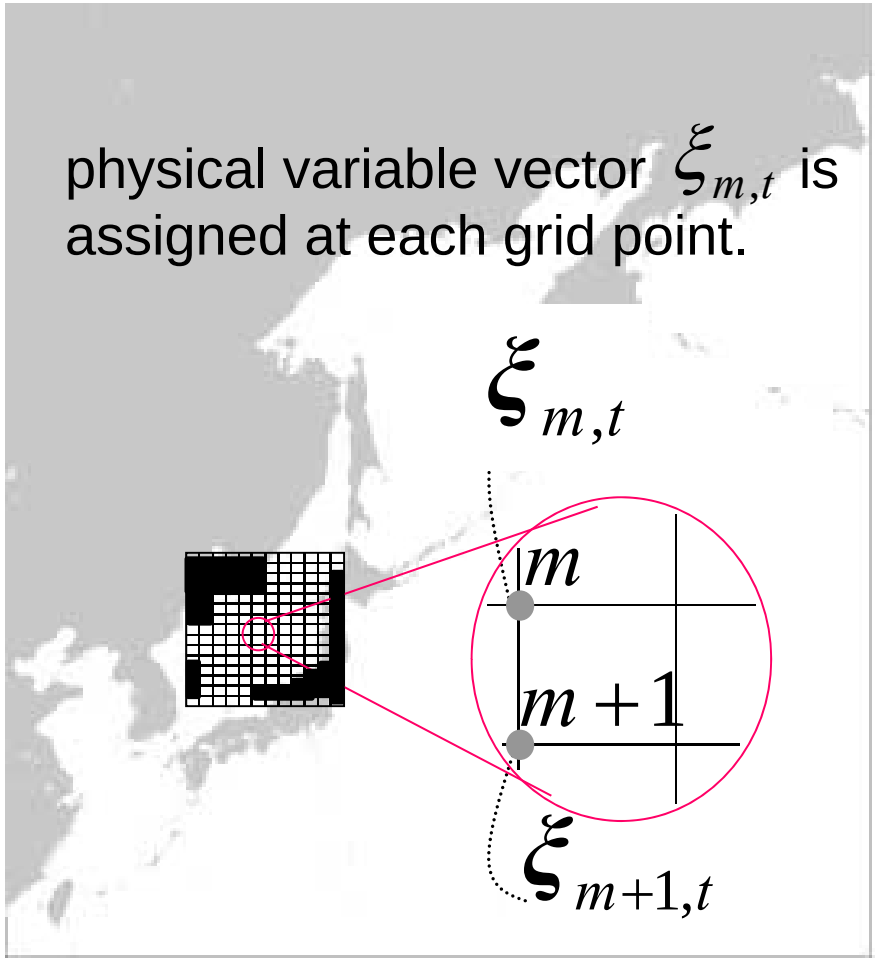
PDE : Partial differential equation

Discrete simulation model
(discrete time/space, FDE)

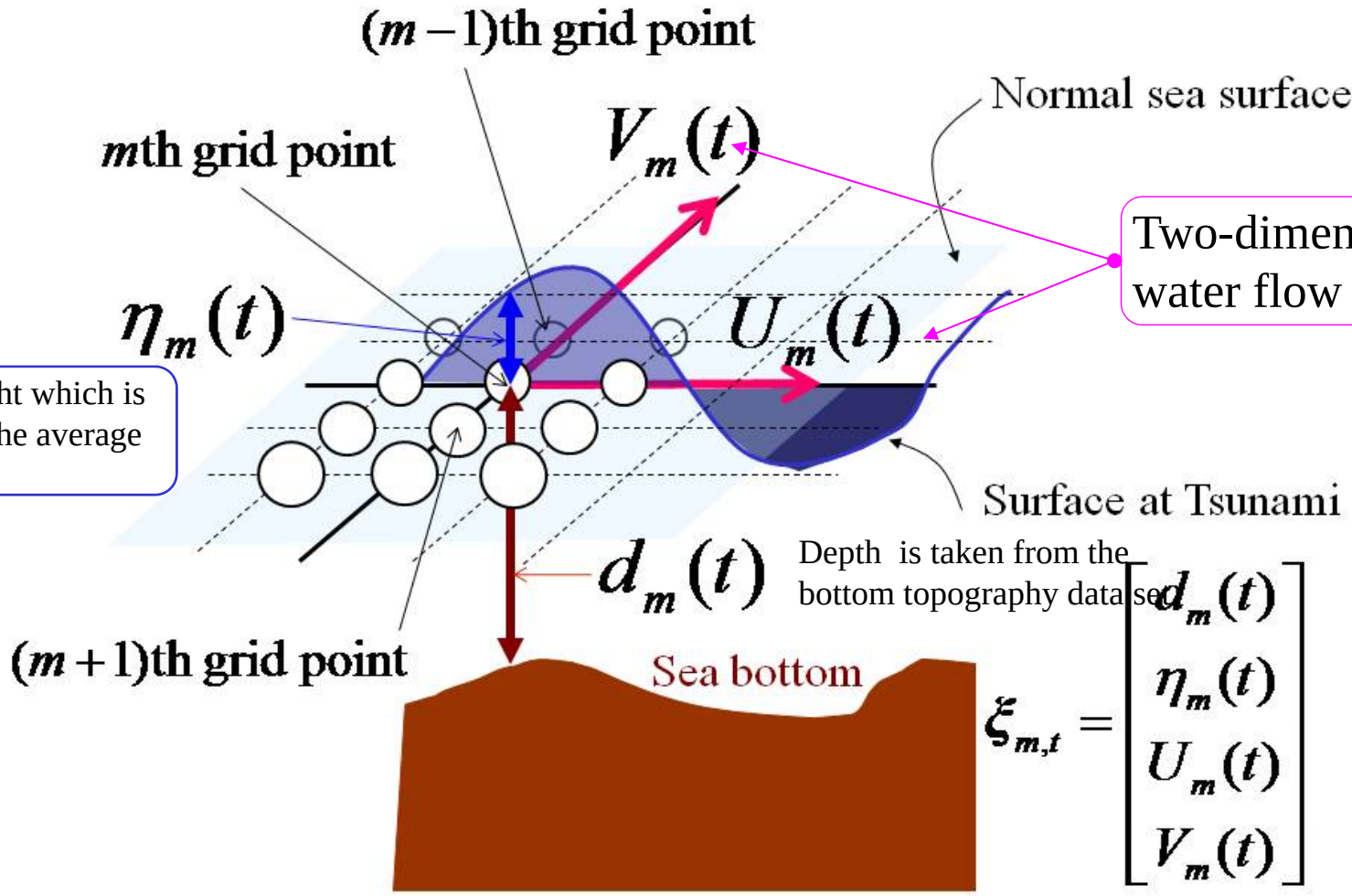
Suppose a case where we conduct a two-dimensional simulation experiment for understanding the flow of shallow water such as the tsunami.



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

State Vector（Simulation variables）

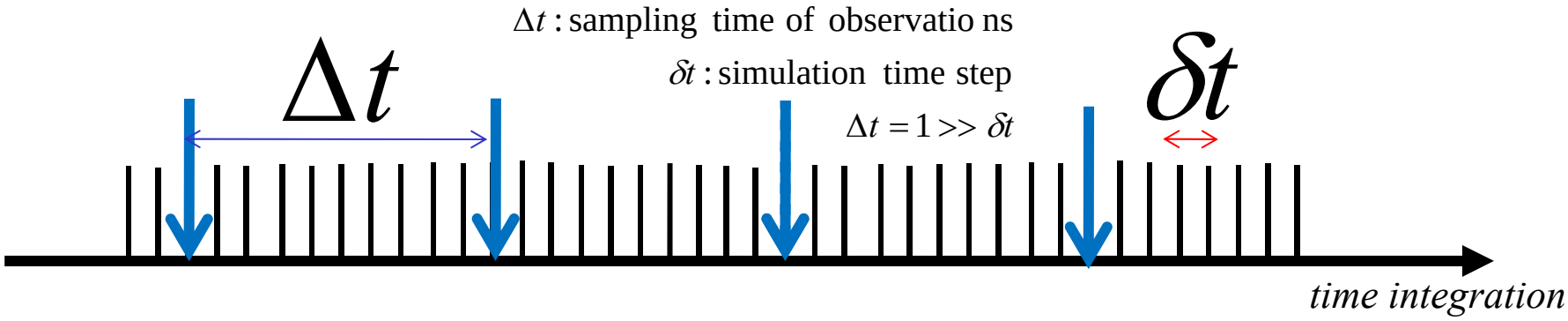
システムモデル
Stochastic simulation model

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

気象・海洋のデータ同化の枠組み

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

Observation model
Measurement model
観測モデル



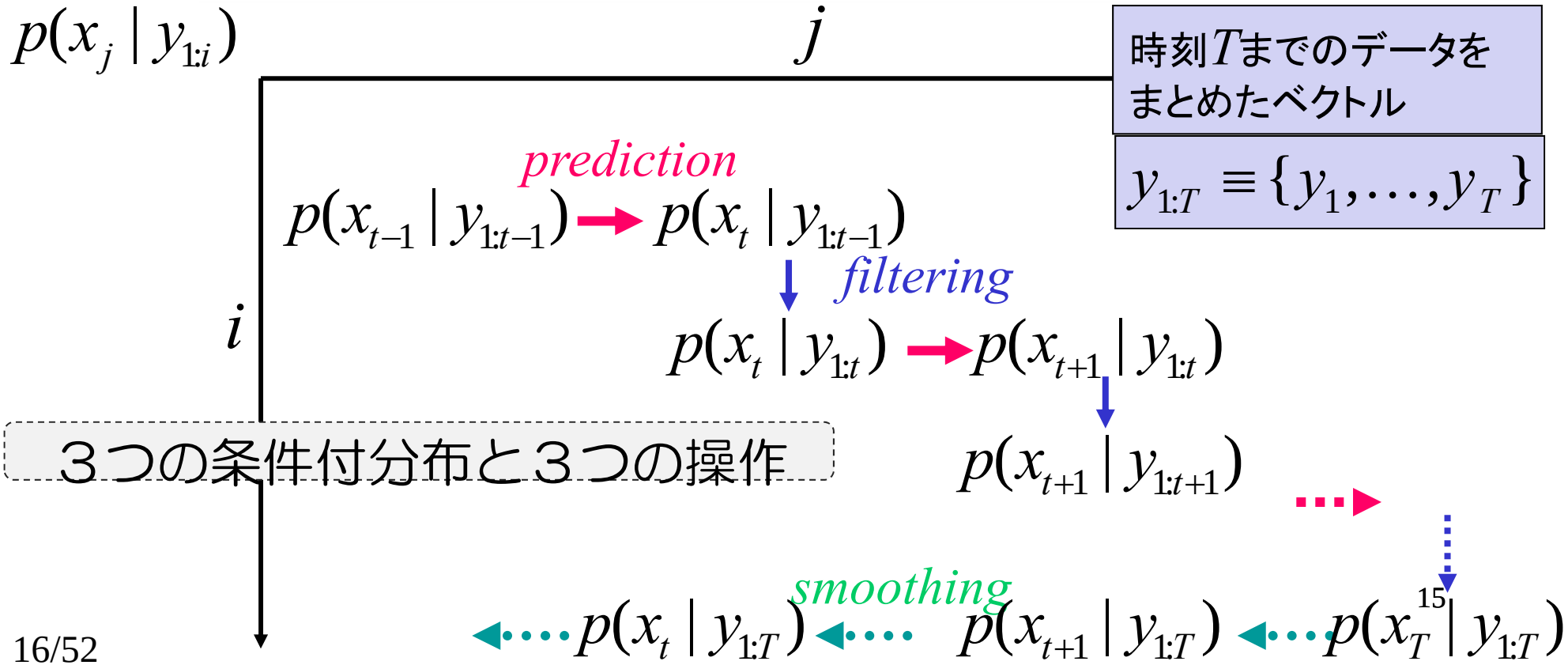
逐次ベイズ計算 --- 日次株価データを考えると ---

条件付き分布

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

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

smoother density: $p(x_t \mid y_{1:T})$
平滑化分布
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



フィルタリングとベイズの定理

x_t：時刻 tの興味のある対象　　$x_{1:T} \equiv \{x_1, \ldots x_T\}$

y_t：データ　　$y_{1:T} \equiv \{y_1, \ldots, y_T\}$

フィルタリング

観測モデル　予測分布

フィルタ分布

$$p(x_t \mid y_{1:T}) = \frac{p(y_t \mid x_t)p(x_t \mid y_{1:t-1})}{\sum_{x_t} p(y_t \mid x_t)p(x_t \mid y_{1:t-1})}$$

