

# 状態空間モデルを利用した パーソナライゼーション技術の研究

樋口知之

情報・システム研究機構 統計数理研究所

## 社会的要請: Personalization

- 背景:
  - 無駄を省く(低価格化, 低コスト)
  - 資源の有効利用のために選択と集中
  - 価値観の多様化
  - “コ”(個人, 個性, 固有, 個別)に特化

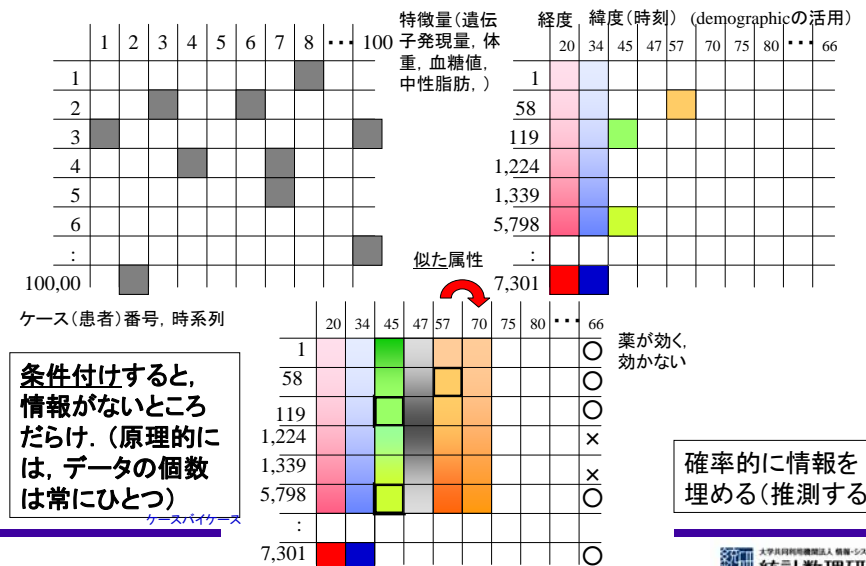
大量生産・大量消費をめざした20世紀→  
個人に焦点をあわせる科学へ

オーダーメイド医療, 副作用の研究, マイクロマーケティング,  
One-to-One \*, Situation \*, 環境に優しい

## 現代のデータ環境下における研究推進の鍵

- 帰納的推論と機能のモデル化
  - 個人化技術 (*Personalization*)
- 新しい  $np$  問題 (NP 古くは Non-deterministic Polynomial time)  
 $n \ll p$   $n$ : サンプル数、 $p$ : 属性変数

## データの有限性→情報の欠損



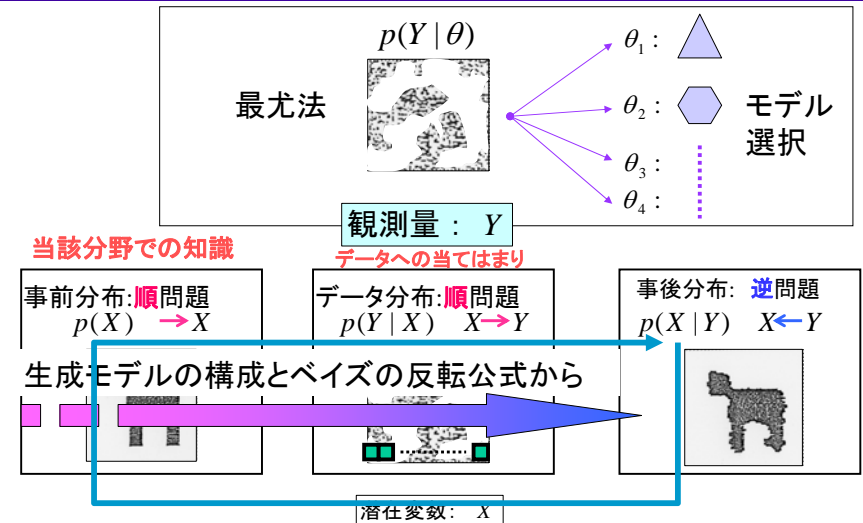
## 見えないものを大規模データで推量する

### わからないもの, 見えないもの

- 人間・生活活動に関連する大量大規模データの出現
- 不十分ではあるが, 一人一人の考え方を間接的に捉えることが可能に.

■ データを生成するモデルを, モデルが持つデータの記述(説明)能力(情報量規準)の比較により、各研究者のアイデアの客観的な相互比較が実現

## 生成モデルと反転公式、そして自然な循環機能



## 話題提供

1. 庶民感覚を活かすマイクロマーケティング
2. エキスパートの感と経験にもとづく高度な技術の伝承法
3. シミュレーションを取り込み融合するデータ同化

## 話題1. 庶民感覚を活かすマイクロマーケティング

相手が知らない情報を推測することで優位に立つ.

