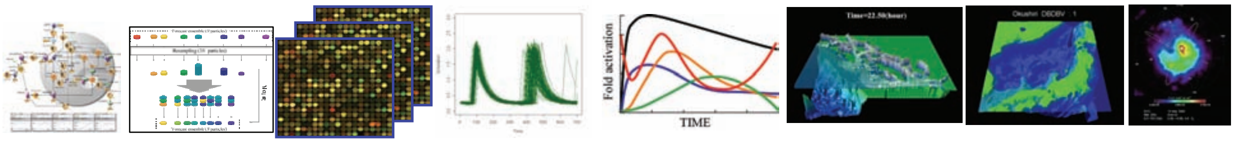
2010/01/25 10:10—11:10 The 3rd Workshop on Integration of EFD and CFD

Introduction to Sequential Data assimilation methods: Their mathematical basis and recent development

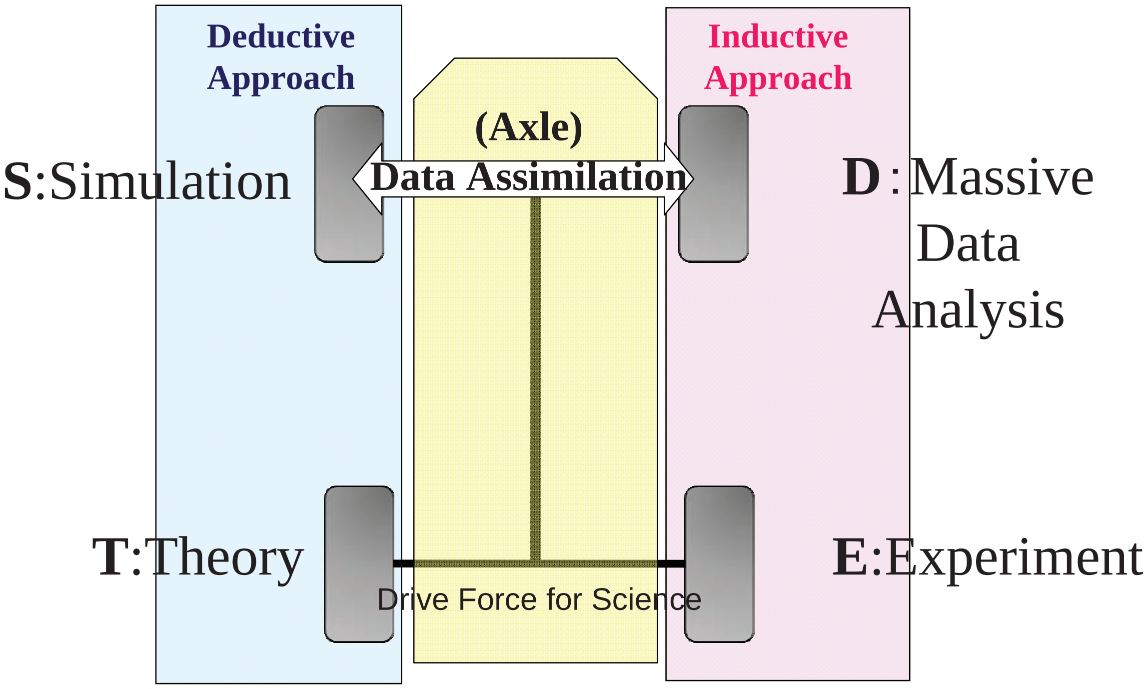
Tomoyuki **Higuchi**
Research Organization of Information and Systems
The Institute of Statistical Mathematics/JST CREST



独立行政法人 科学技術振興機構

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

TESD: Four Kinds of Methodology of Science



Deductive Approach

Inductive Approach

(Axle)

Data Assimilation

S:Simulation

D:Massive Data Analysis

T:Theory

E:Experiment

Drive Force for Science

Red color indicates a slide used in the last year's presentation (2nd. Workshop on Integration of EFD and CFD)

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

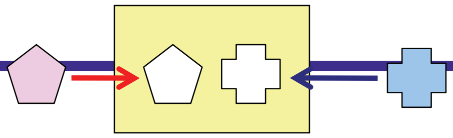
What is Data Assimilation?

- Emerging subject in meteorology and oceanography.
- Methodology to synthesize numerical simulation model and observed data
  - Simulation model can not reflect real physics accurately.
    - （e.g.）Accurate weather forecast needs good initial conditions.
    - Uncertainty in the model (boundary condition, initial condition, unknown parameters, unknown dynamics...) exists.
  - Observation data have some physical/budgetary restrictions.

➡ Correct variables in numerical simulation model using observation data. = Data Assimilation

Simulation model

Observation data



Objects of Data Assimilation from a viewpoint of Meteorology and Oceanography

[1] To produce the best (better) **initial condition** for forecasting. It is actually realized in the real weather forecast (ex., Japan Meteorological Agency).

[2] To find the best (better) **boundary condition** in constructing a simulation model. This procedure includes a setting of appropriate boundary conditions necessary for dealing with a coupled phenomena.

[3] To attain an optimal **parameter** vector that appears in an empirical law (scheme) employed for describing complicated phenomena with the different time and spatial scales. A **validation** of the empirically given values is regarded as this problem.

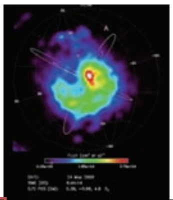
[4] To inter/extrapolate (estimate) an physical quantity at times and locations without observations based on a numerical simulation model. This procedure is called “a **generation of re-analysis dataset** (product)”. This dataset is used to discover a new scientific findings by general geophysical researchers.

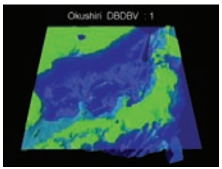
[5] To conduct an experiment with a virtual observation network and perform a **sensitivity analysis** in an attempt to construct an effective observation network system with less budgetary cost and less consuming time.

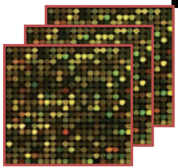
(ex. Kamachi et al., 2006)

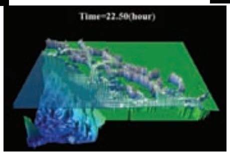
Outline

- Mathematical basis and Bayesian computation
- Sequential data assimilation
- Ensemble-based nonlinear filtering method
  —Particle filter (PF)
- Advanced methods for PF
  —Merging PF
  —Meta PF
  —PF with GPGPU
- Conclusions









Construction of Simulation Model

（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)

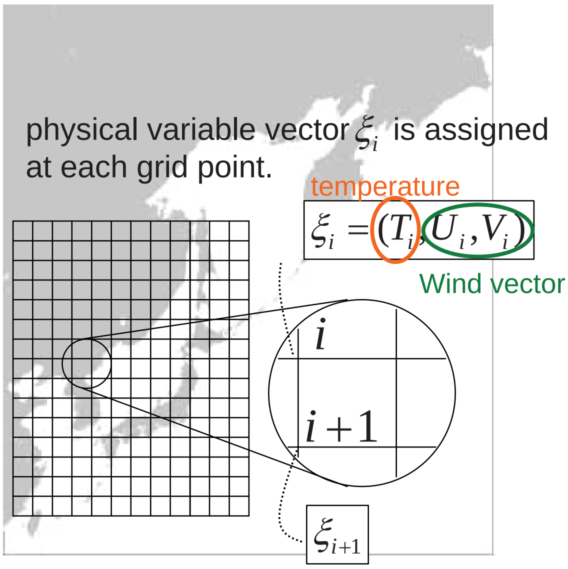
$$x_t = F(x_{t-1})$$

(time varying)
Boundary conditions

v_t

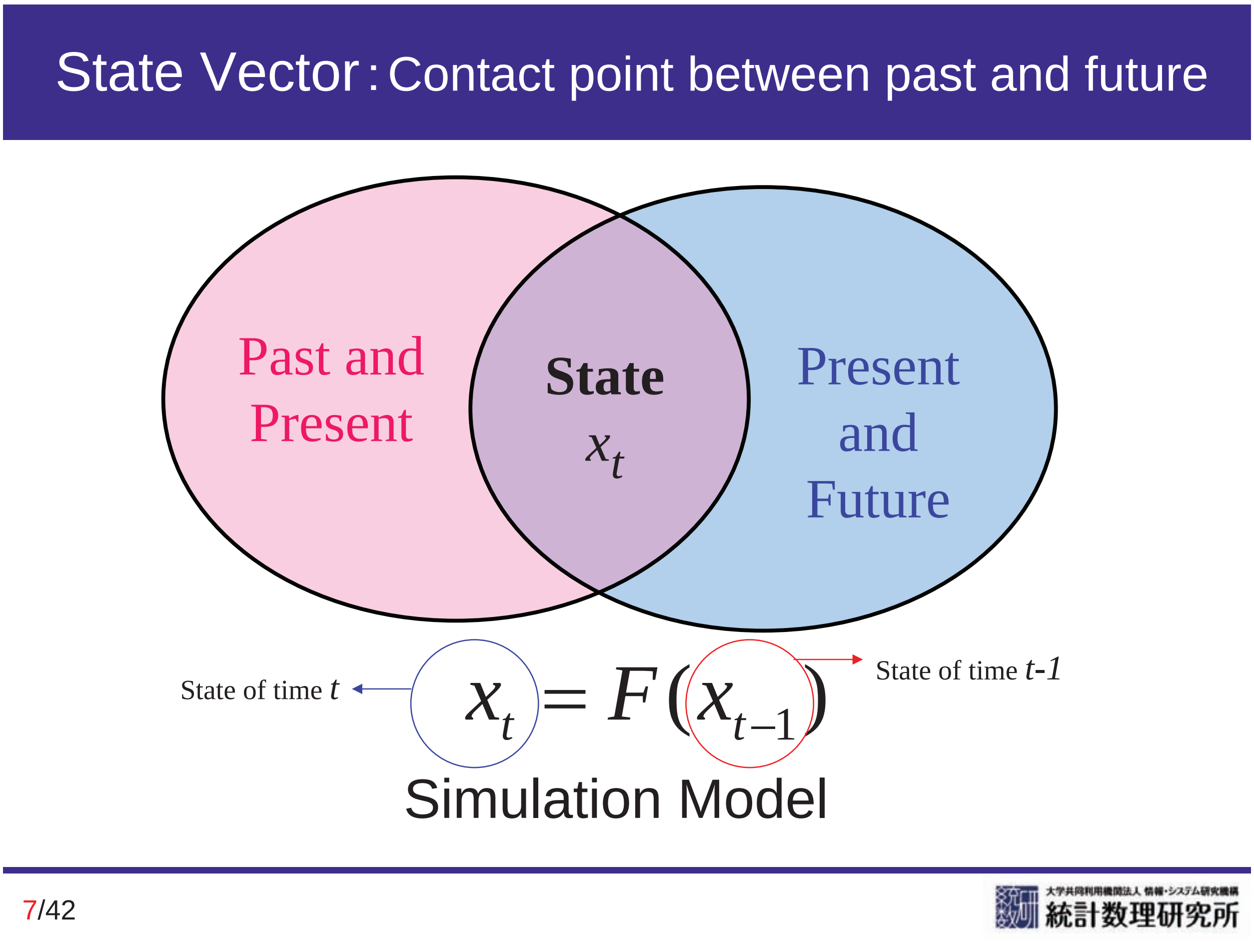
Nonlinear state space model
(discrete time/space, stochastic(SDE))

$$x_t = F(x_{t-1}, v_t)$$



physical variable vector ξ_i is assigned at each grid point.