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

パラメータの事後分布

事前分布

事後分布

$$p(\theta \mid y_{1:T}) = \frac{p(y_{1:T} \mid \theta)p(\theta)}{\sum_{\theta} p(y_{1:T} \mid \theta)p(\theta)}$$

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

ベイズの反転公式

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

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

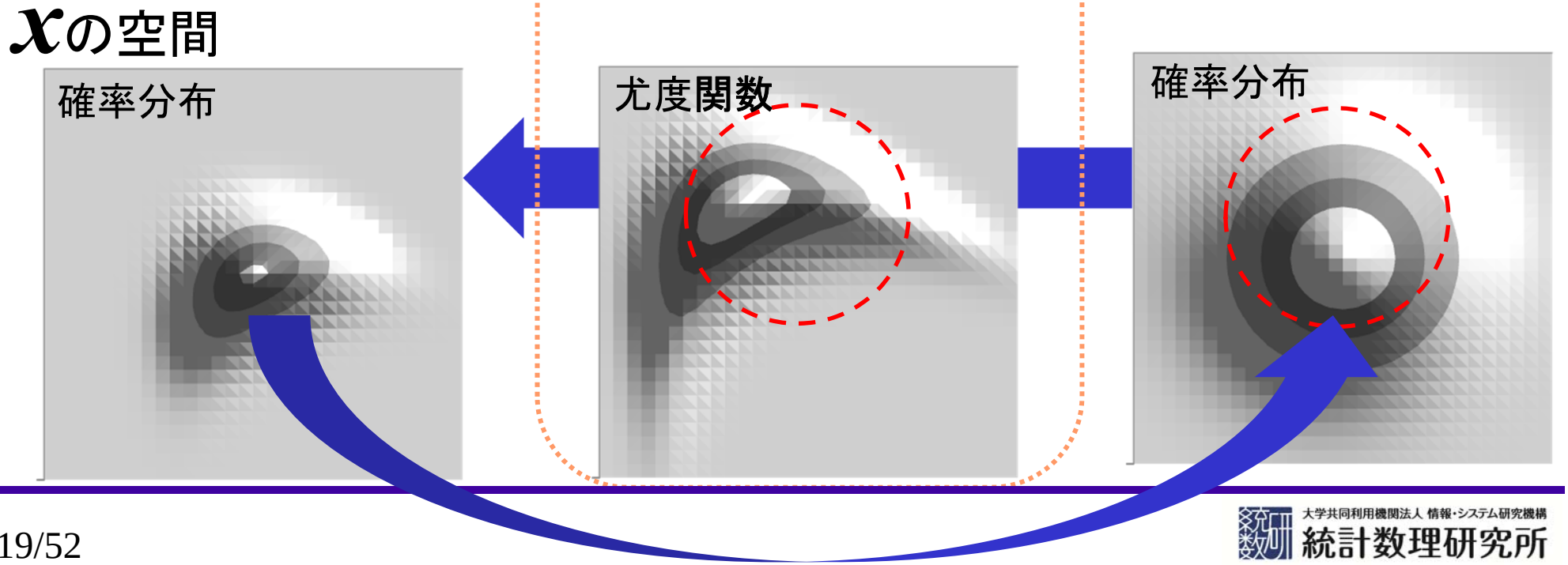
1．膨大な数の積分（和）操作には高速な計算機が必要
コンピュータの性能向上

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

ベイズの定理と情報循環

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

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



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} ←

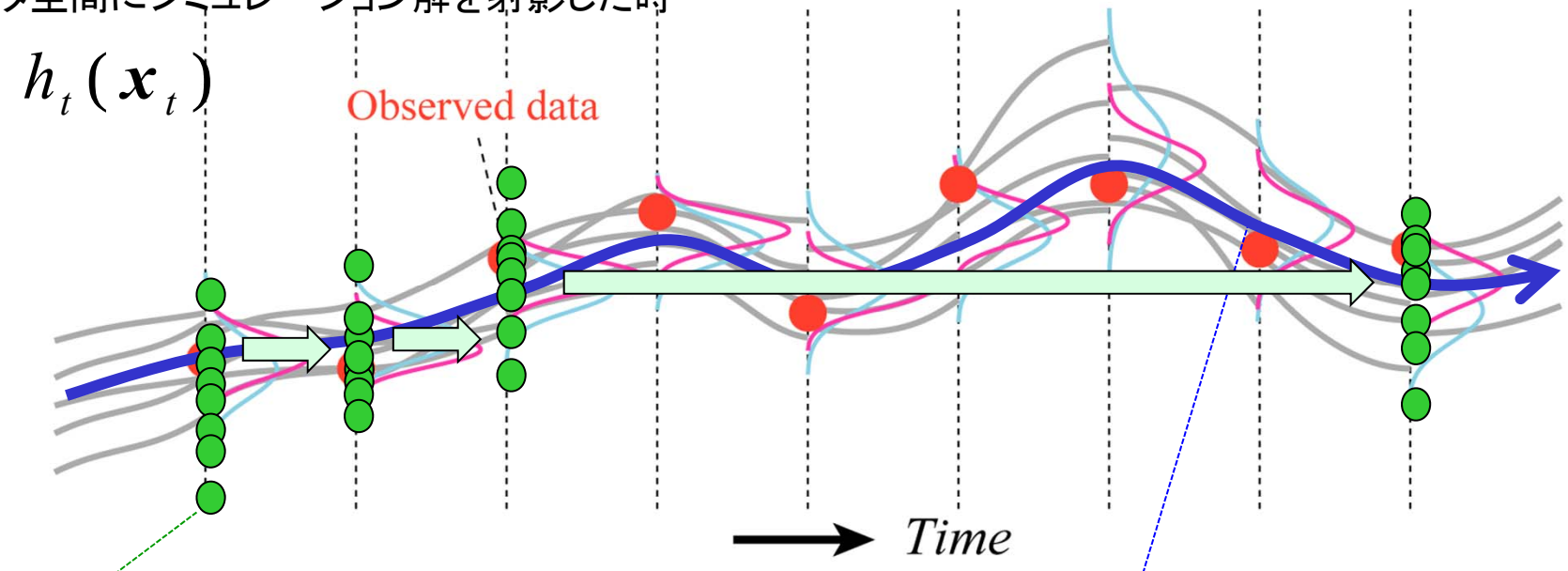
$p(x_{n-1} \mid y_{1:n-1})$

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

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



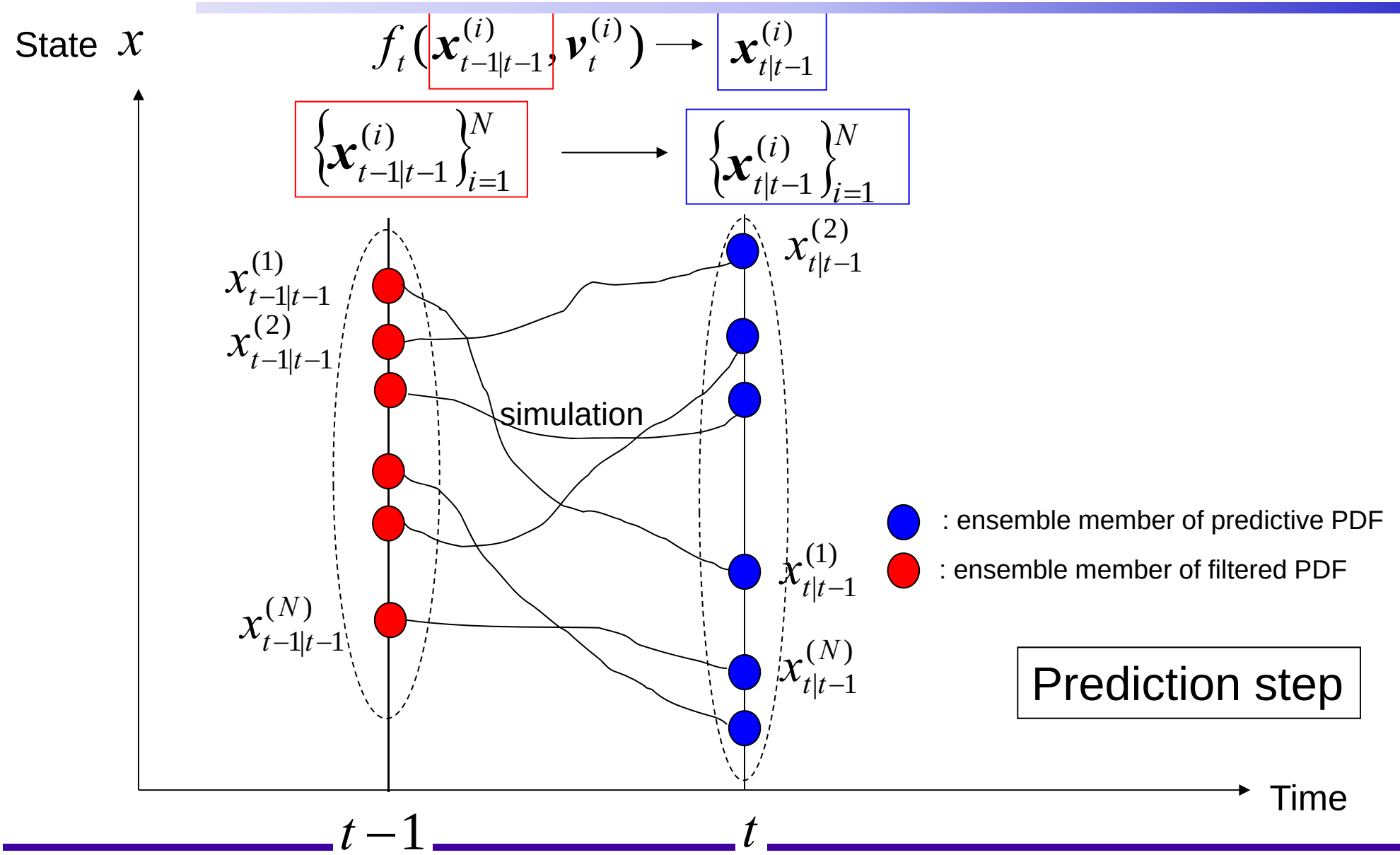
逐次（オンライン）型： 集団の時間発展を追う。つまり、**Swarm Filter**

代表例： EnKF (Ensemble Kalman Filter)