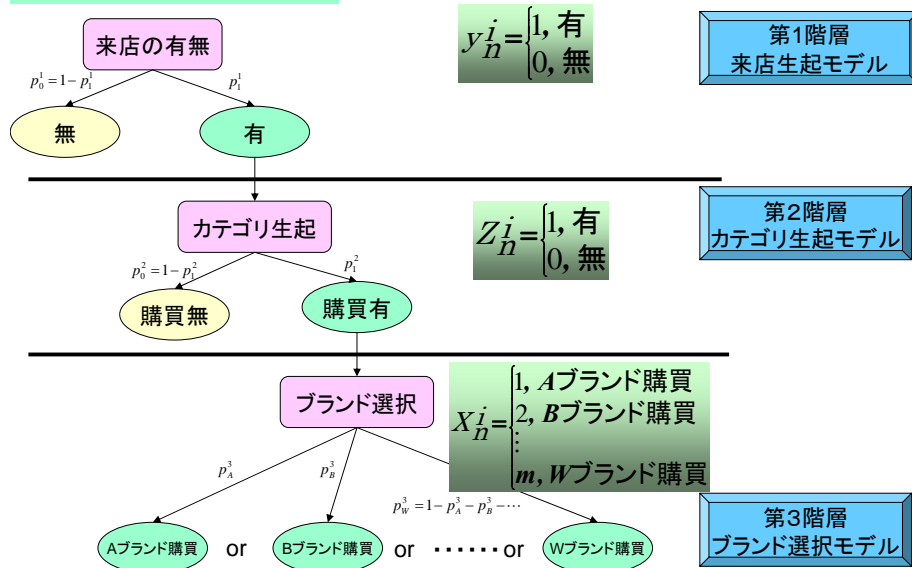
筑波大学大学院ビジネス科学研究科 准教授  
前 (財)流通経済研究所  
佐藤 忠彦



Computer-based Personalized Service

# デモグラフィック情報他の構造を階層化モデリング

個人モデルとして



# 個人毎の来店確率モデル

第 $n$ 期にある消費者 $i$ が来店するかどうか、 $y_n^i$ 、は下記ベルヌーイ過程に従う。

$$P(y_n^i | p_n^i) = p_n^i y_n^i (1 - p_n^i)^{1 - y_n^i}$$

ただし、 $p_n^i = \frac{\exp(u_n^i)}{1 + \exp(u_n^i)}$ とする。

本モデルが本研究で提案する一般状態空間モデルの『観測モデル』になる

# 来店効用

$$u_n^i = (t_n^i) + (w_n^i) + (\beta_{n,1}^i)x_n^1 + (\beta_{n,2}^i)x_n^2 + (\beta_{n,3}^i)x_n^3 + (\beta_{n,4}^i)x_n^4 + (\beta_{n,5}^i)x_n^5$$

$t_n^i$ : トレンド

$w_n^i$ : 週周期成分

時変パラメータ

$x_n^1$ : 基準化変数の対数(前回来店からの日数)

$x_n^2$ : 基準化変数の対数(降水量)

$x_n^3$ : 基準化変数の対数(チラシ掲載商品の平均値引率)

$x_n^4$ : 基準化変数の対数(チラシ掲載商品数)

$x_n^5$ : 前日来店の有無

$\beta_{n,1}^i, \beta_{n,2}^i, \beta_{n,3}^i, \beta_{n,4}^i, \beta_{n,5}^i$ : 時変パラメータ