Observation points and observed variables are limited.



State Vector：Contact point between past and future

Past and Present

State

x_t

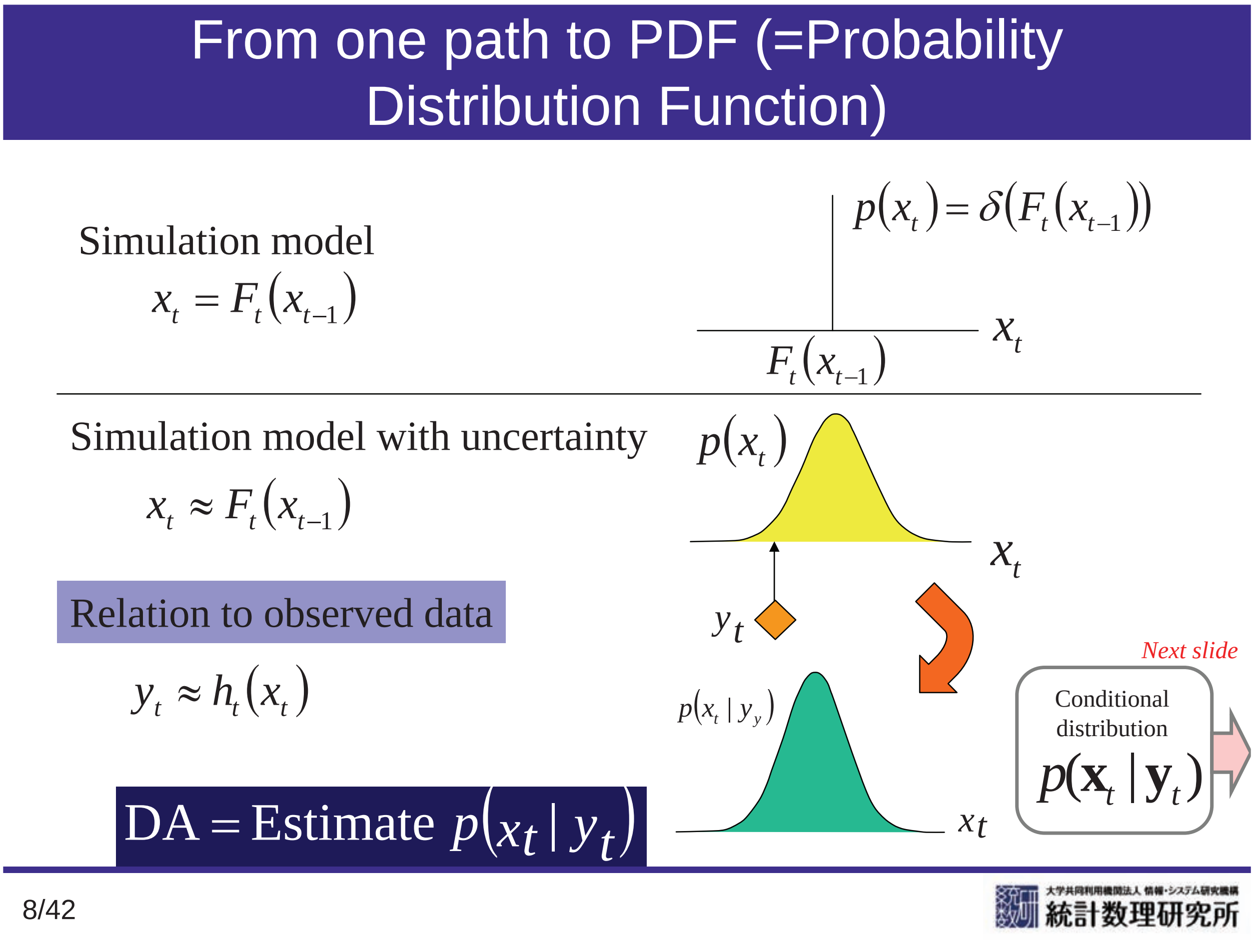
Present and Future

State of time t

$$x_t = F(x_{t-1})$$

State of time t-1

Simulation Model



From one path to PDF (=Probability Distribution Function)

Simulation model

$$x_t = F_t(x_{t-1})$$

$$p(x_t) = \delta(F_t(x_{t-1}))$$

Simulation model with uncertainty

$$x_t \approx F_t(x_{t-1})$$

Relation to observed data

$$y_t \approx h_t(x_t)$$

$$\text{DA} = \text{Estimate } p(x_t \mid y_t)$$

Next slide

Conditional distribution

$$p(\mathbf{x}_t \mid \mathbf{y}_t)$$

Conditional and Joint Probabilities

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

$p(\mathrm{A}\,|\,\mathrm{B}) \equiv$ Probability of **B** given **A** ← Conditional Probability

$p(\mathrm{A},\,\mathrm{B}) \equiv$ Probability of **A** and **B** ← Joint Probability

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

Total: 100

# of consumers to buy a bag of coffee grounds : 30

# of consumers to buy milk: 60

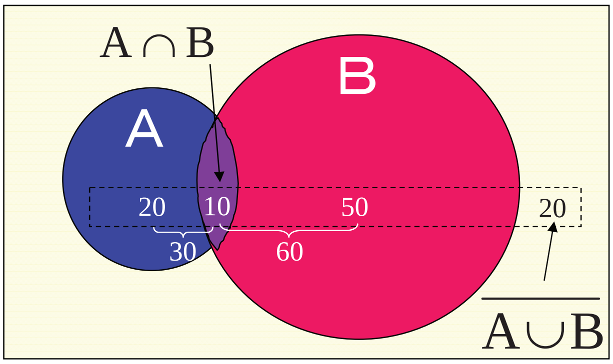
# of consumers to buy a coffee bag and milk: 10

A=1: Buy a bag of coffee grounds

=0: Not buy

B=1: Buy milk

=0: Not buy



Marginalization

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

Joint probability

$$p(\mathrm{A}=i,\mathrm{B}=j)$$

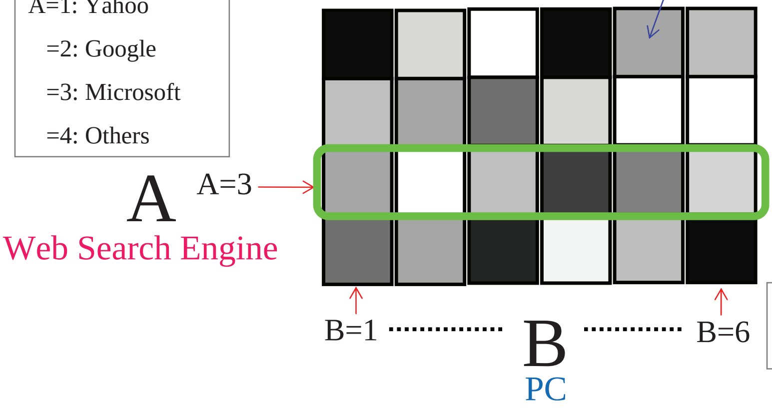
$$\sum_{i=1}^{4}\sum_{j=1}^{6} p(\mathrm{A}=i,\mathrm{B}=j)=1$$

A=1: Yahoo

=2: Google

=3: Microsoft

=4: Others



B=1: NEC, =2: Fujitsu, =3: Dell

=4: Toshiba, =5: Apple , =6: Others

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

Bayes’ Theorem

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

It is **easy** to calculate.

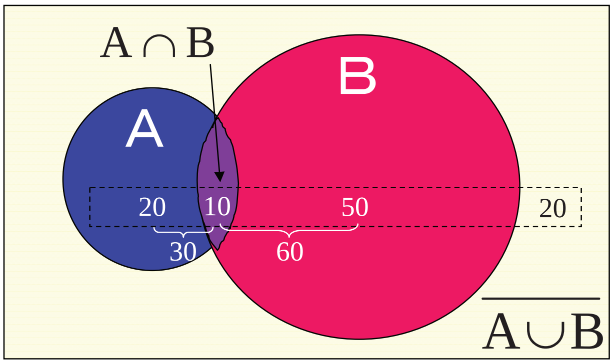
$$\boxed{p(\underline{\mathrm{A}}\,|\,\underline{\mathrm{B}}) = \frac{p(\mathrm{A},\mathrm{B})}{p(\mathrm{B})} = \frac{p(\mathrm{A},\mathrm{B})}{\sum_{\mathrm{A}\in\text{Aのとる可能性}} p(\mathrm{A},\mathrm{B})} = \frac{p(\underline{\mathrm{B}}\,|\,\underline{\mathrm{A}})\,p(\mathrm{A})}{\sum p(\underline{\mathrm{B}}\,|\,\underline{\mathrm{A}})\,p(\mathrm{A})}}$$

A=1: Buy a bag of coffee grounds (Search Engine type)

B=1: Buy milk (PC type)

$$p(\mathrm{A}=1\,|\,\mathrm{B}=1) = \frac{p(\mathrm{B}=1\,|\,\mathrm{A}=1)p(\mathrm{A}=1)}{p(\mathrm{B}=1\,|\,\mathrm{A}=1)p(\mathrm{A}=1)+p(\mathrm{B}=1\,|\,\mathrm{A}=0)p(\mathrm{A}=0)}$$

$$= \frac{\frac{10}{30}\cdot\frac{30}{100}}{\frac{10}{30}\cdot\frac{30}{100}+\frac{50}{70}\cdot\frac{70}{100}} = \frac{10}{10+50}$$



Generative Model, Inversion with Bayes’ theorem, and Data Assimilation

Data dist.（likelihood function）

Posterior dist.

Bayes’ Theorem

$$p(\underline{x}\,|\,y) \equiv \frac{p(y\,|\,x)\cdot p(\underline{x})}{p(y)} \propto p(y\,|\,x)\cdot p(x) = p(x,y)$$