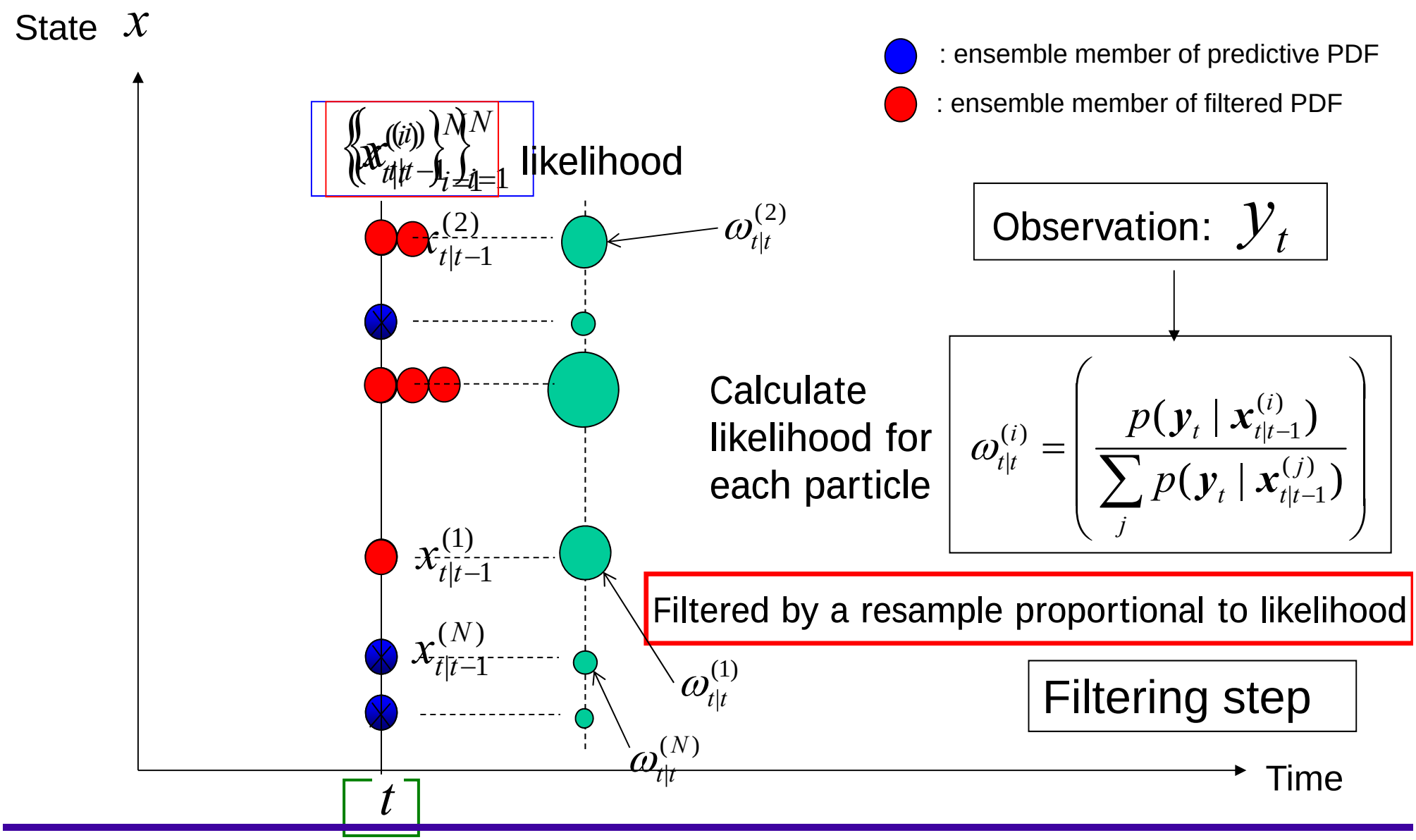
非逐次（オフライン）型：**ベスト初期値をもつパス**を求める

代表例： 4次元変分法（Adjoint法）

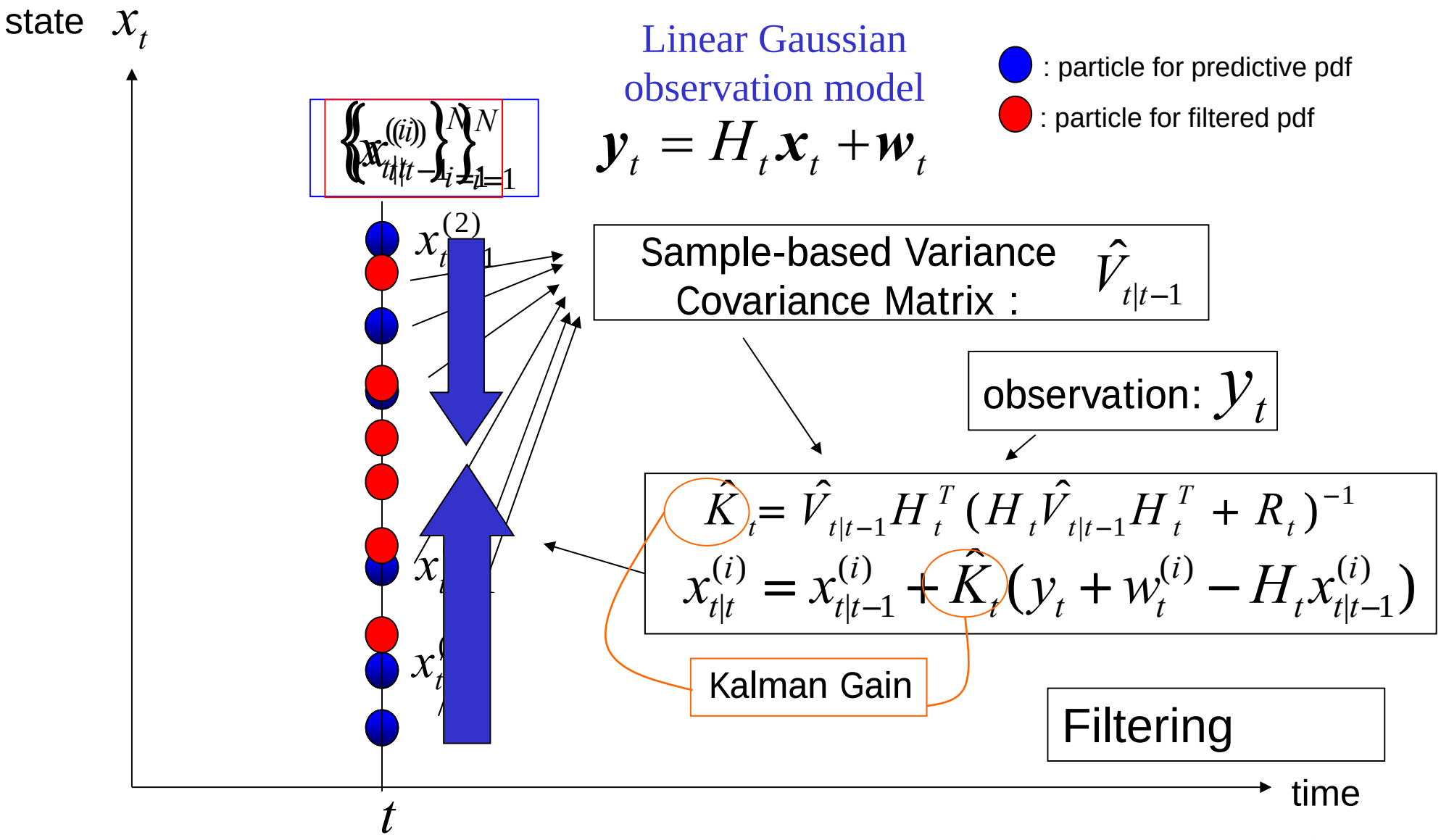
予測のステップ (EnKFとPFで共通)



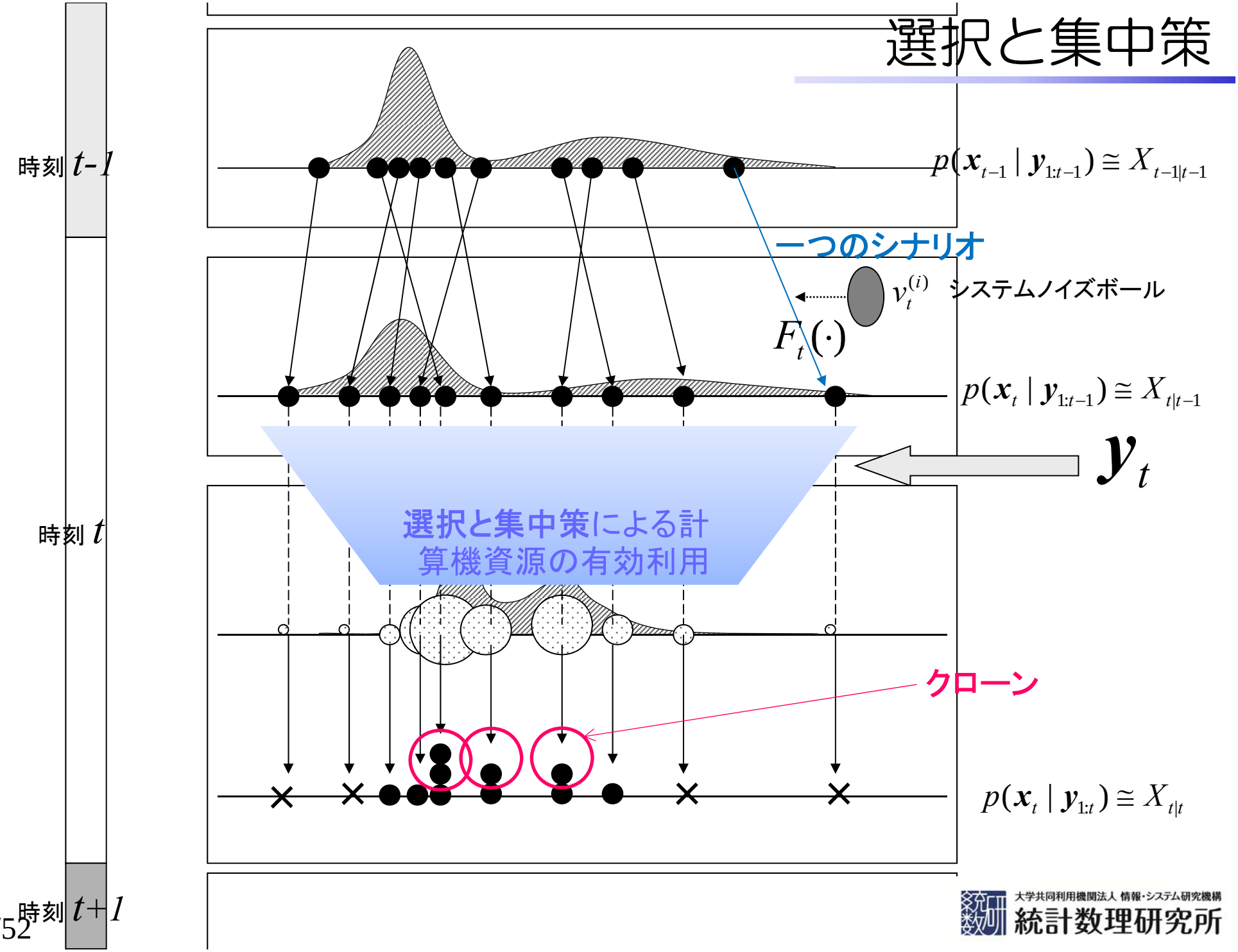
PFでのフィルタリングのステップ



EnKFでのフィルタリングのステップ



選択と集中策



４次元変分法とベイズ統計

予測

システムモデル

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

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

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

初期値のみ最適化

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

$$\hat{\boldsymbol{x}}_0^{(k+1)} \leftarrow \hat{\boldsymbol{x}}_0^{(k)}$$

パラメータの事後分布

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

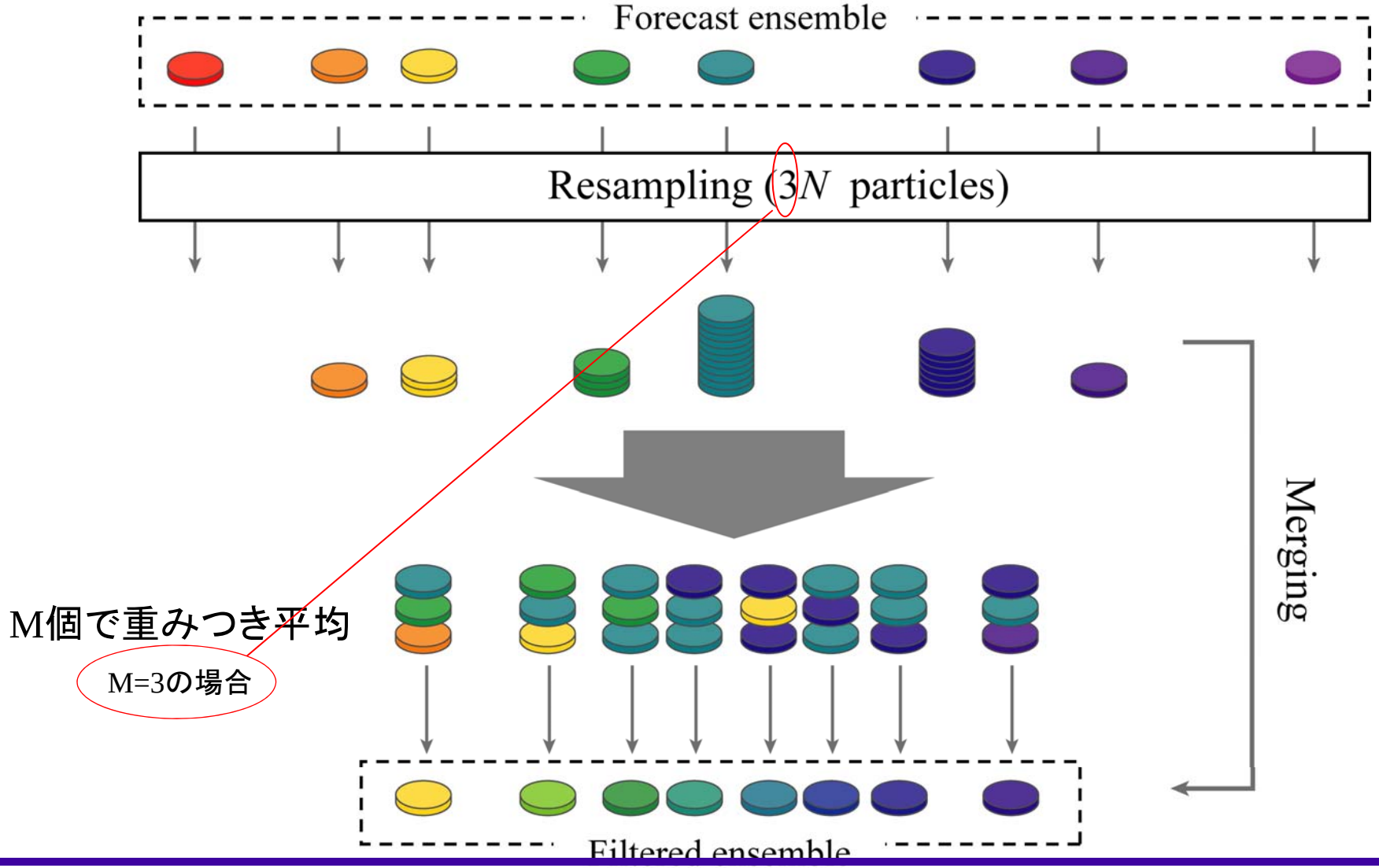
$$\theta^{(k+1)} \leftarrow \theta^{(k)}$$