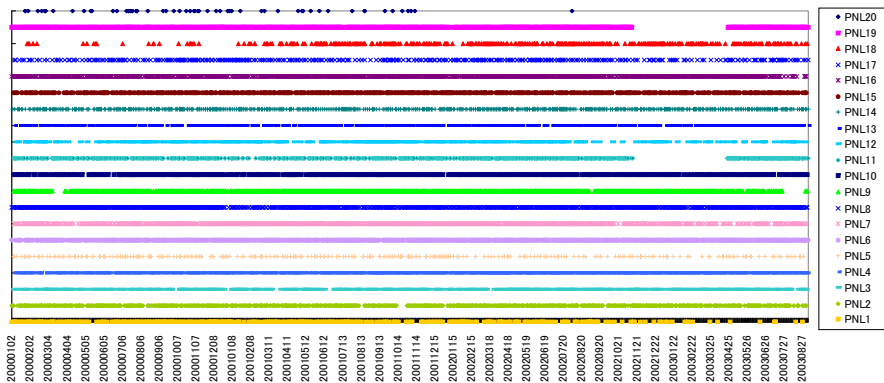
各 $x_{n,j}^i$ の変数毎に、値を規格化

# データ

- 東京にある総合スーパー
- データ期間: 2000年1月2日 ~ 2003年9月5日 (N=1343日)
- データ種別: ID付POSデータ

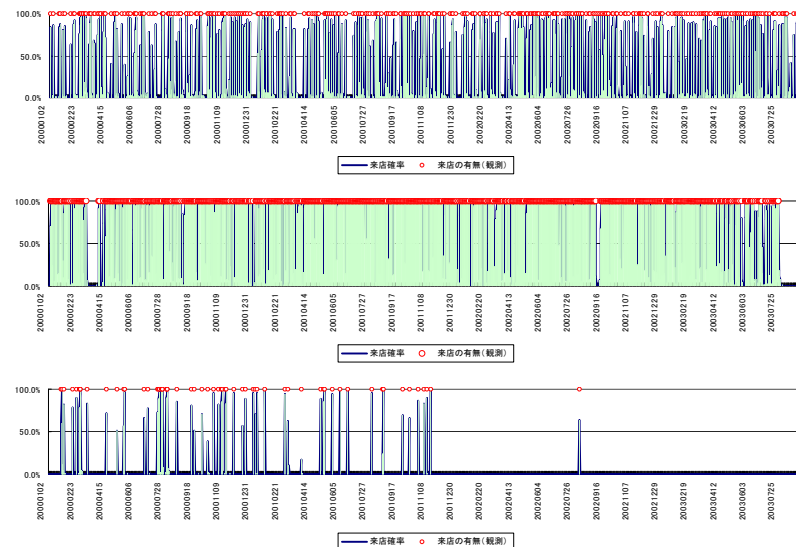
記号	変数
$x_n^1$	前回来店からの日数(基準化後)の対数
$x_n^2$	降水量(基準化後)の対数
$x_n^3$	チラシ掲載商品の平均値引率(基準化後)の対数
$x_n^4$	チラシ掲載商品数(基準化後)の対数
$x_n^5$	前日来店の有無

# 来店時系列(パネル毎)



本提案の枠組みでは、上記の一人一人に対してモデルの推定を行う

# 来店確率(フィルタ推定値)



# 個人モデルの構築

Accepted Model (AIC)	Accepted Model (BIC)					Number of Accepted Models
	No.1	NO.2	NO.6	NO.9	No.10	
No.1	22	18		4	7	51
NO.2		21		5	5	31
NO.6			1			1
NO.9	2	17		30	13	62
No.10	4	7		15	29	55

- No.1:トレンド, 週周期成分, Recency, 降雨量, 昨日来店の有無, チラシ(平均値引率), チラシ(掲載商品数)
- No.2:トレンド, 週周期成分, Recency, 降雨量, 昨日来店の有無
- No.6:週周期成分, Recency, 降雨量, 昨日来店の有無, チラシ(平均値引率), チラシ(掲載商品数)
- No.9:トレンド, 週周期成分, Recency, 降雨量, 昨日来店の有無, チラシ(平均値引率)
- No.10:トレンド, 週周期成分, Recency, 降雨量, 昨日来店の有無, チラシ(掲載商品数)

# 話題2. エキスパートの感と経験にもとづく 高度な技術の伝承法

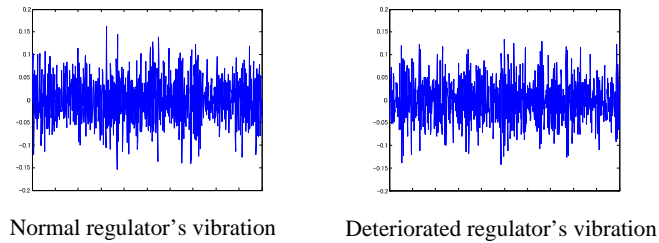


総合研究大学院大学 石垣 司  
現在 産総研・サービス工学研究センター 研究員

法政大学 渡辺嘉二郎

# 見分ける, 聞き分ける

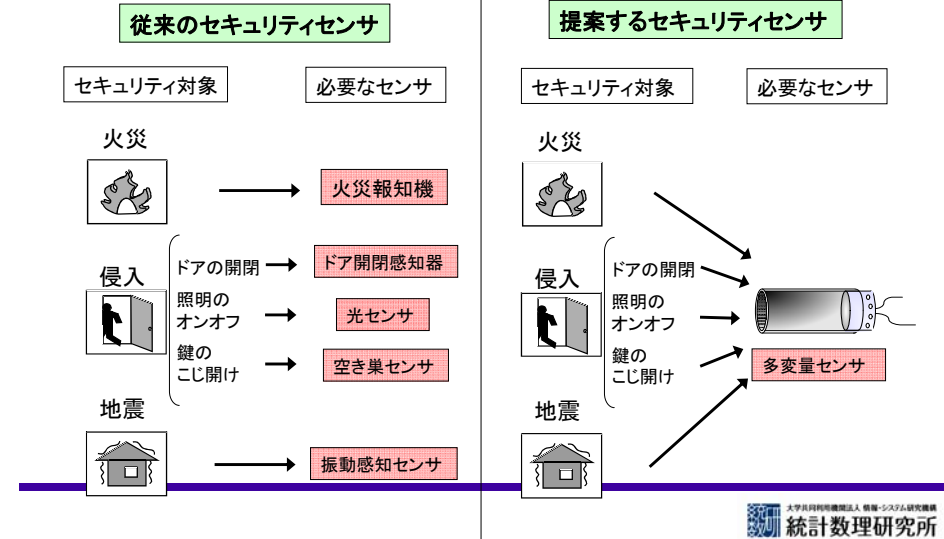
We developed new measurement system of the diaphragm vibration with a microphone sensor. Time series obtain by the sensor of the proposed system. However,