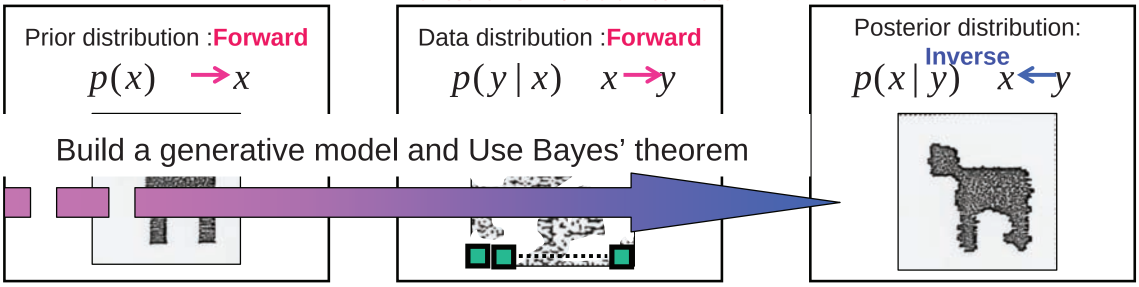
x : simulation variables

y : data

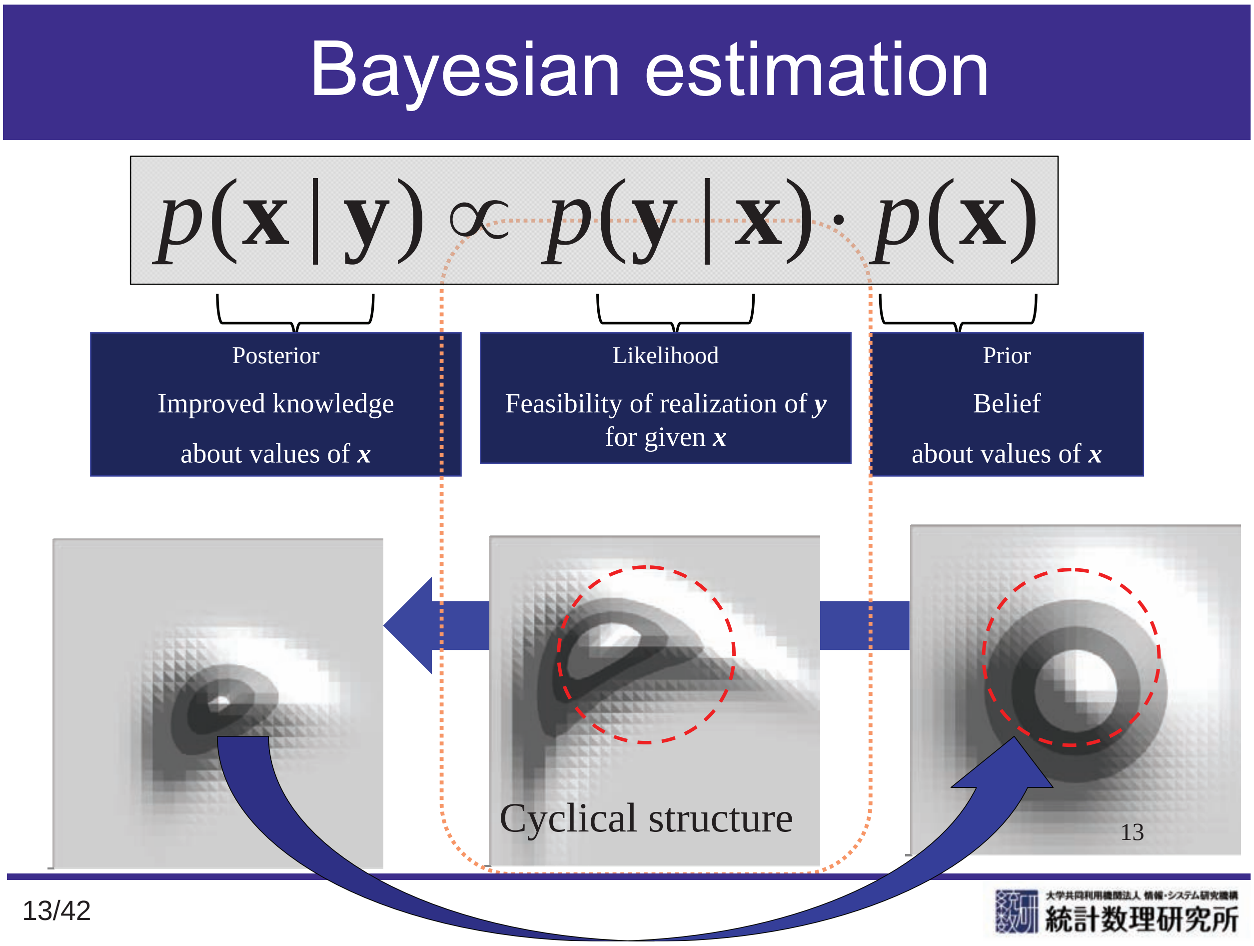
Prior dist.

Simulation

Fitness of Simulation to Data



Latent Variables : x



Bayesian estimation

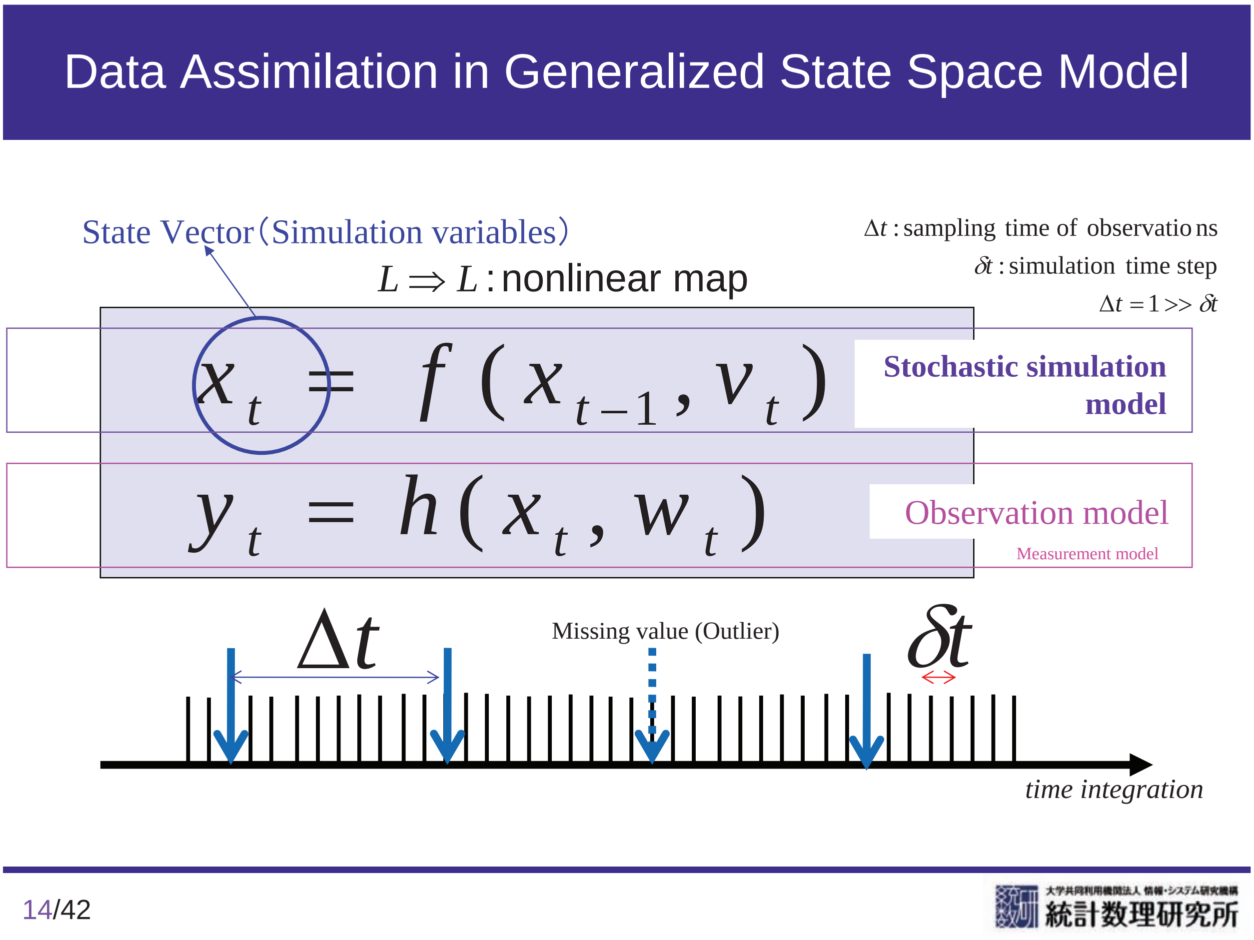
$$p(\mathbf{x}\,|\,\mathbf{y}) \propto p(\mathbf{y}\,|\,\mathbf{x}) \cdot p(\mathbf{x})$$

Posterior
Improved knowledge
about values of *x*

Likelihood
Feasibility of realization of *y*
for given *x*

Prior
Belief
about values of *x*

Cyclical structure



Data Assimilation in Generalized State Space Model

State Vector（Simulation variables）

$L \Rightarrow L$: nonlinear map

Δt : sampling time of observations
δt : simulation time step
$\Delta t = 1 >> \delta t$

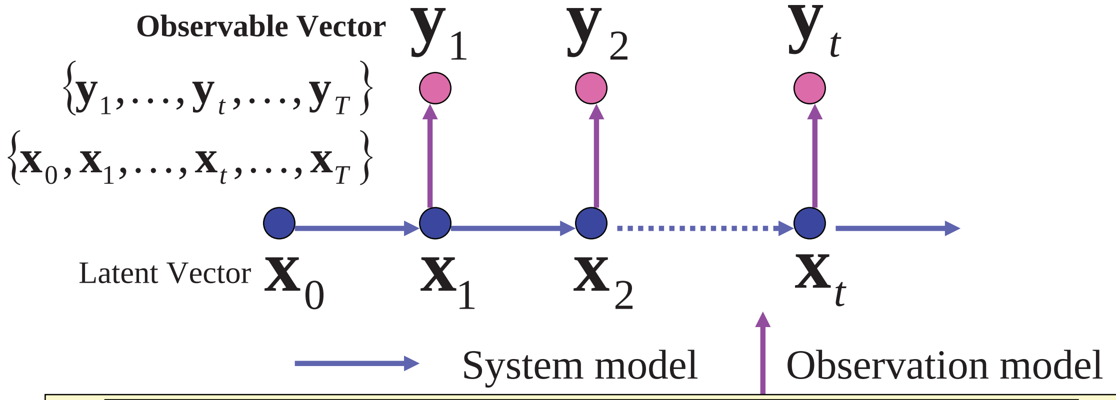
$$x_t = f(x_{t-1}, v_t)$$

Stochastic simulation model

$$y_t = h(x_t, w_t)$$

Observation model
Measurement model

Chain Structure Graphical Model



Suppose a statistical inference problem on a daily economic status given daily stock market data

Today's economic situation given yesterday's stock market data

$$p(\mathbf{x}_t \mid \mathbf{y}_{1:t-1} \equiv [y_1, y_2, \ldots, y_{t-1}])$$

Today's economic situation estimated by the stock market data up to today

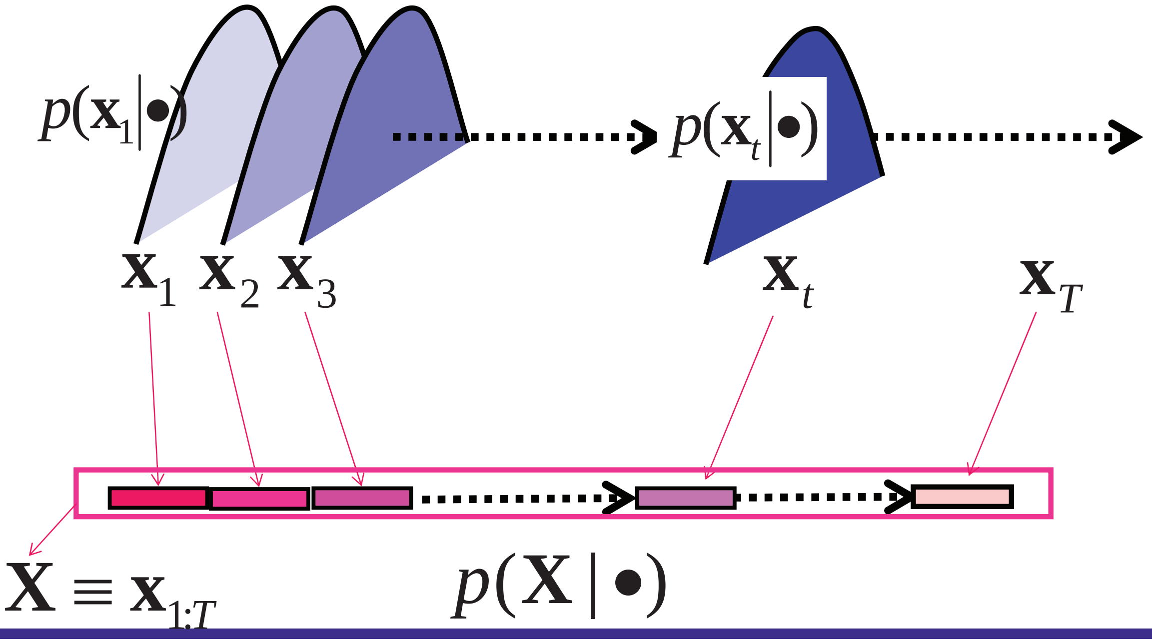
$$p(\mathbf{x}_t \mid \mathbf{y}_{1:t} \equiv [y_1, y_2, \ldots, y_t])$$