Importance Sampling
MCMC

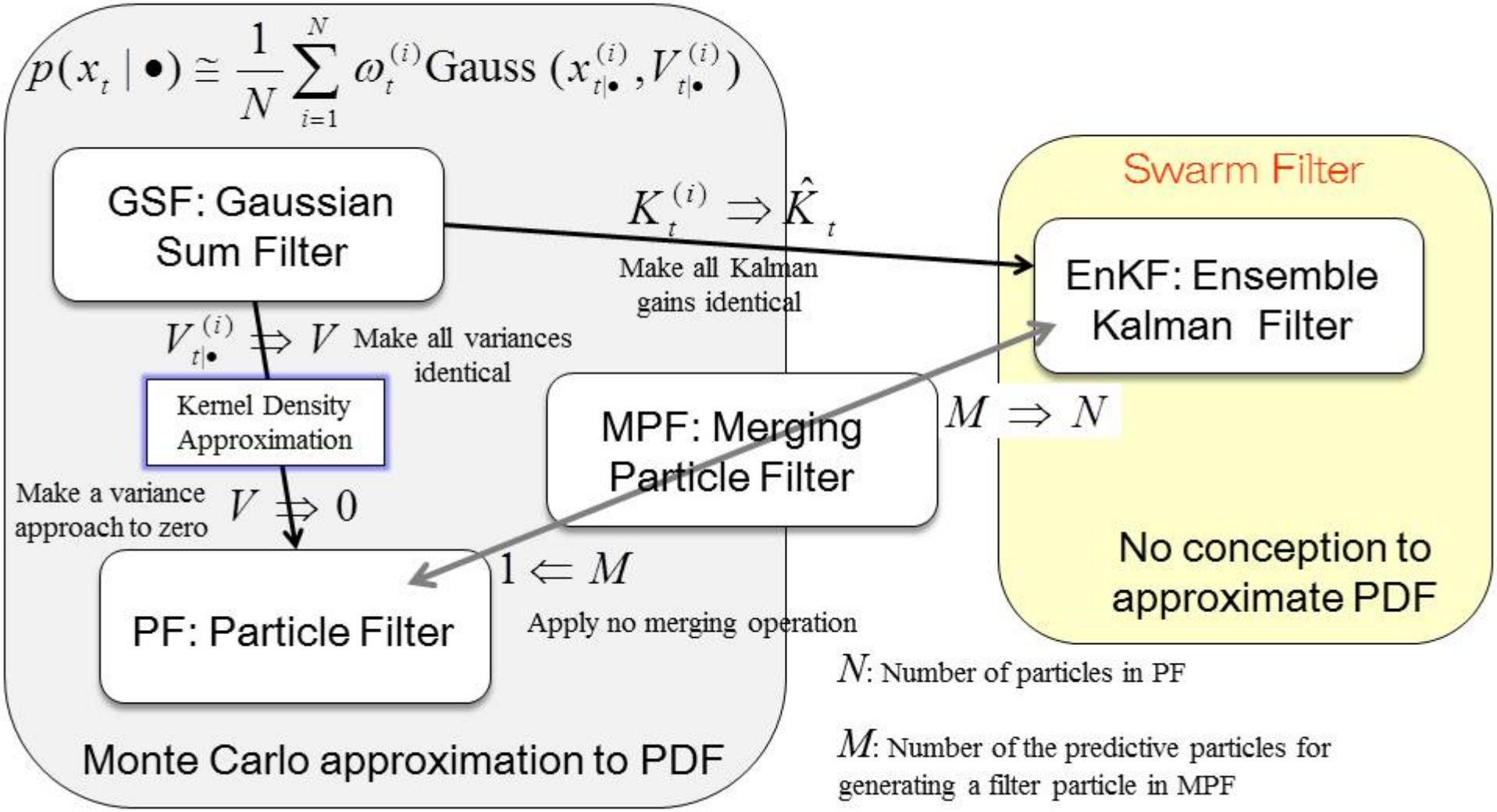
4DVar vs. アンサンブルベースデータ同化法

	非逐次型	逐次型	
	4DVar 力学的バランスを重視（システムノイズ無し）	EnKF 簡便＆安心	原型PF 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。 ・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

Merging particle filter (MPF)



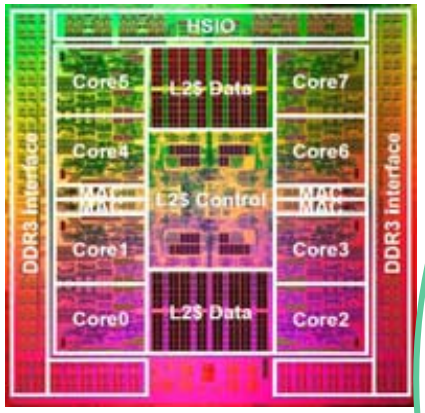
非線形フィルタリング間の関係



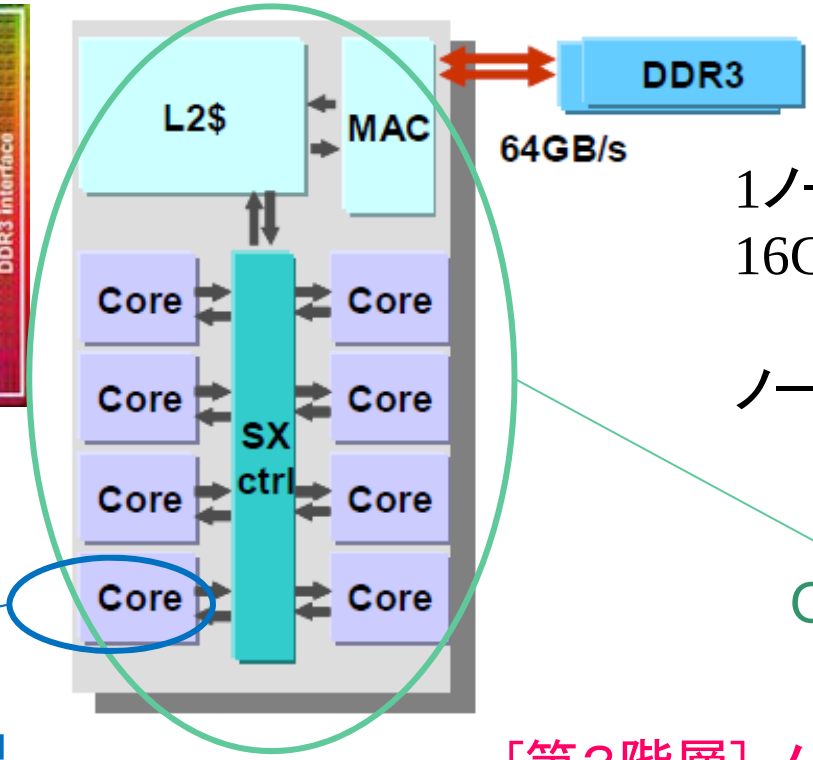
計算機への粒子フィルタの実装

- リサンプリングの際に，node間通信が頻発.
- 任意の2 nodes間でほぼat randomに通信が発生し，並列化しにくい.

①階層構造をはじめから意識



SPARC64　VIIIfx



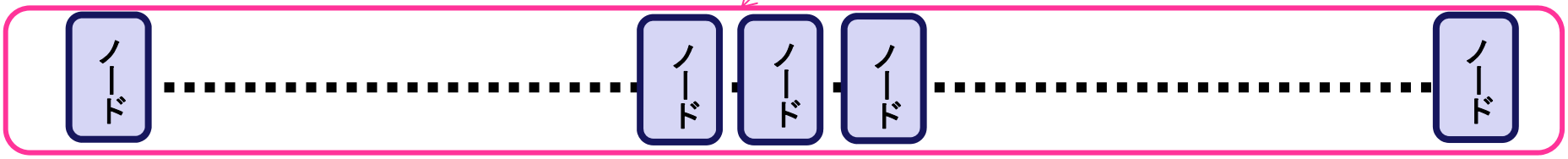
1ノード（CPU）＝　8コア
16GFlops x 8 = 128 Gflops