What's the difference?

ことばの出典: 信学会誌, 小特集「見分ける, 聞き分ける」Vol. 87, No. 3, 2004.

# 応用例2-1: 家庭用スマートセキュリティセンサ



# 特徴の抽出: 時系列の分解

各災害信号をトレンド成分、極低周波成分、低周波成分に分解する。

$$y_n = t_n + q_n^{[1]} + q_n^{[2]} + w_n, \quad w_n \sim N(0, \sigma^2), \quad (n=1, \dots, N)$$

一トレンド成分のモデル化—  
2階差分トレンドモデル

$$t_n = 2t_{n-1} - t_{n-2} + v_n, \quad v_n \sim N(0, \tau_0^2)$$

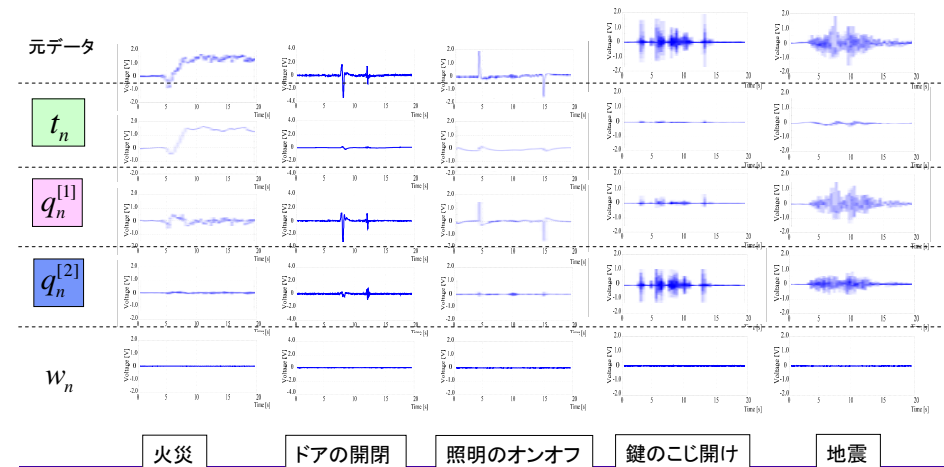
一極低周波成分と低周波成分のモデル化—  
準周期的振動モデル

$$q_n^{[m]} = 2 \cos(2\pi f_m \Delta t) q_{n-1}^{[m]} - q_{n-2}^{[m]} + v_n^{[m]}, \quad v_n^{[m]} \sim N(0, \tau_m^2) \quad (m=1, 2)$$

$y_n$ : 時系列信号  
 $t_n$ : トレンド成分  
 $q_n^{[1]}$ : 極低周波成分  
 $q_n^{[2]}$ : 低周波成分

局所的な確率差分方程式であるので、非定常なデータに対してもデータ適応的に対応することができる

# 分解の様子



## 特徴ベクトルの設定

特徴選択

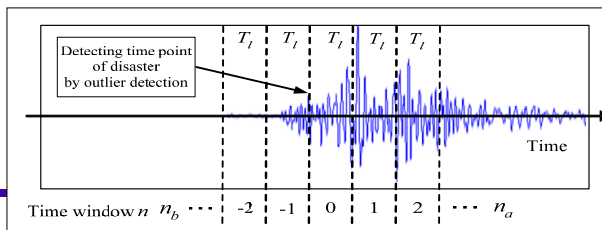
$$p_n(x) = \sum_{i \in n} |x_i|$$

$$s_n(x) = \sum_{i \in n} x_i$$

$$P_n \equiv p_n(t) + p_n(q^{[1]}) + p_n(q^{[2]}) + p_n(w)$$

特徴

$\frac{p_n(t)}{P_n}$	$\frac{s_n(t)}{P_n}$	$\frac{p_n(t) - p_{n-1}(t)}{P_n - P_{n-1}}$
$\frac{p_n(q^{[1]})}{P_n}$	$\frac{s_n(q^{[1]})}{P_n}$	$\frac{p_n(q^{[1]}) - p_{n-1}(q^{[1]})}{P_n - P_{n-1}}$
$\frac{p_n(q^{[2]})}{P_n}$	$\frac{s_n(q^{[2]})}{P_n}$	$\frac{p_n(q^{[2]}) - p_{n-1}(q^{[2]})}{P_n - P_{n-1}}$