Today's economic situation analyzed by using all available data when we look back on the today in future

$$p(\mathbf{x}_t \mid \mathbf{y}_{1:T} \equiv [y_1, y_2, \ldots, y_T])$$

$$p(\mathbf{x}_{1:T} \mid \mathbf{y}_{1:T})$$

State Vector and Concatenated State Vectors



Two ways of DA method

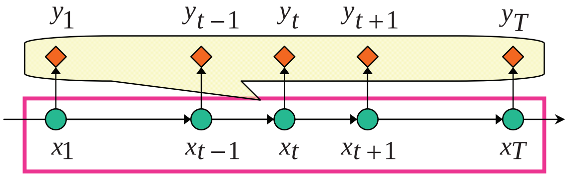
Find **X**=[x_1, …, x_T] that maximize

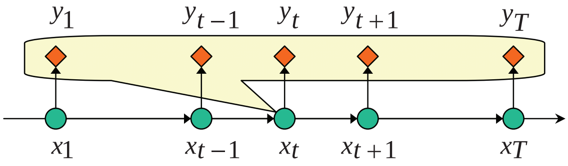
$$p(\boldsymbol{X} \mid y_{1:T})$$

For t = 1, …, T :

estimate $p(x_t \mid y_{1:T})$

Smoothing dist.





Variational DA method

- Adjoint method (4DVar)
- Representer method

Sequential DA method

- Kalman filter (KF), smoother
- Extended KF (EKF)
- Ensemble KF (EnKF)
- Particle filter (PF)

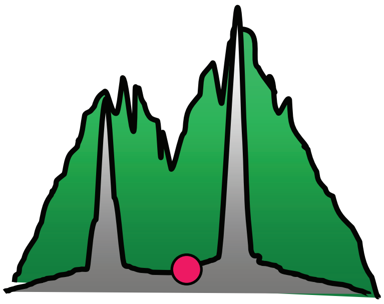
Optimization and Statistical Inference

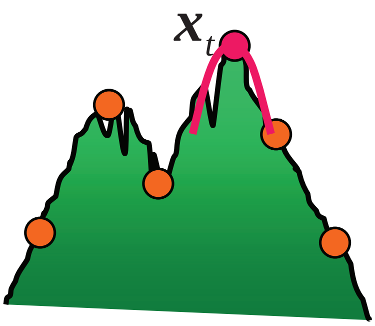
$f(\boldsymbol{X})$ Dimension of $\boldsymbol{X}$ is huge

$$\hat{\boldsymbol{X}} = \max\{- f(\boldsymbol{X})\}$$

$$- f(\boldsymbol{x}_t) \Rightarrow p(\boldsymbol{x}_t)$$

$$\hat{\boldsymbol{x}}_t = \int p(\boldsymbol{x}_t)\cdot \boldsymbol{x}_t d\boldsymbol{x}_t$$





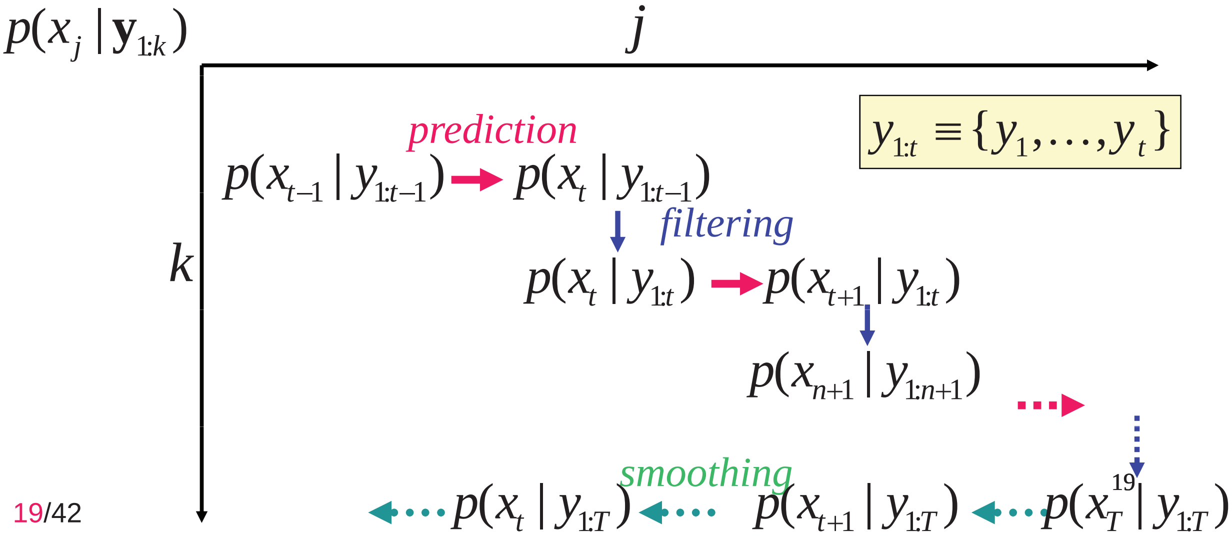
Conditional Distribution

Recursive formula

predictive density: $p(x_t \mid \mathbf{y}_{1:t-1})$ Today's economic situation given yesterday's stock market data

filter density: $p(x_t \mid \mathbf{y}_{1:t})$ Today's economic situation estimated by the stock market data up to today

smoother density: $p(x_t \mid \mathbf{y}_{1:T})$ Today's economic situation analyzed by using all available data when we look back on the today in future



$p(x_j \mid \mathbf{y}_{1:k})$

j

k

$y_{1:t} \equiv \{y_1, \ldots, y_t\}$

prediction

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

filtering

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

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

smoothing

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

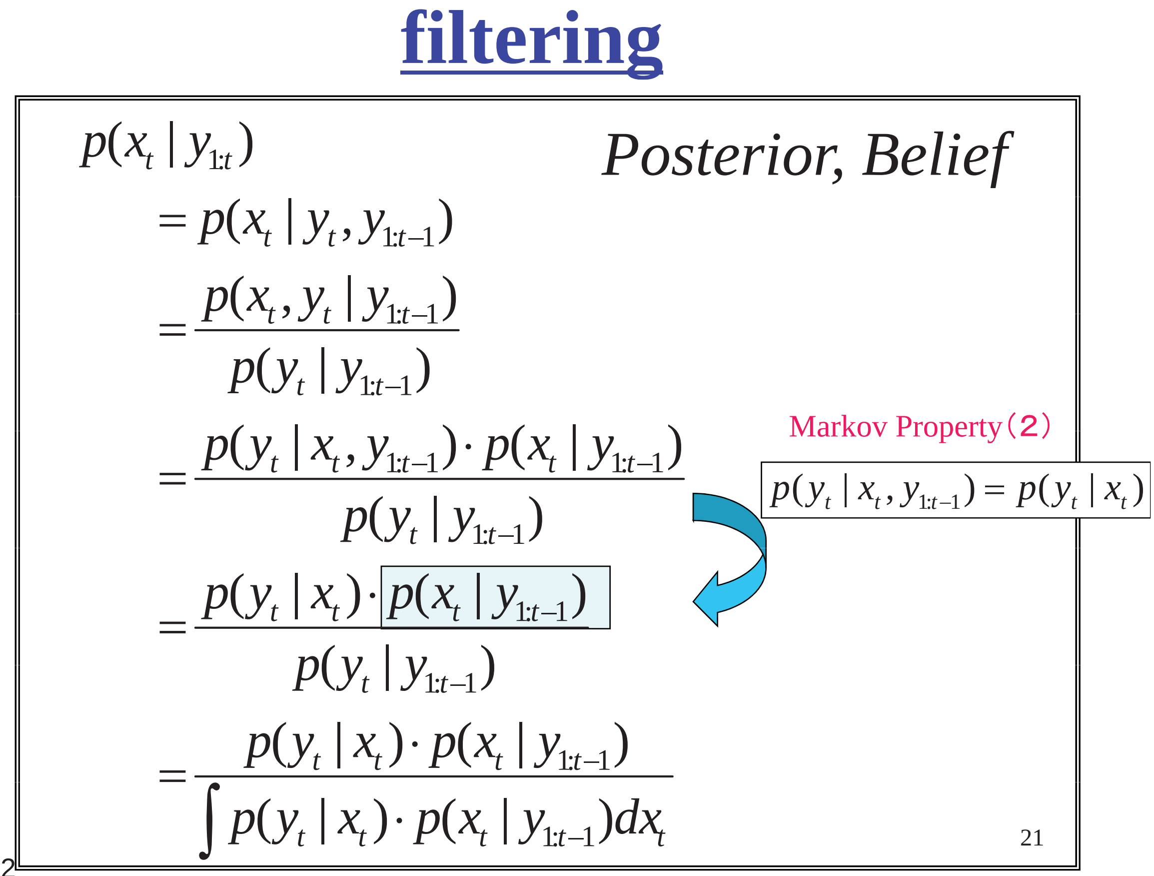
Prediction

$$
\begin{aligned}
p(x_t \mid y_{1:t-1}) &= \int p(x_t, x_{t-1} \mid y_{1:t-1}) dx_{t-1} \\
&= \int p(x_t \mid x_{t-1}, y_{1:t-1}) p(x_{t-1} \mid y_{1:t-1}) dx_{t-1}
\end{aligned}
$$