ノード数は8万以上

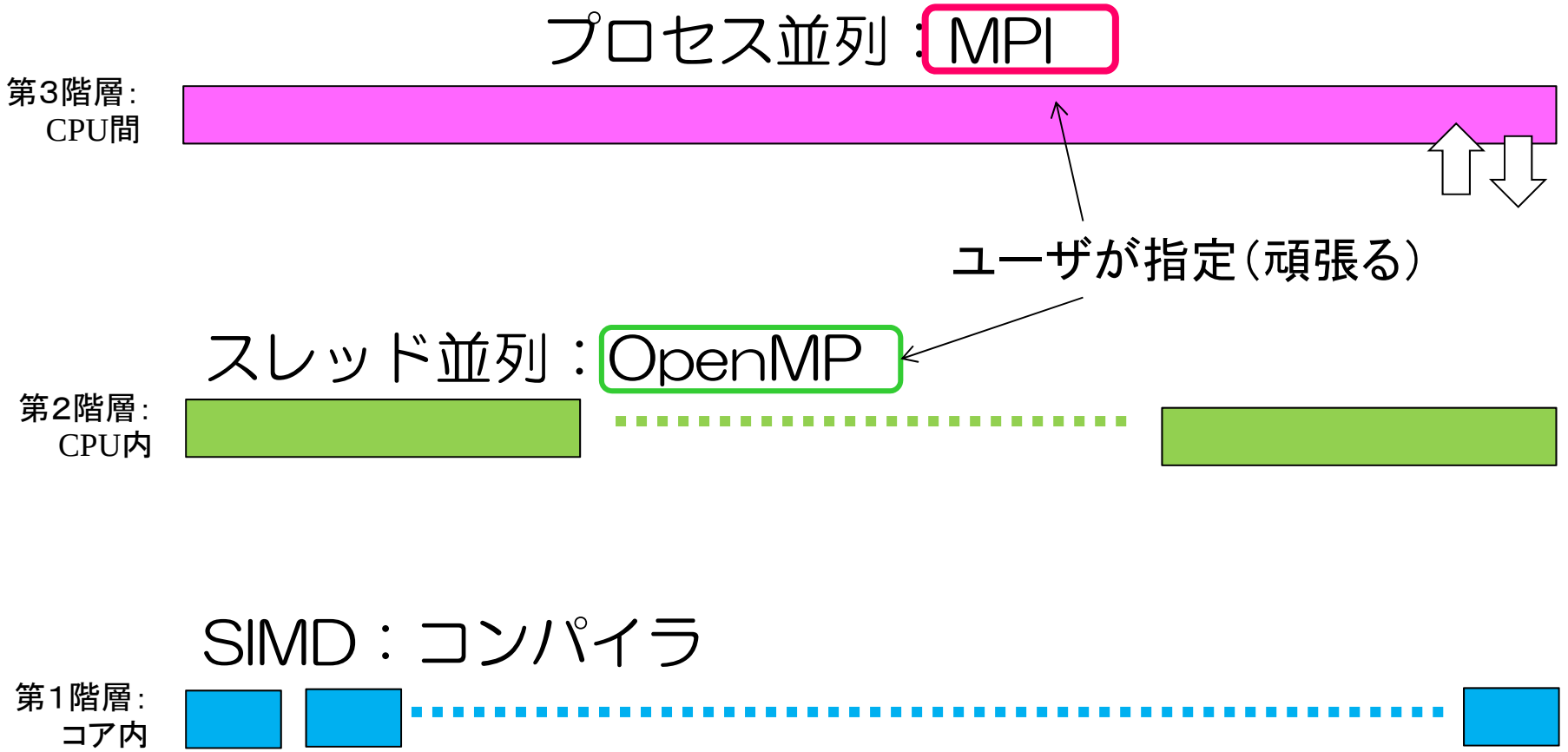
[第2階層]ノード内
OpenMP：スレッド並列

[第1階層]
コア内：SIMD化（コンパイラが対応）

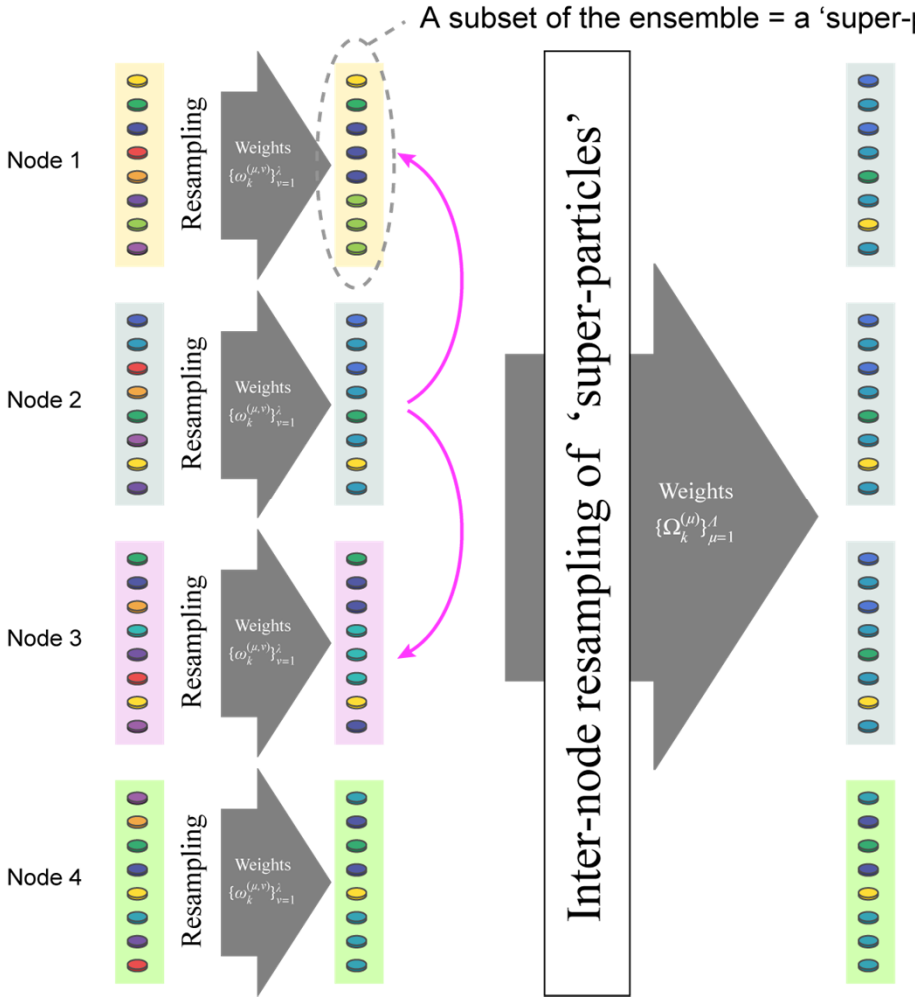
[第3階層] ノード間　MPI：プロセス並列



①階層構造をはじめから意識



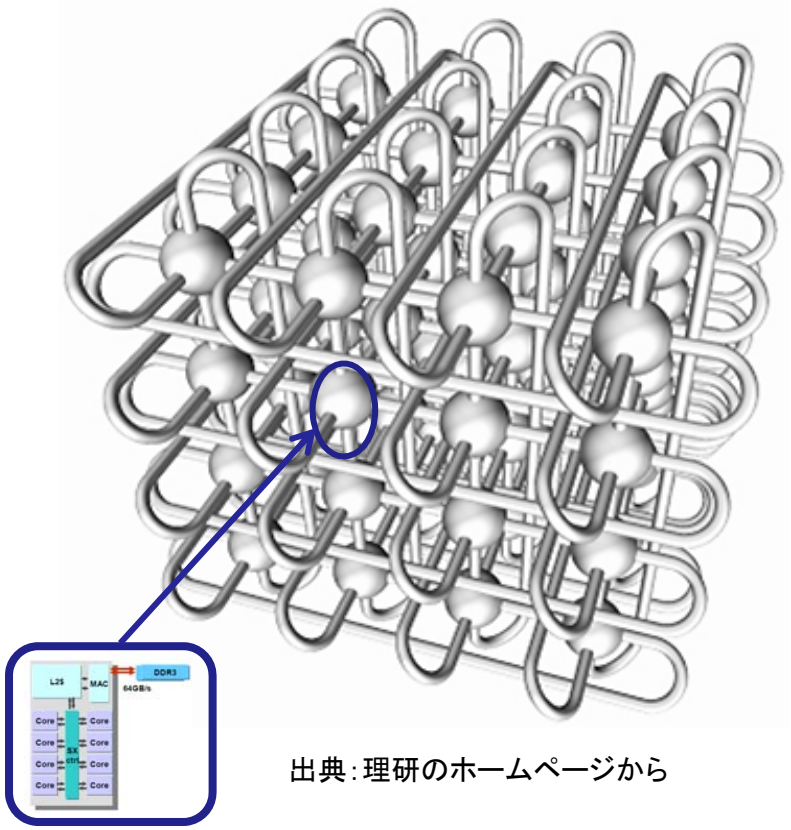
MetaPF: Meta-particle filter



Nodeに割り当てられたensemble の subset (=超粒子)をresamplingする.

現状では，nodeに割り当てられた重みがどれか1つでも0.3を超えたらresamplingすることにしている.

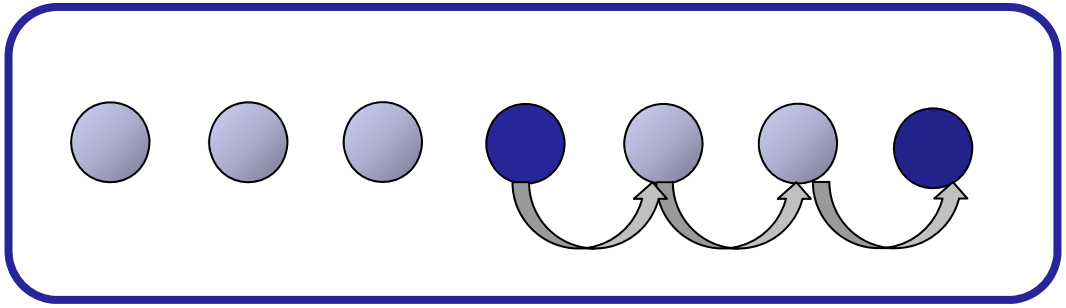
②ネットワーク構造をはじめから意識



出典：理研のホームページから

IBM Sequoia (BlueGene/Q)
Cray XT5(Jaguar)　も3次元トーラス

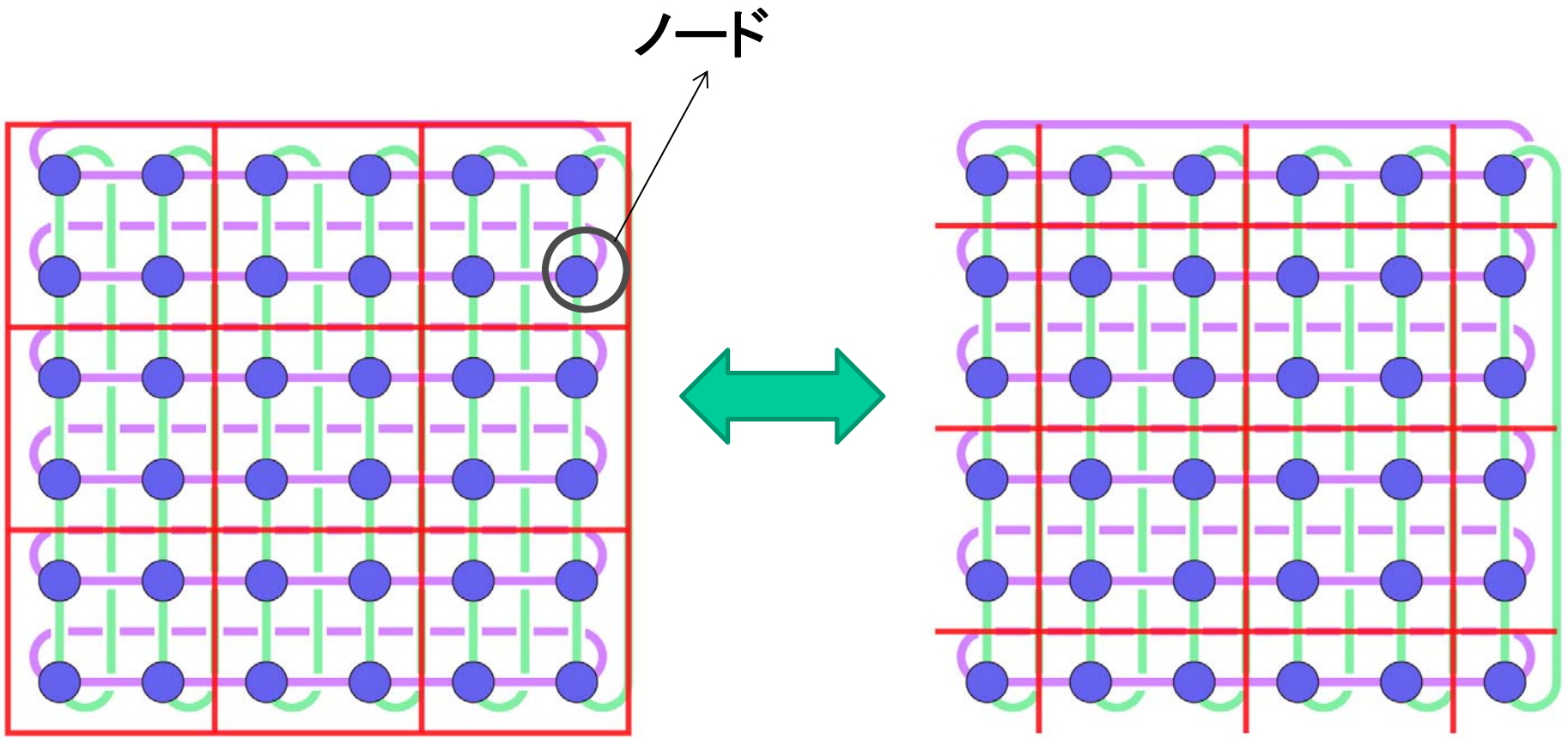
6次元メッシュトーラス：3次元の直方体に配置したノードをそれぞれ6方向で結合し、各次元がそれぞれリング状に結合されるネットワーク構成



Q. ネットワークについて、三次元トーラスで隣り合うノードどうしの通信が速いとのことですが、隣り合わないノード間の通信はその通信回数分遅くなるのでしょうか。

A. 隣り合わないノード間の通信は、基本的には**バケツリレー方式**で行われるので、**経由するノードの数が多ければ多いほど、通信時間（レイテンシ）は長く**なります。

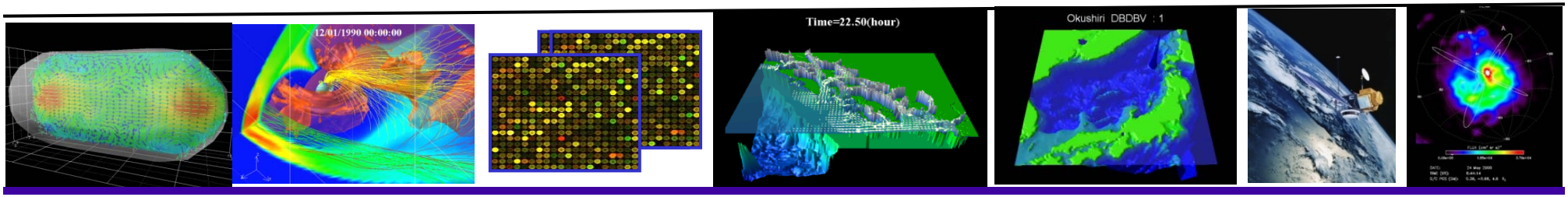
Alternate lattice-pattern switching



ここでは，上図のような2種類の格子パターンでノードをグループ化することを考え，2種類のパターンを毎回切り替えるものとする.