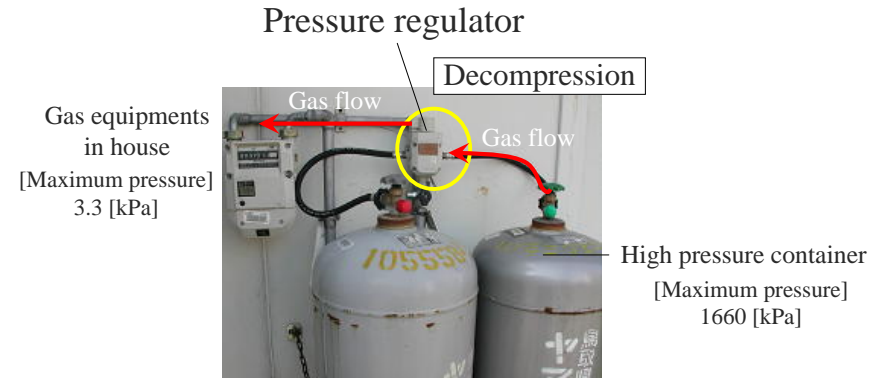


$$n_b = -1, n_a = 3, T_l = 1 [\text{sec}]$$

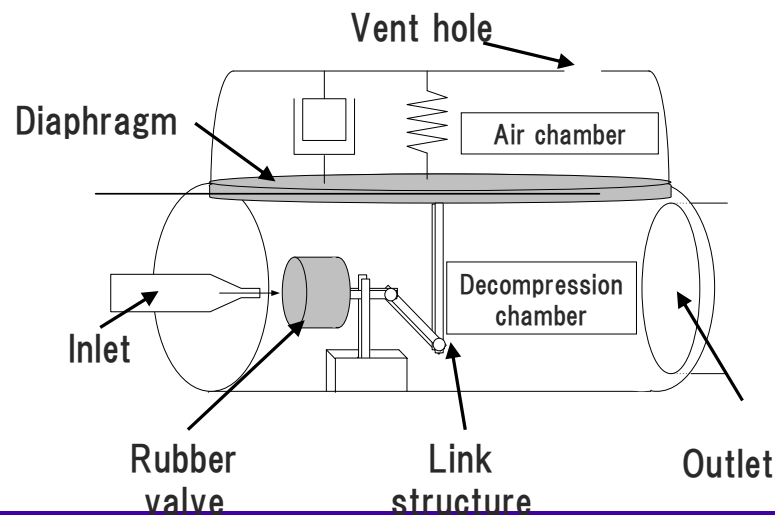
特徴ベクトル→42次元

## 応用例2-2:プロパンガス

Fault diagnosis must be performed in order to ensure the safety of machines and instruments

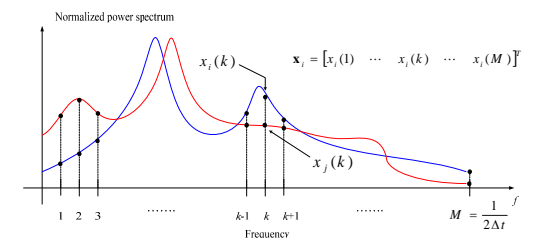


## ガス圧調整器



## 正規化周波数スペクトル判別

- 周波数スペクトルを正規化したものの判別問題
- 音声認識、故障診断、地震波解析などに利用
- 適切なカーネルを設計することで高精度なシステムの実現。





# Kullback-Leibler ダイバージェンス

- 確率分布間の近さを測る指標として統計学において頻繁に使用され、非常に有用なことが知られている。

– KLダイバージェンス

$$KLD(p(x), q(x)) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

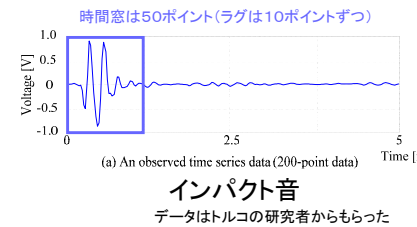
- これを対称化したダイバージェンス

– 対称化KLダイバージェンス

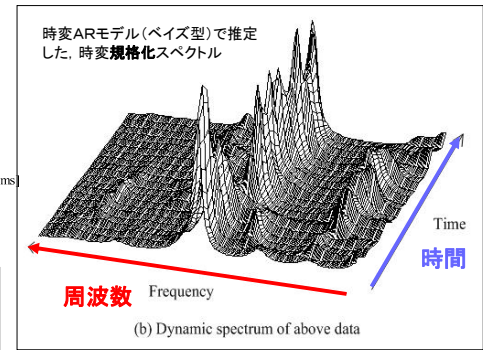
$$SKLD(p(x), q(x)) = \int \{p(x) - q(x)\} \log \frac{p(x)}{q(x)} dx$$