$\boxed{p(x_t \mid x_{t-1}, y_{1:t-1}) = p(x_t \mid x_{t-1})}$ Markov property（1）

$$
= \int p(x_t \mid x_{t-1}) \boxed{p(x_{t-1} \mid y_{1:t-1})} dx_{t-1}
$$

Filter pdf at time *t-1*



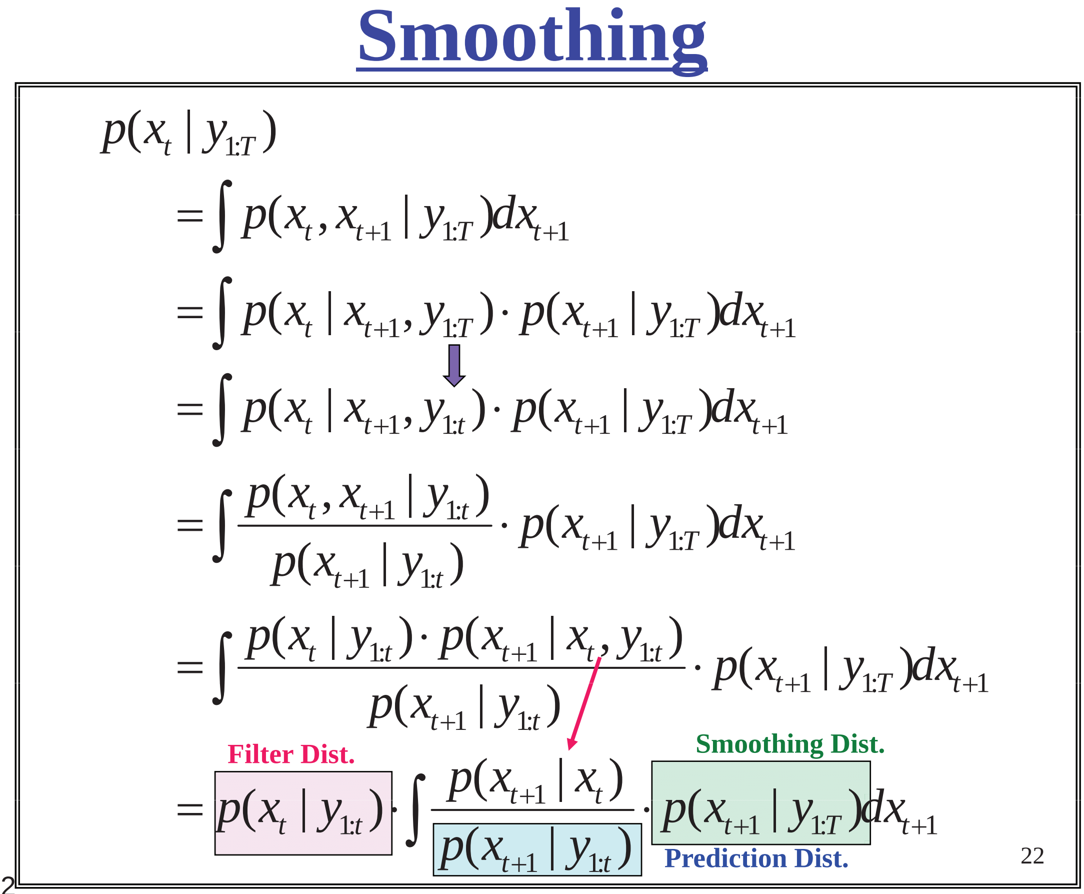
filtering

Posterior, Belief

$$
\begin{aligned}
p(x_t \mid y_{1:t}) \\
&= p(x_t \mid y_t, y_{1:t-1}) \\
&= \frac{p(x_t, y_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})} \\
&= \frac{p(y_t \mid x_t, y_{1:t-1}) \cdot p(x_t \mid y_{1:t-1})}{p(y_t \mid y_{1:t-1})} \\
&= \frac{p(y_t \mid x_t) \cdot \boxed{p(x_t \mid y_{1:t-1})}}{p(y_t \mid y_{1:t-1})} \\
&= \frac{p(y_t \mid x_t) \cdot p(x_t \mid y_{1:t-1})}{\int p(y_t \mid x_t) \cdot p(x_t \mid y_{1:t-1}) dx_t}
\end{aligned}
$$

Markov Property（２）

$$\boxed{p(y_t \mid x_t, y_{1:t-1}) = p(y_t \mid x_t)}$$



Smoothing

$$
\begin{aligned}
p(x_t \mid y_{1:T}) \\
&= \int p(x_t, x_{t+1} \mid y_{1:T}) dx_{t+1} \\
&= \int p(x_t \mid x_{t+1}, y_{1:T}) \cdot p(x_{t+1} \mid y_{1:T}) dx_{t+1} \\
&= \int p(x_t \mid x_{t+1}, y_{1:t}) \cdot p(x_{t+1} \mid y_{1:T}) dx_{t+1} \\
&= \int \frac{p(x_t, x_{t+1} \mid y_{1:t})}{p(x_{t+1} \mid y_{1:t})} \cdot p(x_{t+1} \mid y_{1:T}) dx_{t+1} \\
&= \int \frac{p(x_t \mid y_{1:t}) \cdot p(x_{t+1} \mid x_t, y_{1:t})}{p(x_{t+1} \mid y_{1:t})} \cdot p(x_{t+1} \mid y_{1:T}) dx_{t+1} \\
&= \underbrace{\boxed{p(x_t \mid y_{1:t})}}_{\text{Filter Dist.}} \cdot \int \frac{p(x_{t+1} \mid x_t)}{\underbrace{\boxed{p(x_{t+1} \mid y_{1:t})}}_{\text{Prediction Dist.}}} \cdot \underbrace{\boxed{p(x_{t+1} \mid y_{1:T})}}_{\text{Smoothing Dist.}} dx_{t+1}
\end{aligned}
$$

Sequential Data Assimilation

Estimate PDF of state vector x_t or its moments (mean, variance, …)
sequentially on each observation

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

$$\begin{tikzcd}
y_{t-1} \arrow[d] & & \\
\boxed{p(x_{t-1} \mid y_{1:t-1})} \arrow[r, "\text{Simulation}"'] & \boxed{p(x_t \mid y_{1:t-1})} \arrow[d, "y_t"] & \\
& \boxed{p(x_t \mid y_{1:t})} \arrow[r, "\text{Simulation}"'] & \boxed{p(x_{t+1} \mid y_{1:t})} \arrow[d, "y_{t+1}"] \\
& & {}
\end{tikzcd}$$

Challenging problem: Huge dimension and inversion

- Data Assimilation = Estimation problem of state vector x_t:

(system model) $x_t = F_t(x_{t-1}, v_t | x_0)$

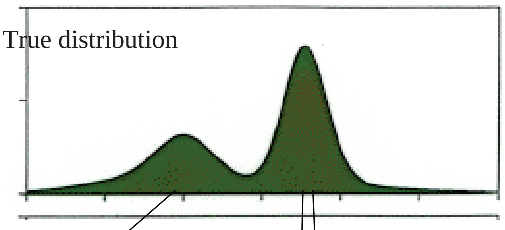
(observation model) $y_t = H_t x_t + w_t$ or $y_t = h_t(x_t) + w_t$

- x_t : All variables in simulation model
- y_t : All observed variables
- v_t : Stochastic part to represent uncertainty of model (boundary condition, …)
- w_t: Observation error
- v_t, w_t : Normally Gaussian x_0: Initial condition

dimension x_t :10⁴~10⁶ y_t :10²~10⁵ $\dim(x_t) >> \dim(y_t)$

Numerical representation of distribution

$p(x_t \mid y_{1:t-1}), p(x_t \mid y_{1:t}), p(x_t \mid y_{1:T})$



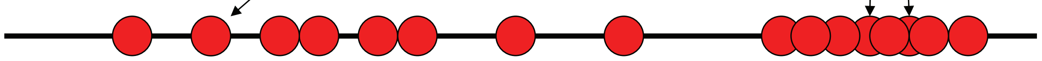
Monte Carlo approximation

Represent pdf by the actual realizations.

N: # of particles

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

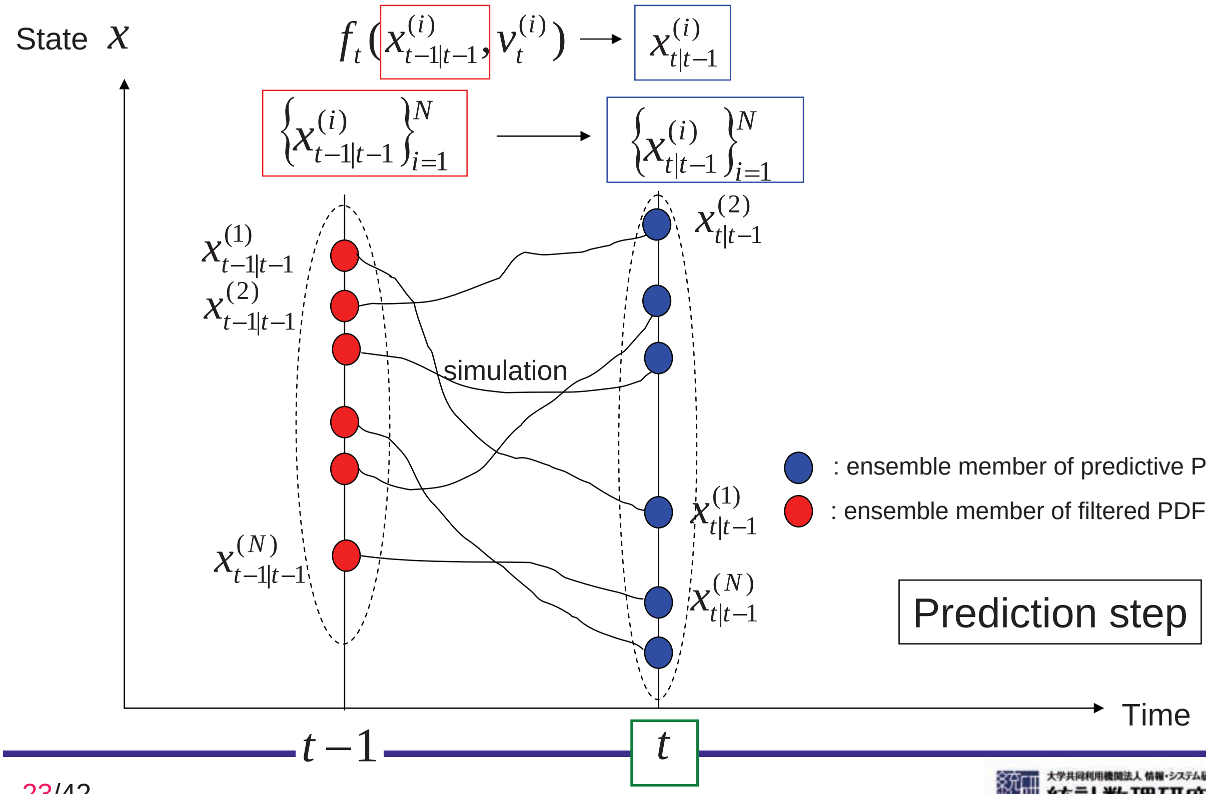
$$p(x_t \mid y_{1:t}) \cong X_{t|t} \equiv \left[x^{(1)}_{t|t}, x^{(2)}_{t|t}, \ldots, x^{(N)}_{t|t}\right]$$



Prediction Step (Common in EnKF and PF)

$$f_t\left(x^{(i)}_{t-1|t-1}, v^{(i)}_t\right) \longrightarrow x^{(i)}_{t|t-1}$$

$$\left\{x^{(i)}_{t-1|t-1}\right\}_{i=1}^{N} \longrightarrow \left\{x^{(i)}_{t|t-1}\right\}_{i=1}^{N}$$