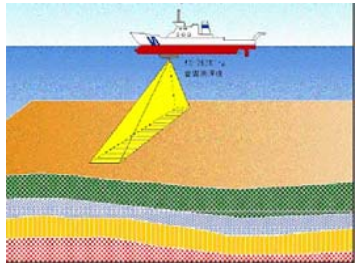
データ同化研究開発センターでの研究プロジェクト

- Coupled Ocean-Atmosphere model
- Tsunami model
- Ocean tide
- 3D structure of ring current
- Genome informatics
- Marketing (with multi-agent simulations)

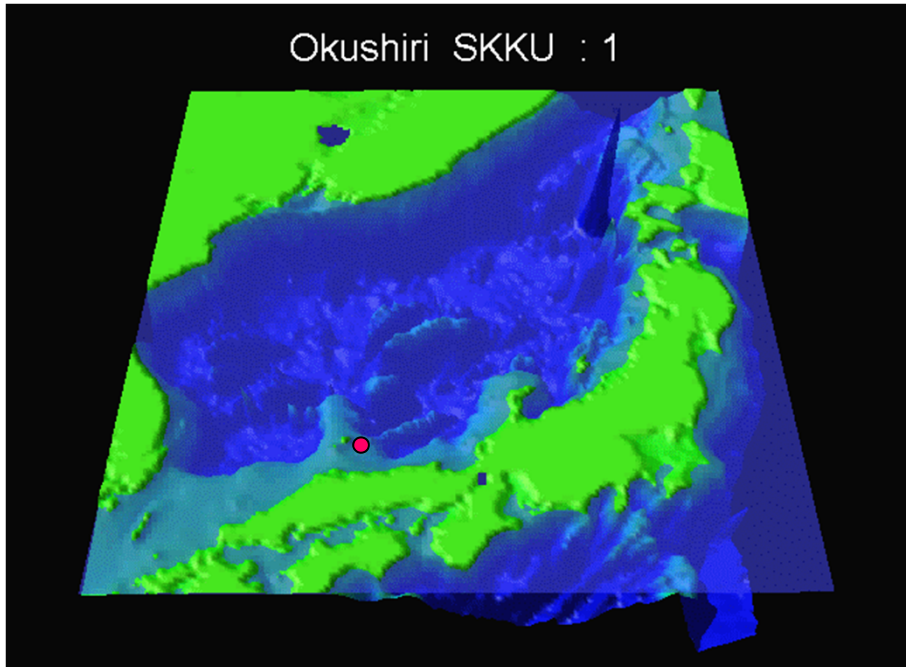


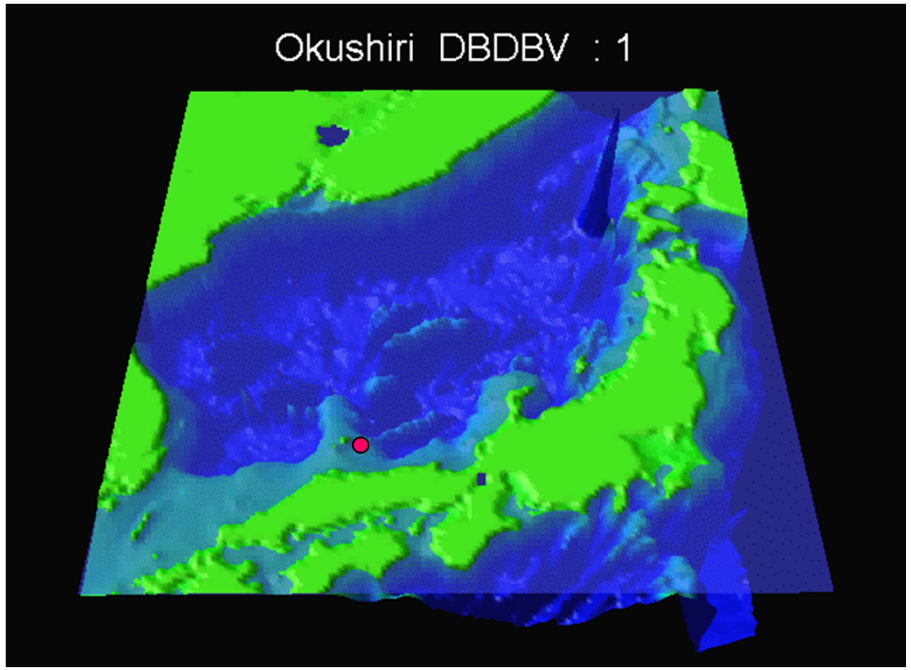
シミュレーション (同化結果でない)

- Simulation of Okushiri Tsunami
  - Simulation based on topographies made by different organizations.
  - It looks similar, but time series of sea surface displacement at a point ( ) is ...



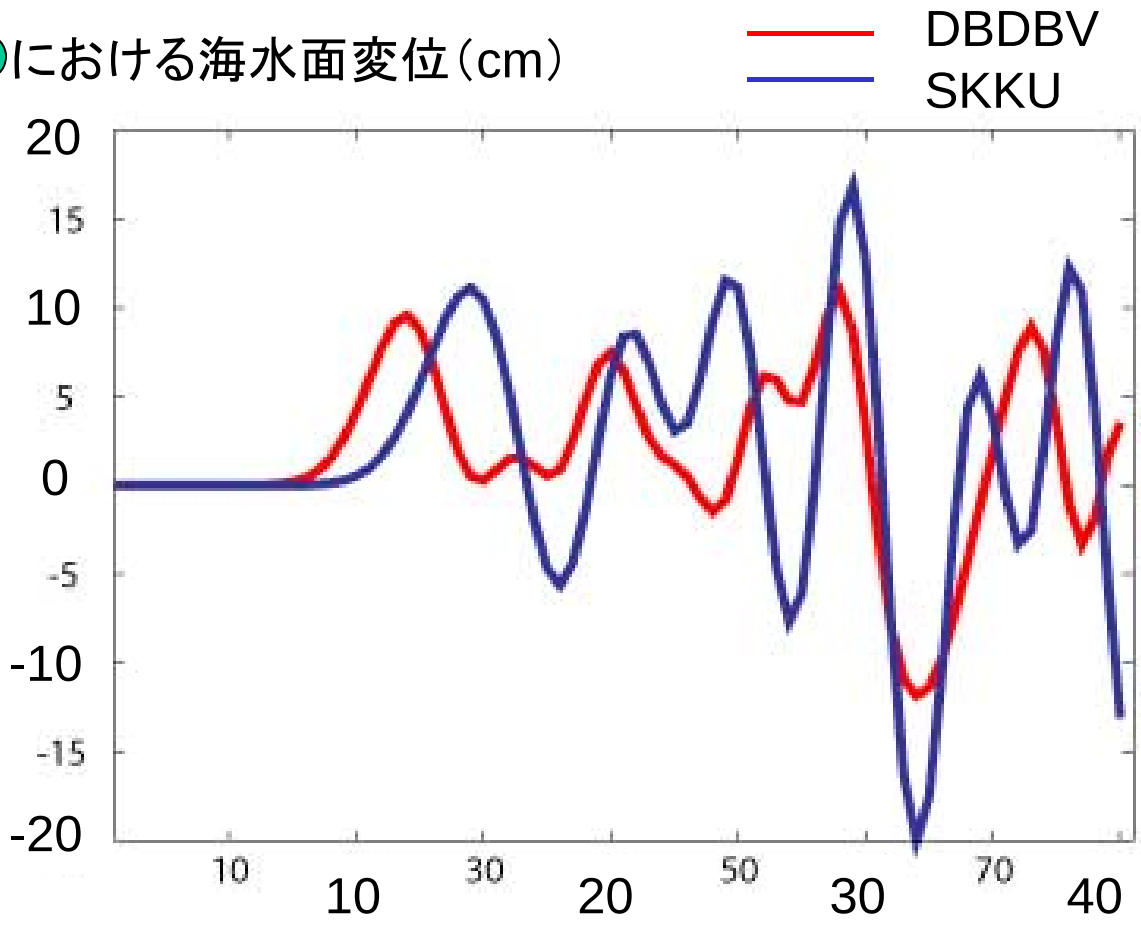
(From Japan Coast Guard)