※注 このスライドのみ  $x$  は確率変数. また,  $p, q$  は確率密度関数

# 時変スペクトルへの応用(1):ヘーゼルナッツの判別



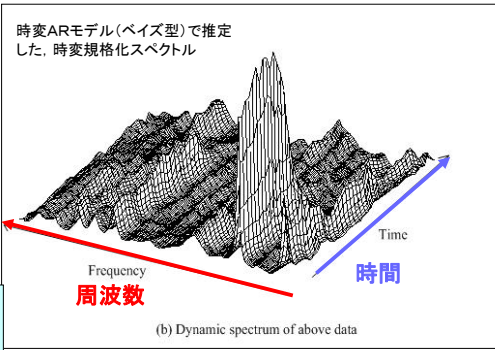
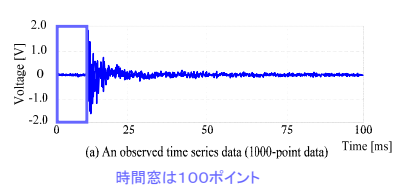
KLカーネル 判別結果 91.7%  
 Gaussianカーネル → 90.4%  
 3次多項式カーネル → 87.2%  
 既存手法 → 90.1%



## KLカーネルへのインプット

- 各時間窓毎にスペクトルを規格化
- 全体のスペクトルを規格化

# 時変スペクトルへの応用(2):疲労紙幣の判別



KLカーネル → 99.4%  
 Gaussianカーネル → 93.4%  
 3次多項式カーネル → 87.0%

# 玄人の眼, 聴を造る

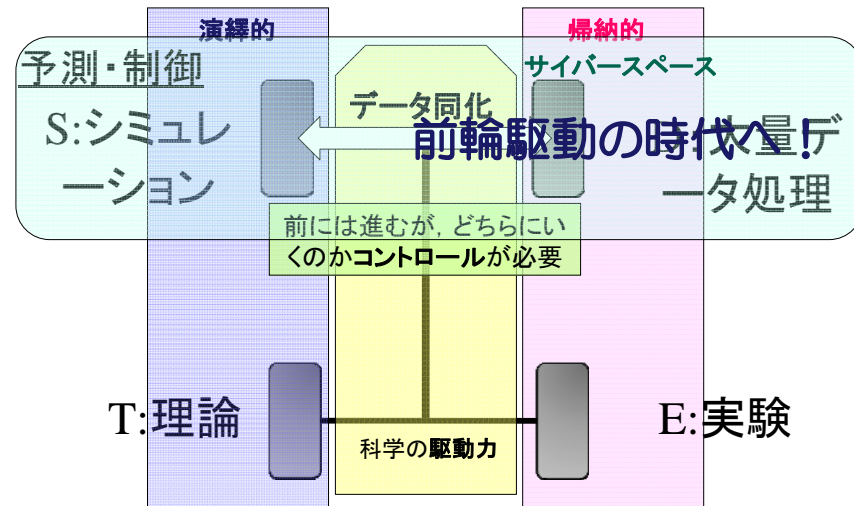
- 団塊の世代のエキスパート(熟練者)がどんどん退職。
- 微妙な判断を確実にできる能力を伝承していくために, ノウハウ習得用のマシンをつくれる。

# 話題3. シミュレーションを取り込み融合するデータ同化

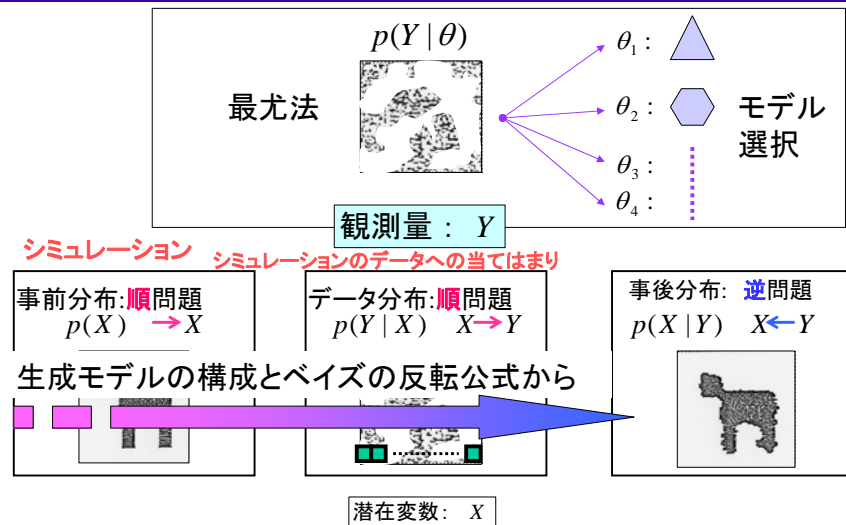
- 気象学・海洋学の分野で発達
- 物理数値シミュレーションモデルと実際の観測を統合する手法
  - シミュレーションのみでは適切に物理現象を再現できない
    - シミュレーションモデルには、モデルの不完全性や境界条件が正確にはわからないなどの不確かさが存在
    - (たとえば) 正確な気象予測には適切な初期条件の構成が必要
  - 観測データは物理的・社会的制約により得られる情報に限界がある