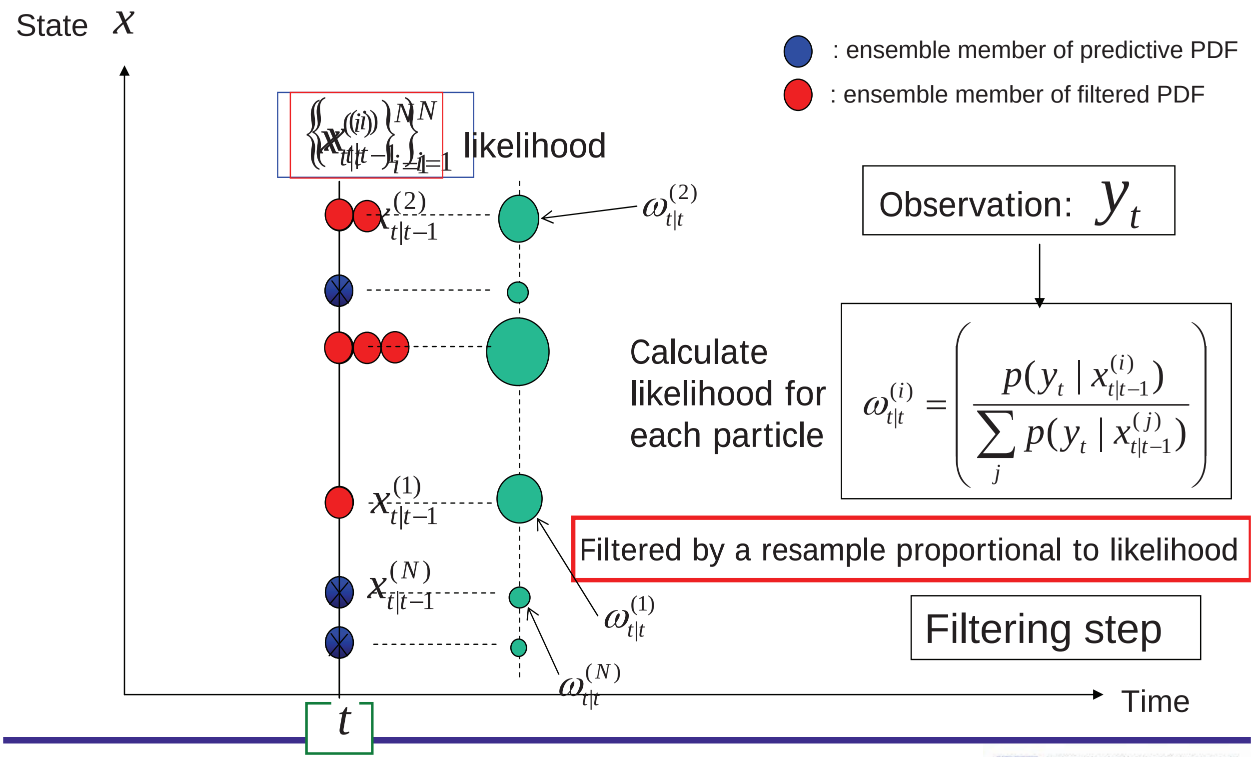


● : ensemble member of predictive PDF

● : ensemble member of filtered PDF

Prediction step

Filtering Step of PF



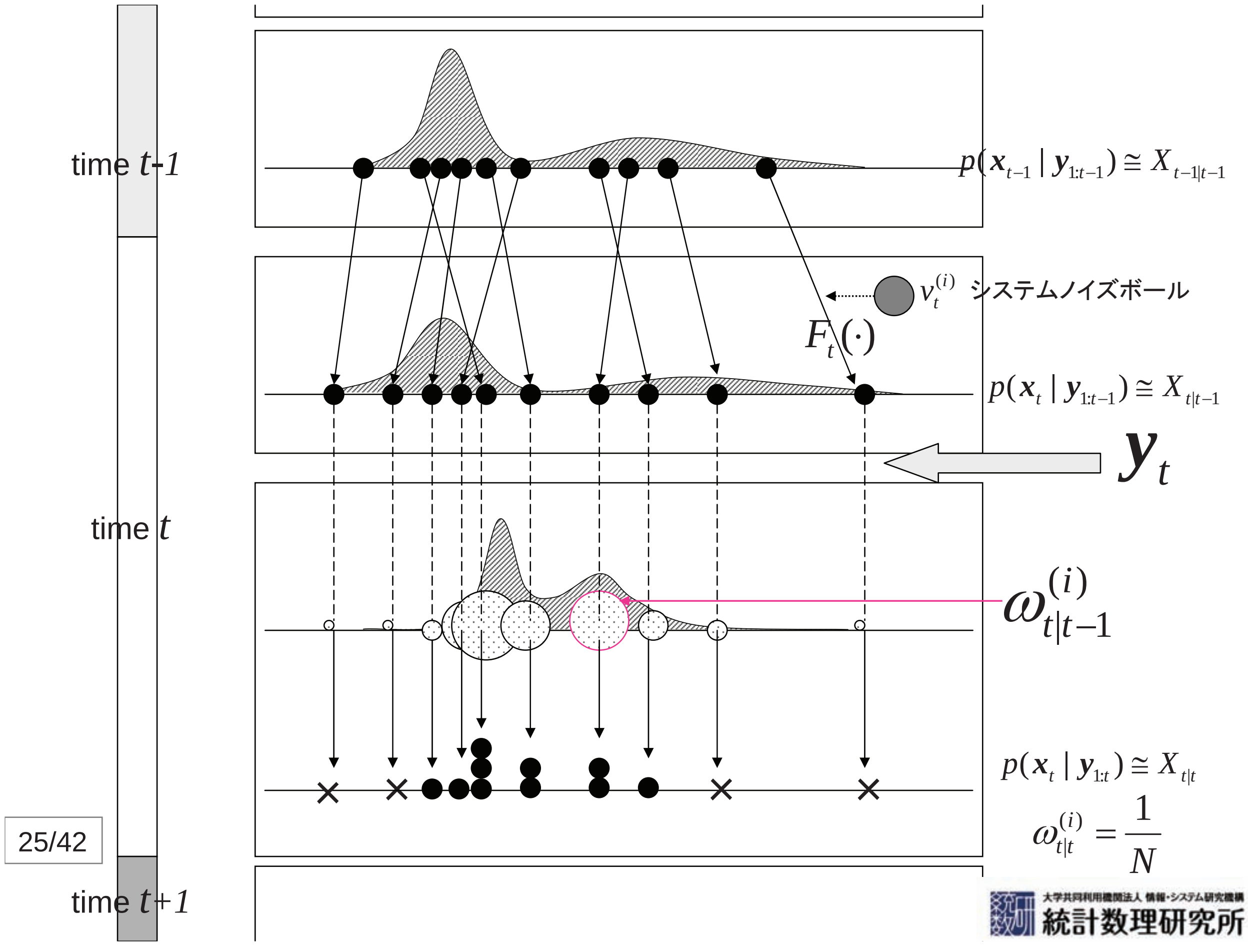
Observation: y_t

Calculate likelihood for each particle

$$\omega_{t|t}^{(i)} = \left(\frac{p(y_t \mid x_{t|t-1}^{(i)})}{\sum_j p(y_t \mid x_{t|t-1}^{(j)})} \right)$$

Filtered by a resample proportional to likelihood

Filtering step



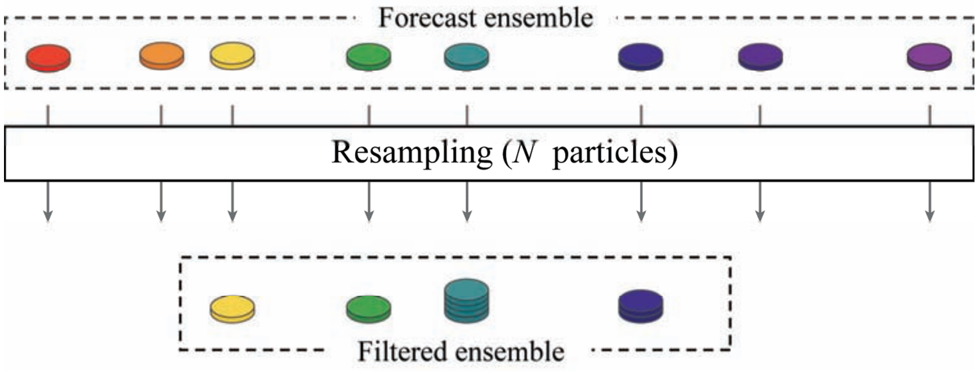
Family of particle filters

- Kalman filter (EKF: Extended Kalman filter)
- EnKF: Ensemble Kalman filter (Evensen 1994)
- Particle filter
  - SIR filter (e.g., Gordon et al. 1993, Kitagawa 1993)★
  - Gaussian particle filter (e.g., Kotecha and Djuric 2003)
  - Kernel filter (e.g., Hurzeler and Kunsh 1998)
  - Merging particle filter (MPF)

★ It encounters a problem called ‘degeneration’ in applying to high-dimensional models. (i.e., the diversity of an ensemble is lost after repeating resampling procedures.)

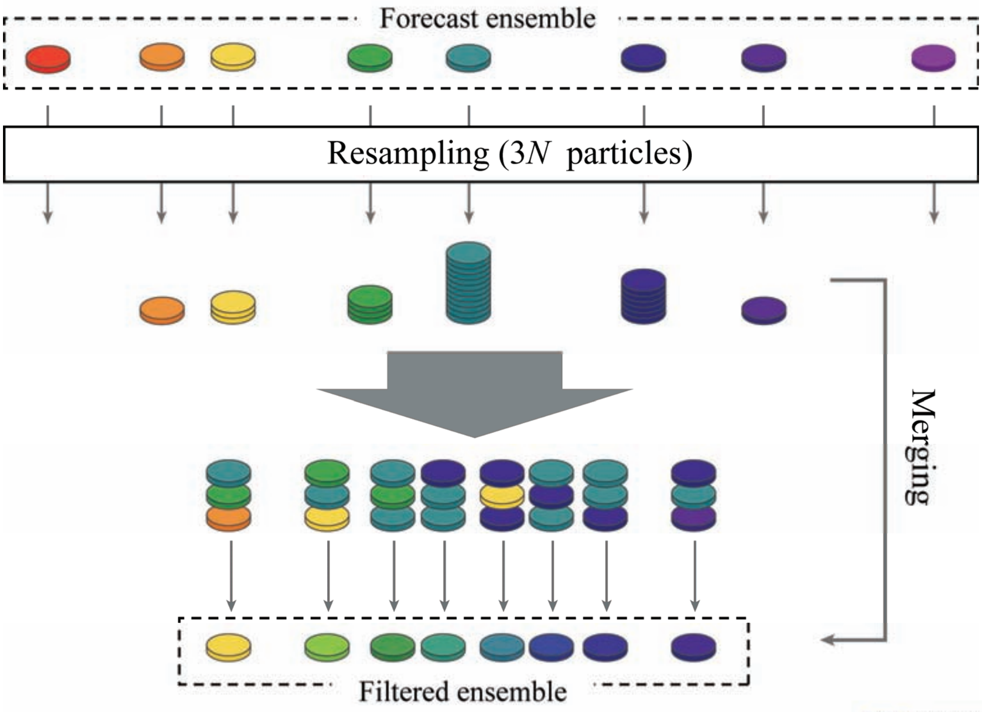
※Gaussian PF: 1) Each particle in filtered ensemble is drawn from a Gaussian function with the mean and covariance of the forecast ensemble. 2) It requires high computational cost due to a factorization of a high-dimensional covariance matrix in generating Gaussian samples.

Particle filter
(SIR: **S**equential **I**mportance **R**esampling)



A posterior (filtered) ensemble is obtained by resampling the forecast ensemble with weights of likelihood. Thus, an ensemble member is duplicated in the filtered ensemble according to its likelihood.

Merging particle filter (MPF)



MPF algorithm (1)

We draw $\boxed{n}\times N$ samples from the forecast ensemble with weights of w_i, and obtain an ensemble: $\{\hat{x}_{t|t}^{(1,1)},\cdots,\hat{x}_{t|t}^{(n,1)},\cdots,\hat{x}_{t|t}^{(1,N)},\cdots,\hat{x}_{t|t}^{(n,N)}\}$.

A subset $\{\hat{x}_{t|t}^{(j,1)},\cdots,\hat{x}_{t|t}^{(j,N)}\}$ from this $\boxed{n\times N}$ samples satisfies

$$p(x_t \mid y_{1:t}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta\left(x_t-\hat{x}_{t|t}^{(j,i)}\right)$$

because it is a filtered ensemble obtained using normal PF.