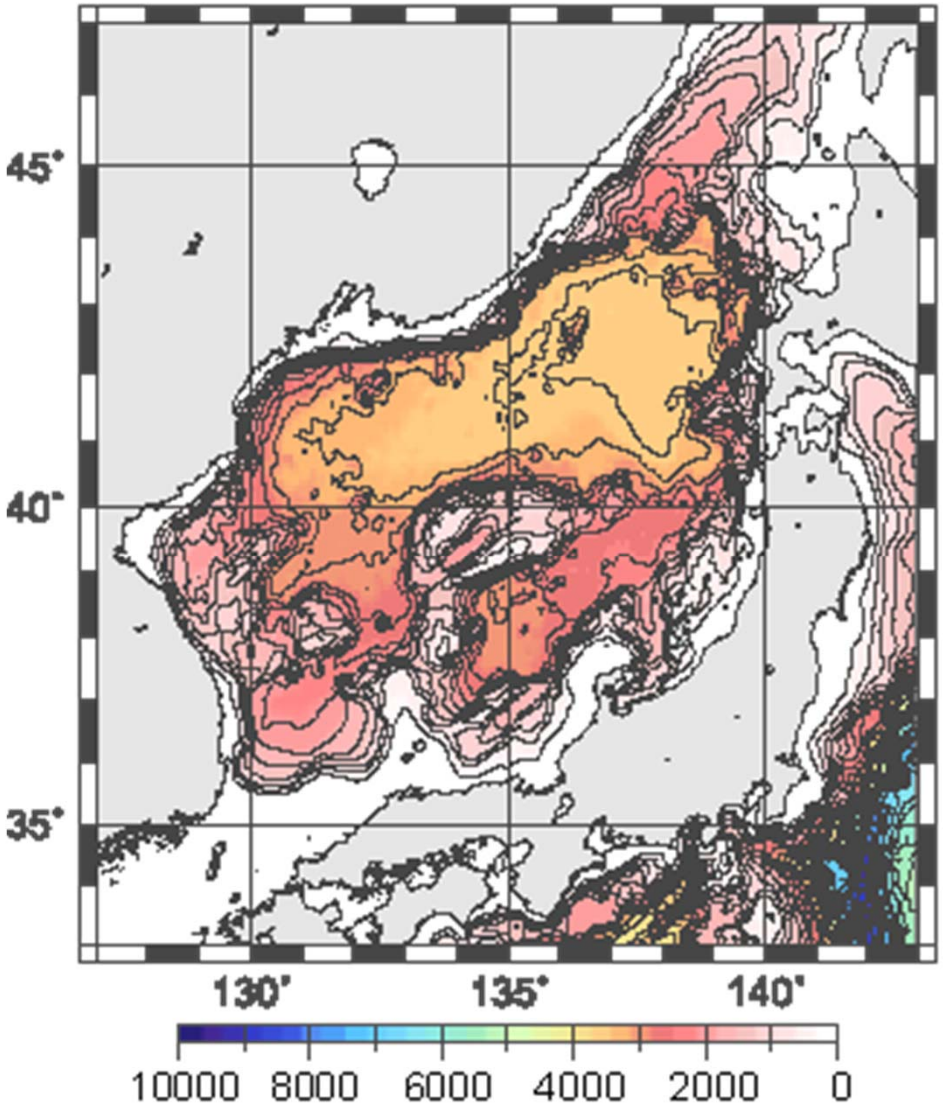


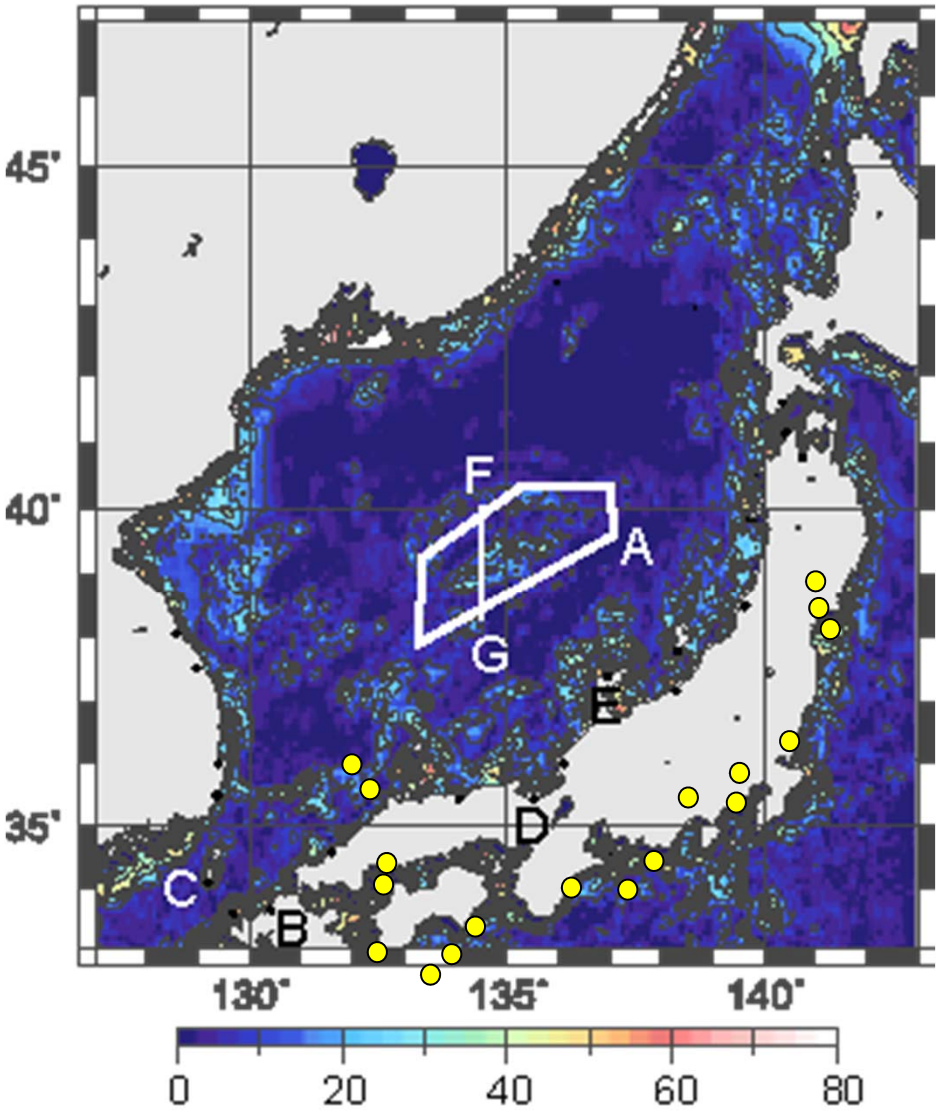
海水面変位

●における海水面変位（cm）



大きい相対的誤差：
深さに関する4つのデータベース間の比較



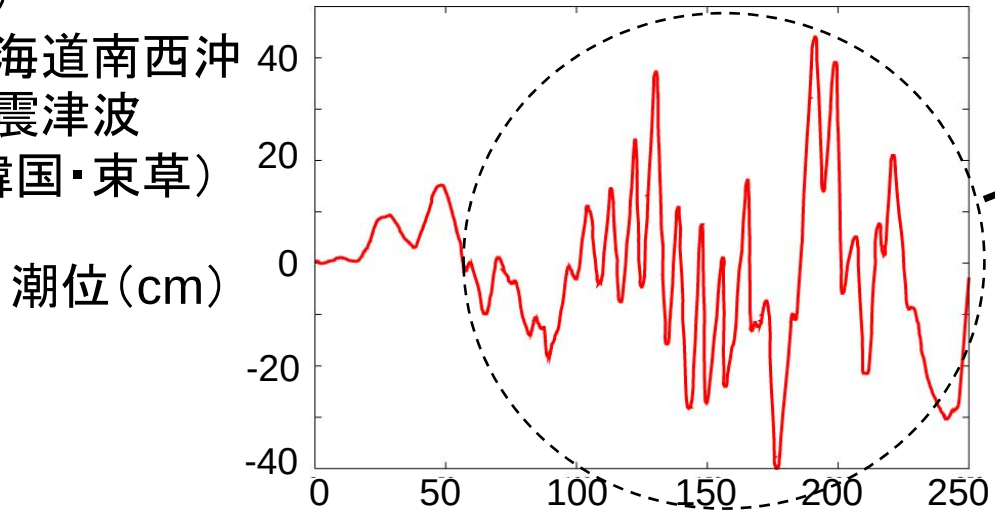


● 潮位計設置点

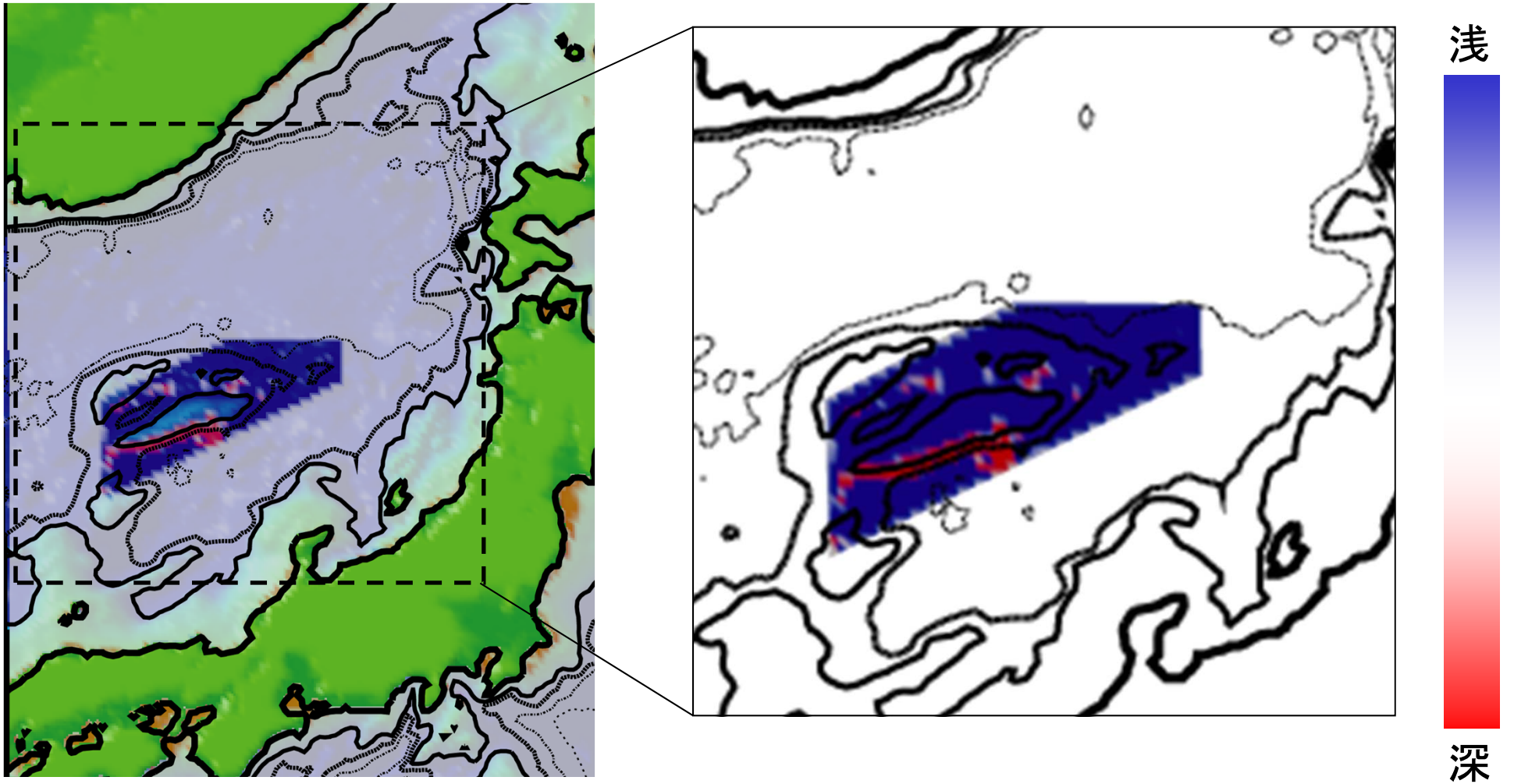
観測データ

- 潮位計データ
  - 設置地点付近の海面変位を反映
  - 得られるデータは1次元時系列
- 設置地点は右図
  - 30点程度

例）
北海道南西沖
地震津波
（韓国・束草）



同化結果



- 大和海嶺周辺は4種類の海底地形データの平均よりもやや浅いと推定される
- 南斜面に平均よりも深いと推定される部分がある

“個”にマッチしたシミュレーション：
境界条件の設定機能をパーソナライズする

“個”によって異なる形状，形態情報をシミュレーションモデルに取り込む 『メタシミュレーションモデル』

海底摩擦項

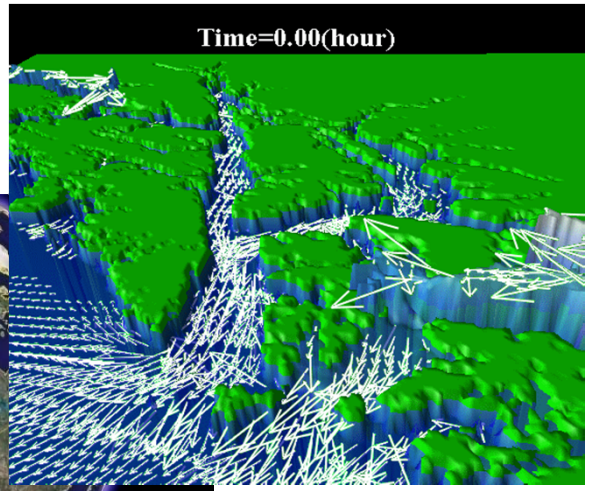
運動方程式：
$$\frac{\partial \boldsymbol{v}}{\partial t}+(\boldsymbol{v}\cdot\nabla)\boldsymbol{v}+\boldsymbol{f}\times\boldsymbol{v}=-g\nabla\eta-\gamma_b\frac{\boldsymbol{v}|\boldsymbol{v}|}{H}+A_H\nabla^2\boldsymbol{v}$$