→ 観測データを用い数値シミュレーション内の変数を修正  
= データ同化

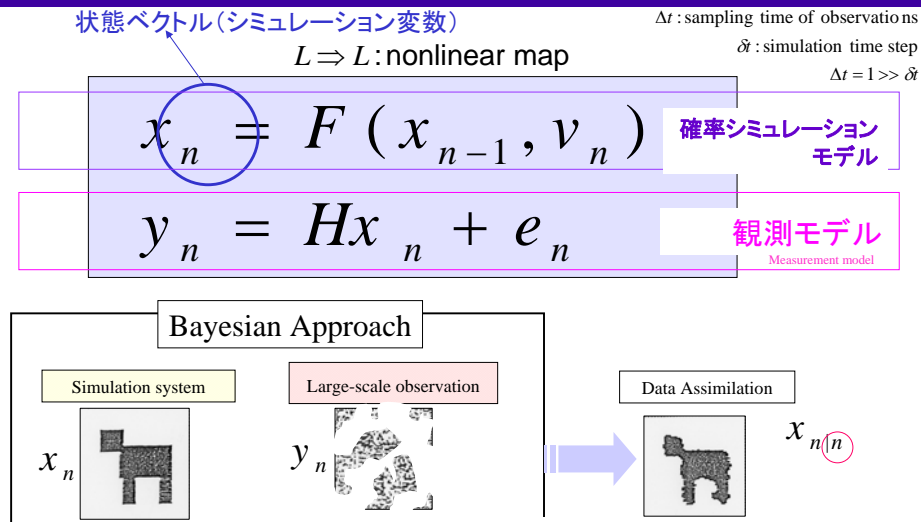
# TESD: 第4の科学, 第4の方法論



# 生成と反転, そしてデータ同化へ



# シミュレーションとデータ同化



<単一パスから分布へ>

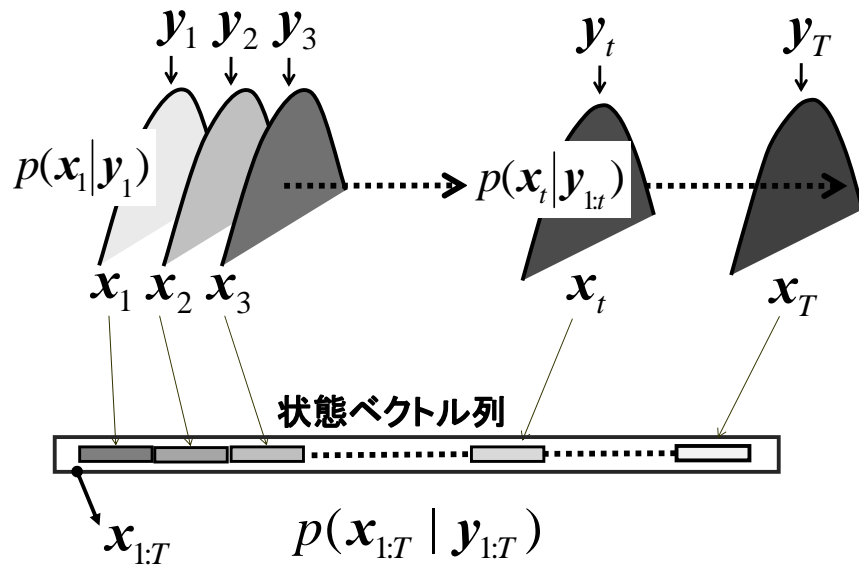
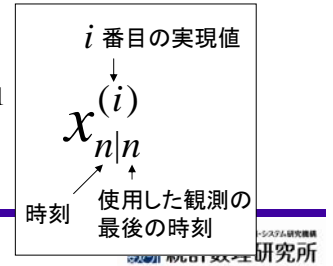