Each number of a filtered ensemble is generated as a weighted sum of n samples from the $n\times N$ sample set as:

$$x_{t|t}^{(i)}=\sum_{j=1}^{n}\boxed{\alpha_j}\hat{x}_{t|t}^{(j,i)}$$

MPF algorithm (2)

In order to ensure that the newly generated ensemble approximately preserves the mean and covariances of the filtered PDF, the merging weights α_j are set to satisfy

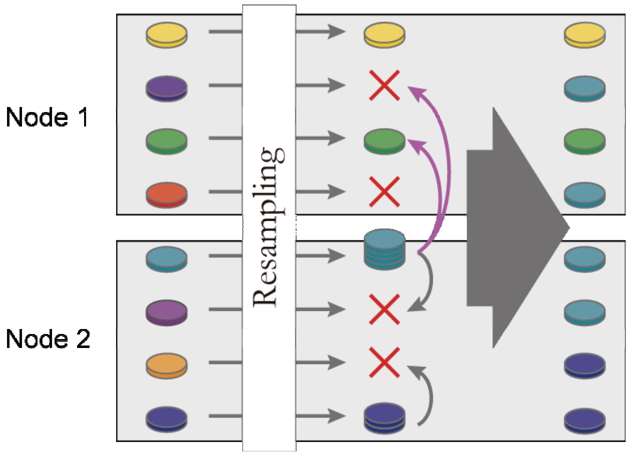
$$\boxed{\sum_{j=1}^{n}\alpha_j = 1, \quad \sum_{j=1}^{n}\alpha_j^2 = 1} \quad (n \leq 3 \text{ such that } \alpha_j \neq 0 \text{ for all } j)$$

where each α_j is a real number.

Then, a new ensemble approximation of the filtered PDF $p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t})$ is obtained as

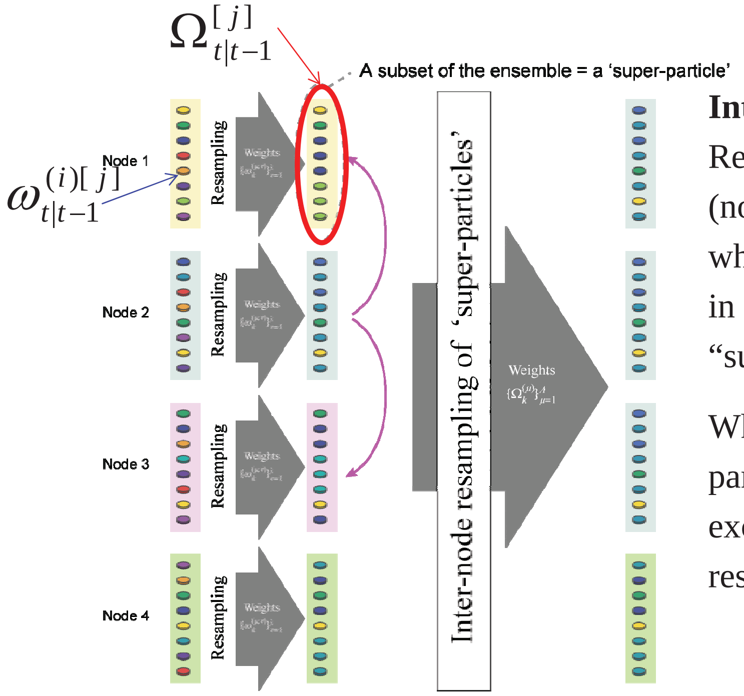
$$p(\boldsymbol{x}_t \mid \boldsymbol{y}_{1:t}) \approx \frac{1}{N}\sum_{i=1}^{N}\delta\left(\boldsymbol{x}_t - \hat{\boldsymbol{x}}_{t|t}^{(j,i)}\right)$$

Flowchart of PF



A concept of the “Islands” in GA is similar, but different.

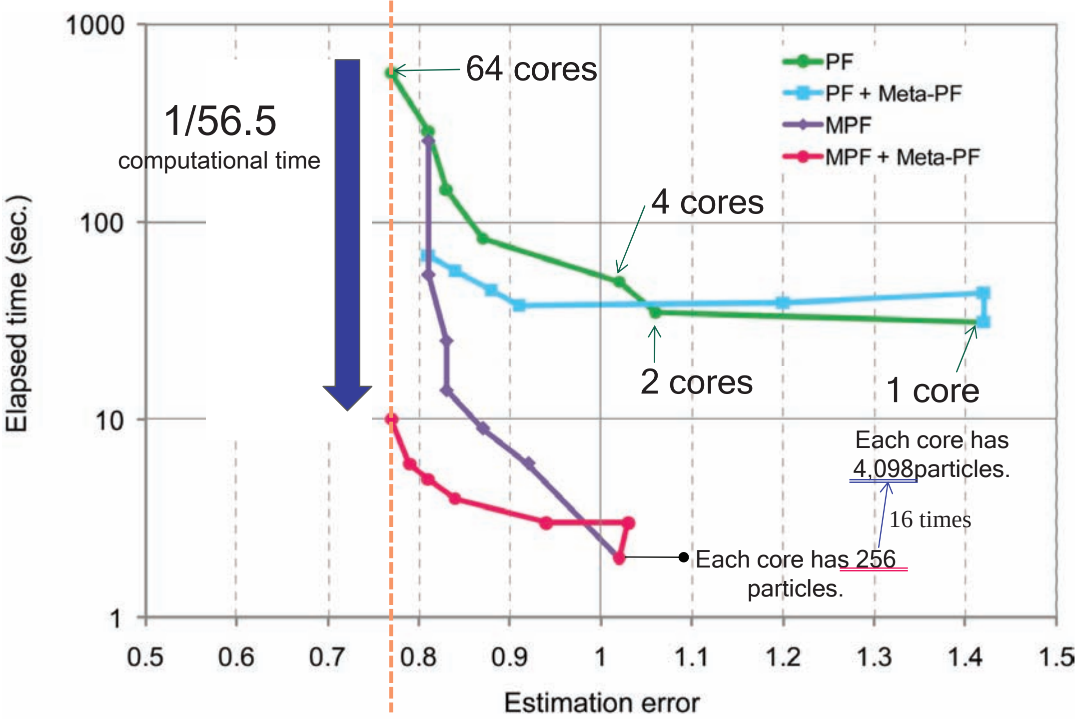
Meta particle filter
(DPF: Distributed particle filter)



Inter-node resampling: Resampling between ensembles (not ensemble member) each of which consists of many particles in a node. This ensemble is called "super-particle".

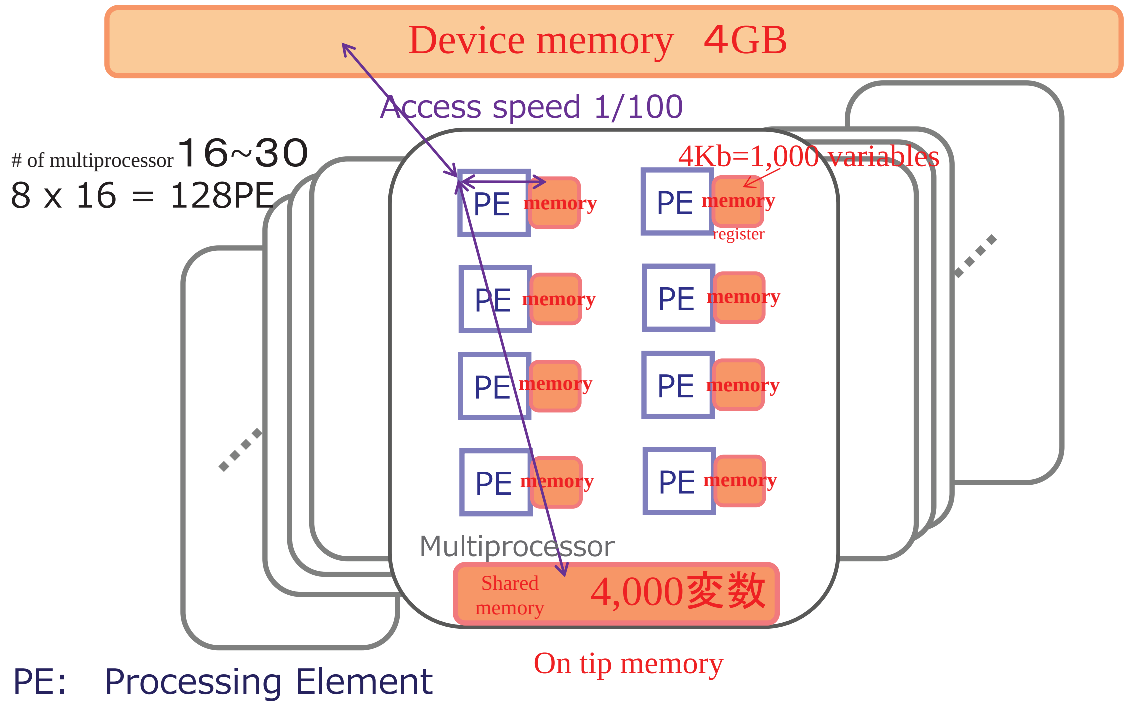
When a weight for any super-particles (i.e., nodes) $\Omega^{[j]}_{t|t-1}$ exceeds 0.3, the inter-node resampling procedure is applied.