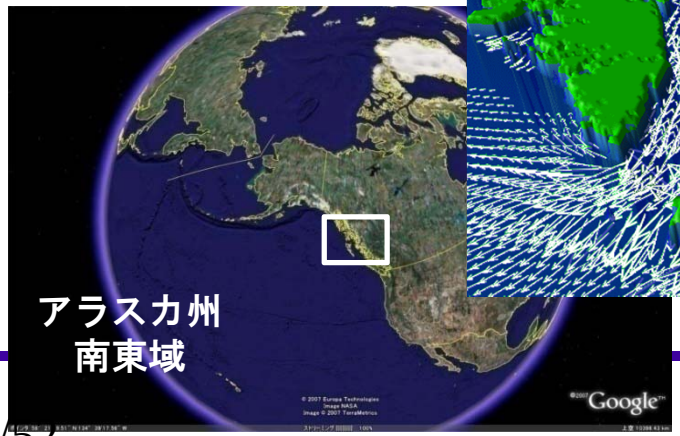
海底摩擦係数（地域依存性）

水深

連続式：
$$\frac{\partial \eta}{\partial t}+\nabla\cdot(\boldsymbol{v}H)=0$$

我々研究チームによる，
潮汐シミュレーションの例





$\mathbf{v}$：水平（2次元）流速ベクトル
η：海面水位
H：水深，$\mathbf{f}$：コリオリパラメータ

多品種少量生産を基本とする製品開発現場でのステップの簡略化や、患者一人一人に合った治療サービスの提供と期間の短縮化

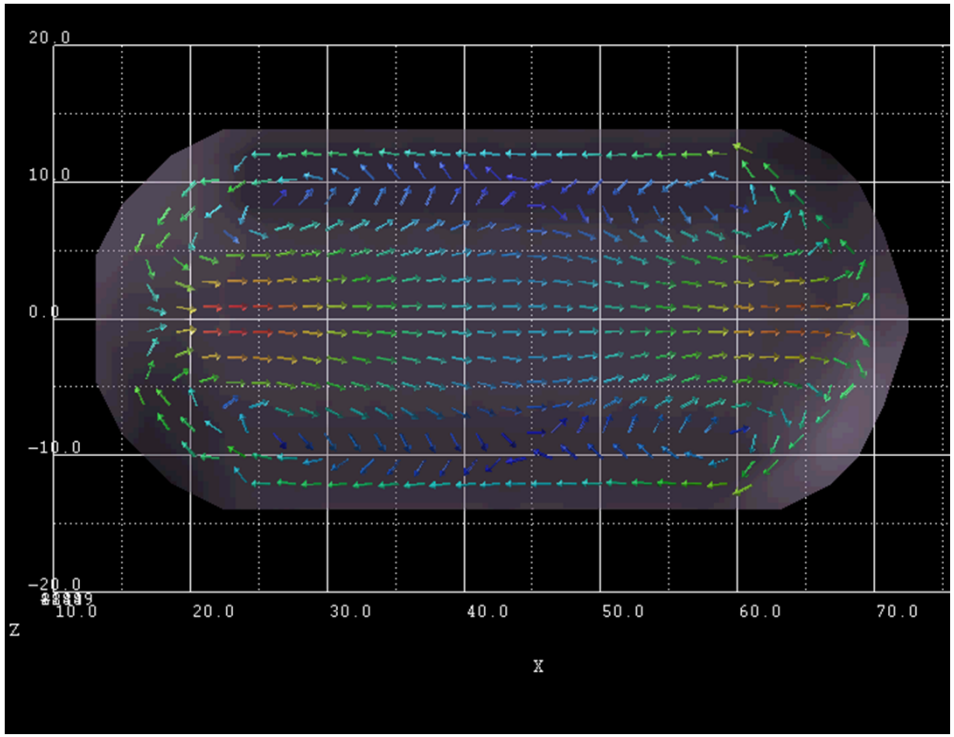
流動の原動力のパラメータ推定

国立遺伝学研究所・木村研究室と共同研究

細胞質流動（線虫）

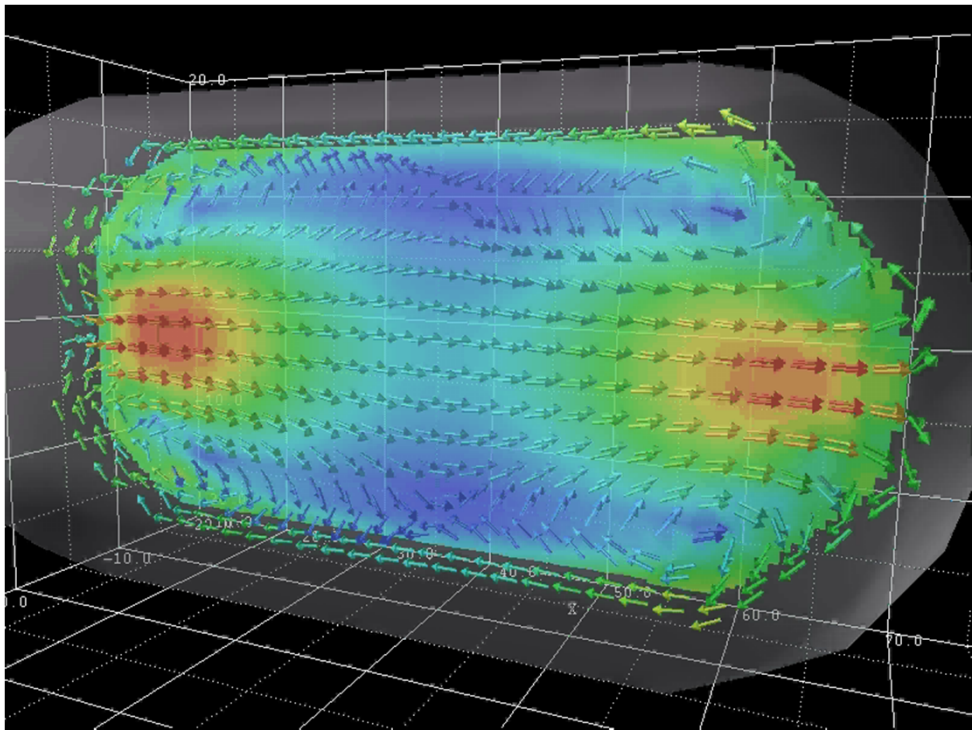
・境界条件の推定： ミオシンがつくる剪断応力値を推定

細胞質流動の定量データ



PIV（**P**article **I**mage **V**elocitmetry）計測

細胞質流動のシミュレーション



数値解法はMPS（**M**oving **P**article **s**emi-implicit）法を採用。連続体を多数の粒子で近似．計算格子を生成しなくていい．自由表面などの移動境界がある問題の扱いが簡単

東北地方太平洋沖地震（東日本大震災）に伴う地震波・地震音波伝搬シミュレーション

物理モデル

- 地球中心から地表までの固体地球および地表から高度1000kmまでの大気を考慮した1次元地球構造モデル
- 気象庁が決定した震源解（モーメントテンソル解）を長さ150kmの断層に沿って配置した断層破壊モデル

各計算グリッドにおける地震波（表面波）および音波の応答波形をノーマルモード法（Kobayashi [2007]）によって計算

グリッド幅は**緯度・経度方向0.1度**および**高度方向10km**

断層を含む緯度・経度方向10度および高度方向120kmの範囲の可視化した結果を示す.

統計数理研究所が所有するスーパーコンピュータ（富士通Dシステム）を約4000ノード時間（512並列で約60時間）使用

津波よりも伝搬速度が速い音波を利用した将来型津波警報システムの構築

オーダーメイド医療・創薬に向けたPersonalized Simulation

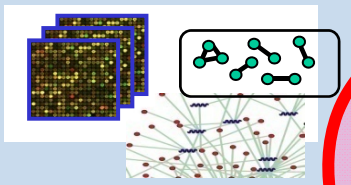
東京大学・医科学研究所・宮野研究室と共同研究

データ同化

- 生体システムを統合的に理解・予測するための有望な方法論
- 実験科学と統計・シミュレーション科学の融合

10万コアの超並列計算に成功！

Order-Made Medicine, Drug discovery with Personalized Simulation



DATA

データ同化

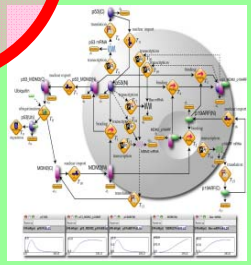
Validation

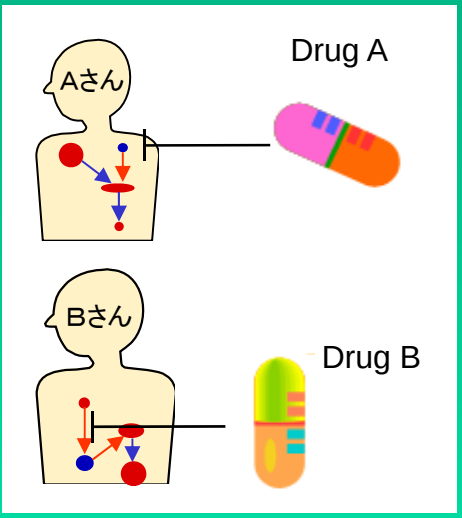
Parameter Estimation
Model Selection
Super Computer

Experiments

Design

Hypothesis Creation





 本スライドは東京大学（医科研）・山口類氏から借用，改変

“ものづくり”分野：

地盤シミュレーションのデータ同化

京都大学（農学研究科）・村上教授グループとの共同研究

■実学的観点から： 地盤沈下の最終沈下量の予測

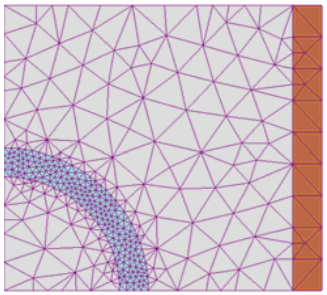
・関西国際空港の埋め立てにおける沈下量予測など、途中までの計測を用いて最終沈下量の予測をしたい。

・データ同化の観点からは，シミュレーション予報のための初期値生成作業。

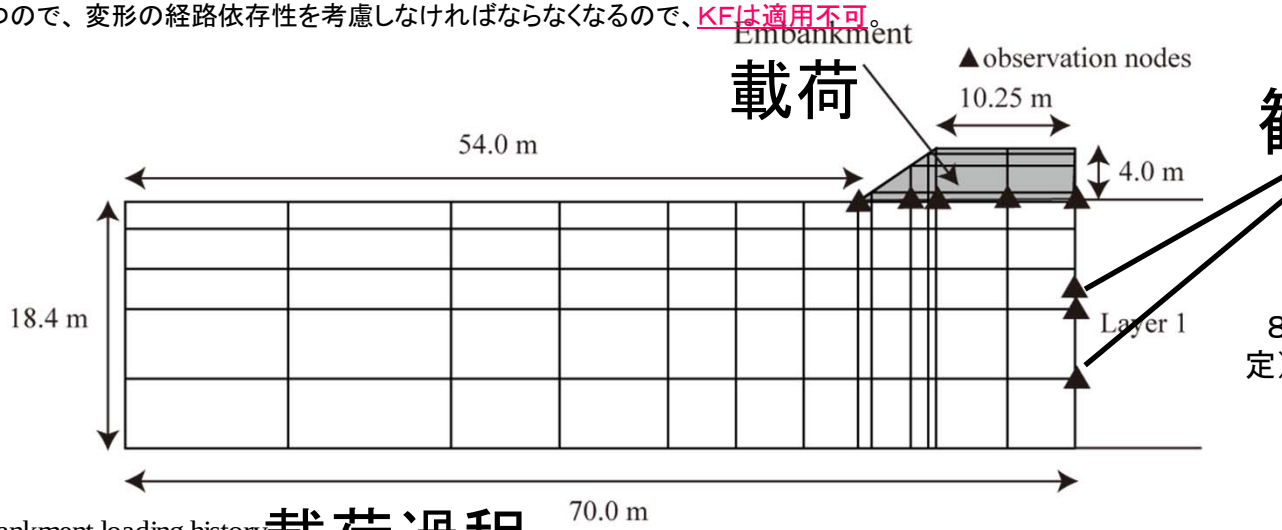
■計算手法の観点から： 弾性体であれば，線形性・復元性によりデータ同化は簡単。（カルマンフィルタ(KF)が有効。 実際の土は塑性をもつので、変形の経路依存性を考慮しなければならなくなるので、KFは適用不可。

構造シミュレーション：

有限要素法、境界要素法

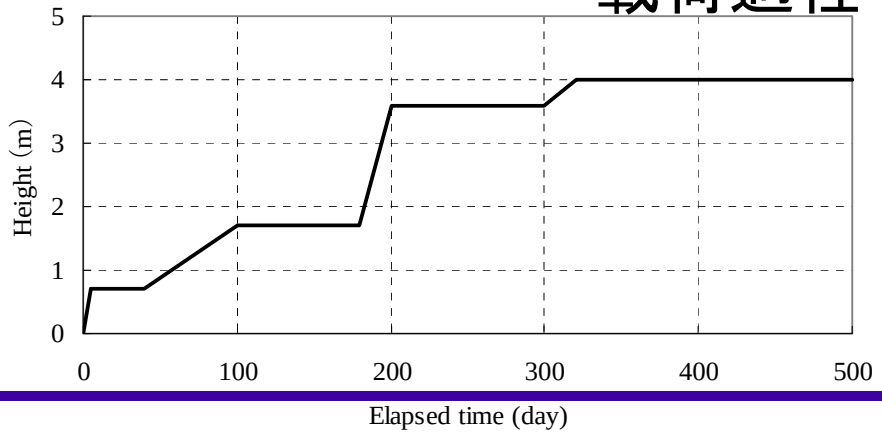


(Wikipediaから)

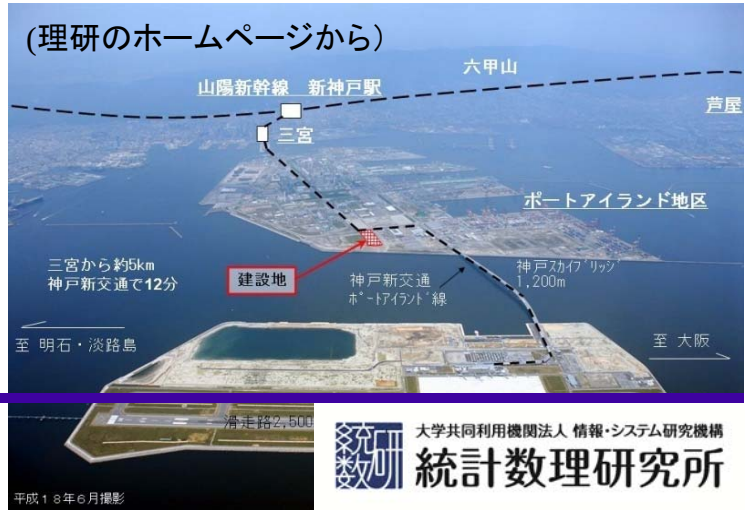


８点の観測（沈下量測定）を用いて，弾塑性パラメータを推定する

Embankment loading history 載荷過程



・塑性変形も加わると，非線形性に伴うシミュレーション計算の破綻（保存則を満たさない等）の問題が出てきてしまい，KF では難しかったが SIS（逐次データ同化の一種）で解決できた。



富を産む仕組み

前世紀：　物質（「もの」）を均質に大量に生産するシステム

↓

21世紀：　個人化された情報サービスを提供するシステム

個人をターゲットにした商品・サービスの提供を効率的に行えるシステム

“コ”ー個人，個性，個別，固有ーが大切！

ご静聴ありがとうございました。

バーチャルにリアルを埋め込むデータ同化：
ベイズ、アンサンブル予測
ヒューマンモデリング、個人化情報サービス

(2011年4月刊行)

(2011年9月刊行)

(2011年11月刊行)

データ同化研究開発センターの紹介ビデオ（10分）
を作成しました。YouTubeにアップしています。