Result

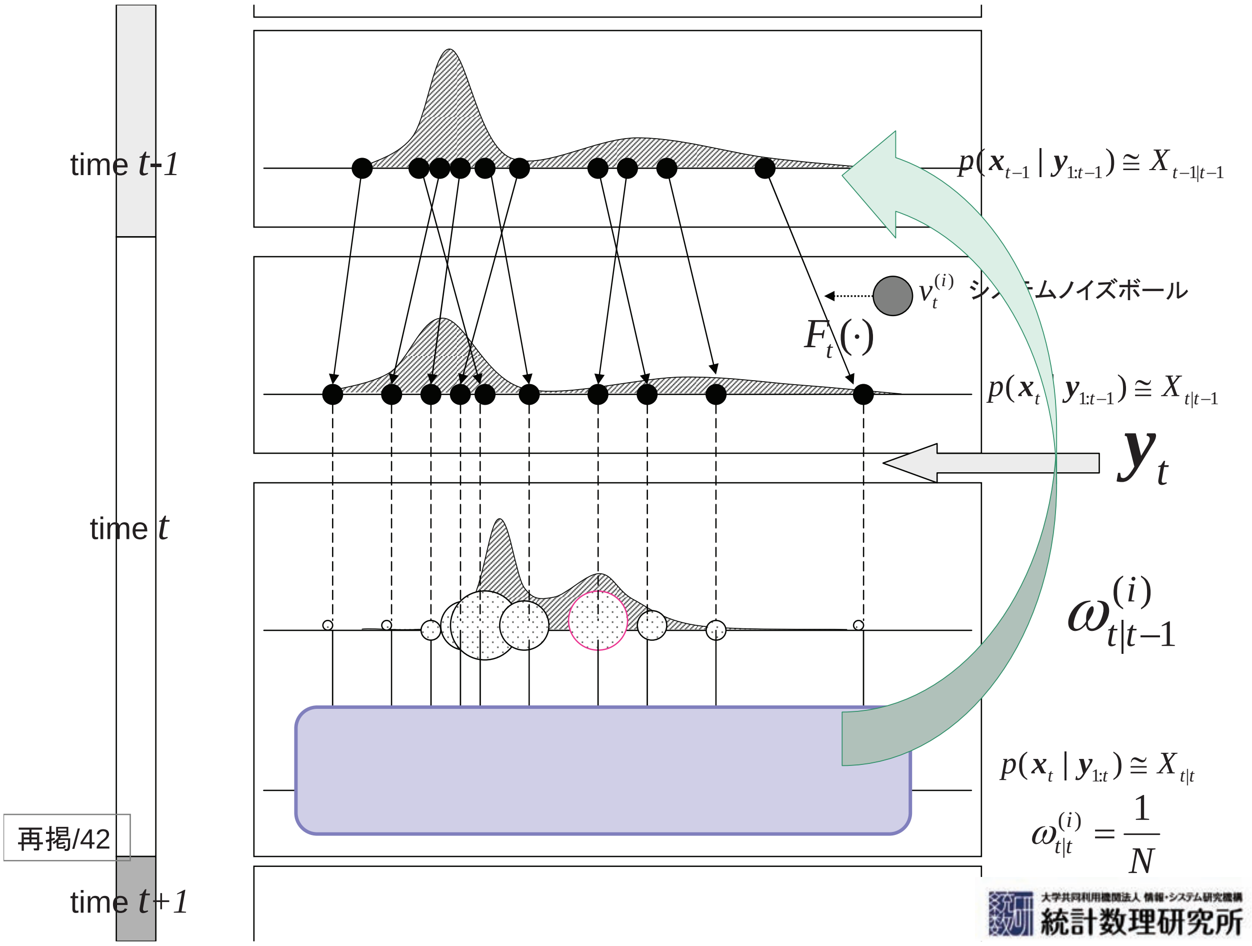


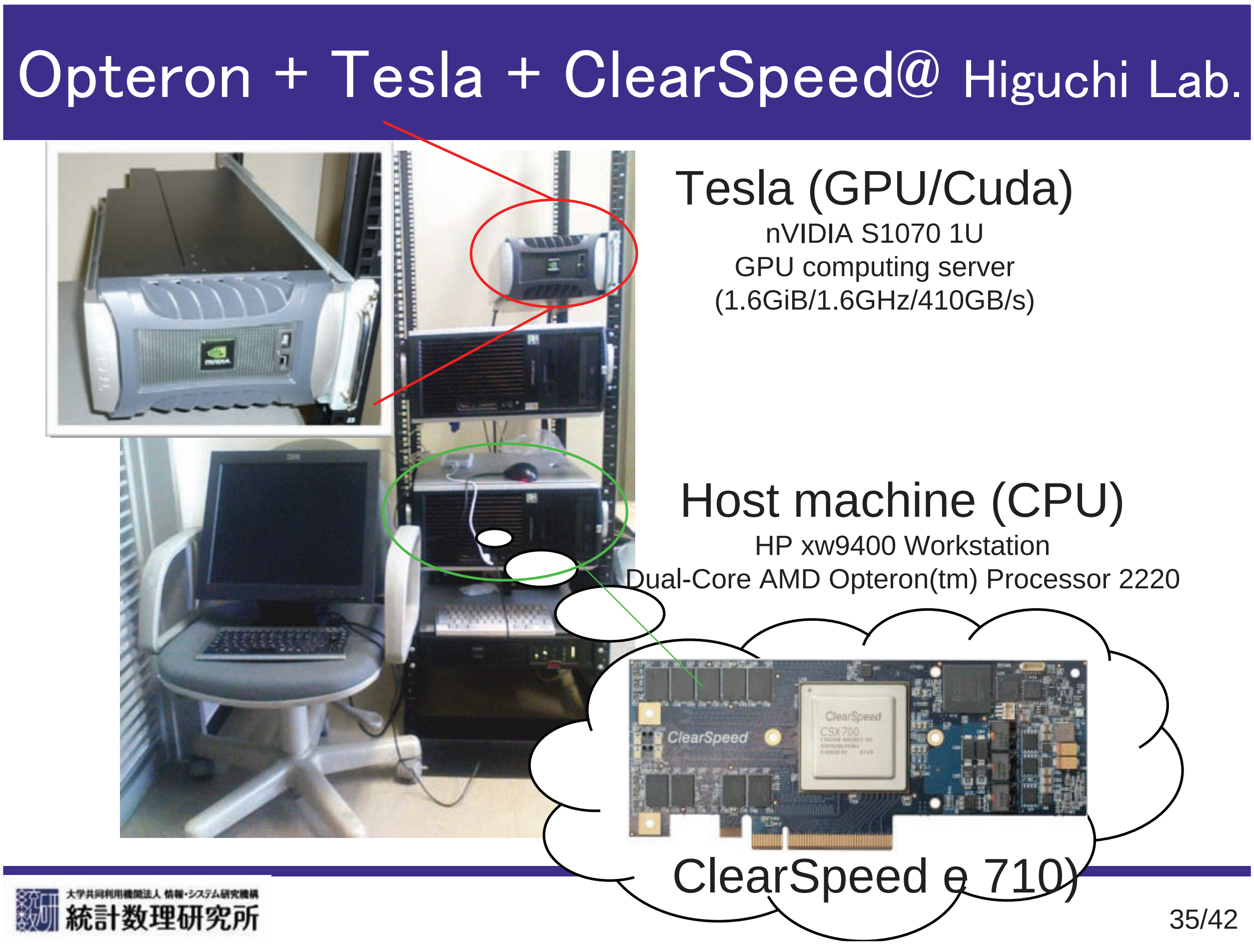
Inside GPU:128～240 parallel computing

GPGPU’s power is equivalent to a computer with 128 1.2GHz processors



PE: Processing Element





Opteron + Tesla + ClearSpeed@ Higuchi Lab.

Tesla (GPU/Cuda)

nVIDIA S1070 1U
GPU computing server
(1.6GiB/1.6GHz/410GB/s)

Host machine (CPU)

HP xw9400 Workstation
Dual-Core AMD Opteron(tm) Processor 2220

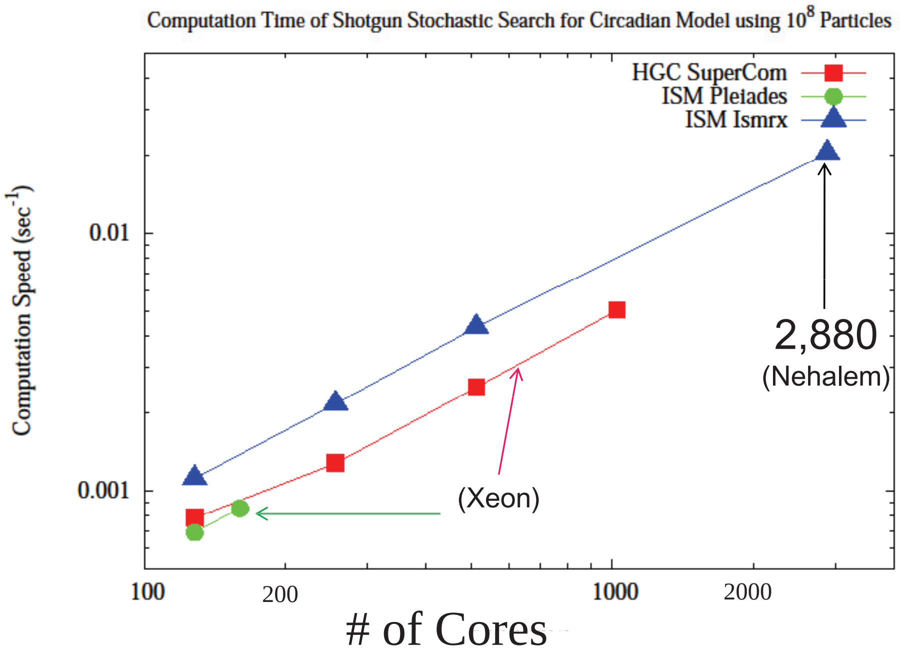
ClearSpeed e 710)

SIS meets GPGPU.

- SIS on GPGPU designed for parameter estimation.
  - Simulation is carried out on GPGPU.
  - Parameter estimation is carried out on CPU.

particles	**PF** **Opteron 2220 1core** (Nakamura et al., 2009)	**PF** **Opteron2220 1core** **+ GPGPU(Tesla C870)**	**SIS** **Opteron2220 1core** **+ GPGPU(Tesla C870)**
100,000,000 (1億)	8Days (6.8 × 10^5sec)	12Hours (4.5 × 10^4sec)	3Hours (1.0 × 10^4sec)
	1	×15	×67
1,000,000,000 (10億)	2.6 Months?	5 Days?	28Hours (1.0 × 10^5sec)

HPC of our group



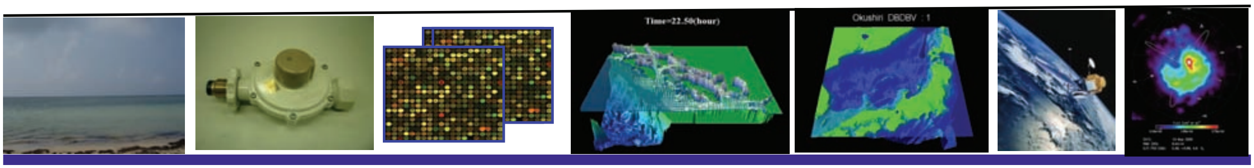
Next-Generation of Supercomputer in Japan at Kobe

■Grand Challenge:
-- Nanotech (Institute for Molecular Science)
-- Life Science (RIKEN)

Japanese Government will spend more than 1 billion US$ for this national project. It has more than **600,000** cores.

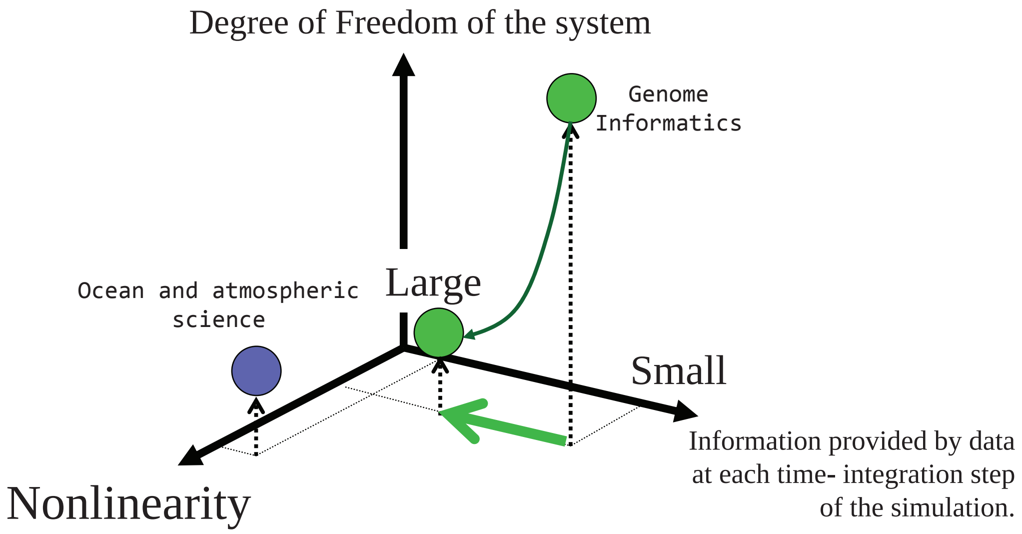
Research projects in progress by our group

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



43

TIPS: A choice of the data assimilation methods
Reduction of degree of freedom



Contact

Email: higuchi@ism.ac.jp

Homepage:
http://daweb.ism.ac.jp/