## 逐次データ同化の手法

- 逐次データ同化においては、アンサンブルカルマンフィルタ(EnKF)がよく用いられる。
  - 条件付分布を実現値の集合(アンサンブル)として保持
  - フィルタリングにはカルマンフィルタの式を利用
- 粒子フィルタ(Particle Filter (PF))の応用例は少ない。
  - 実現値集合を用いる点でEnKFと同様

$$p(x_n | y_{1:n-1}) \cong \frac{1}{N} \sum_{i=1}^m \delta(x_n - x_{n|n-1}^{(i)}) \quad \left\{ x_{n|n-1}^{(i)} \right\}_{i=1}^m$$

$$p(x_n | y_{1:n}) \cong \frac{1}{N} \sum_{i=1}^m \delta(x_n - x_{n|n}^{(i)}) \quad \left\{ x_{n|n}^{(i)} \right\}_{i=1}^m$$



## 研究の実施体制: [ ] が主担当者

5分野 +  $\alpha$  で進行中

共同研究者

### ・ [中野] リングカレントの3次元構造の解明

海老原祐輔(名大), M.-C Fok(NASA)  
大谷晋一, P.C.Brandt(ジョンホプキンス大)



### ・ [上野] 気象・海洋結合

鍵本崇(地球環境フロンティア研究センター)  
広瀬直毅(九州大)



### ・ [中村] 津波データ同化

広瀬直毅(九州大), B.H. Choi (Sungkyunkwan大学(韓国))



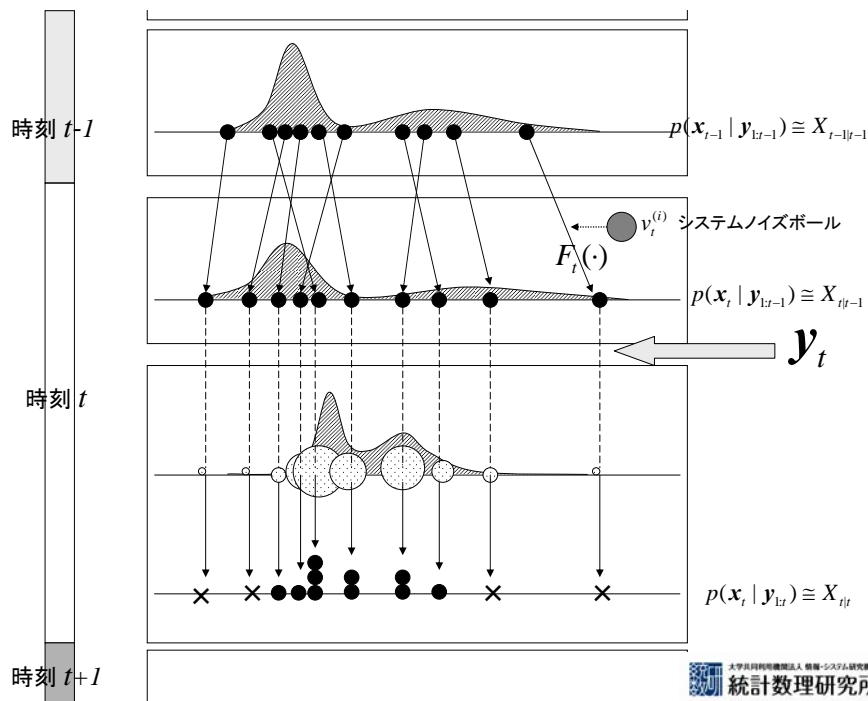
### ・ [稲津] 潮汐データ同化

佐藤忠弘、三浦哲(東北大) 他

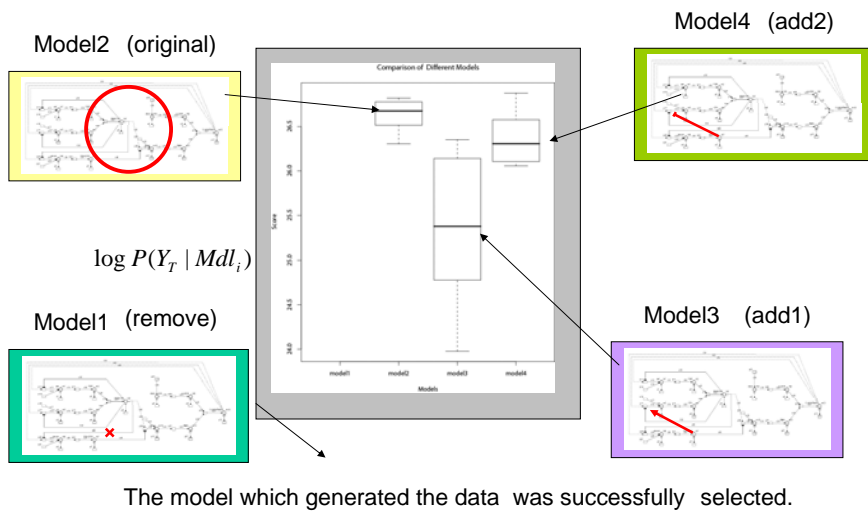


### ・ [吉田] ゲノム情報(GDA project)

東大・医科研・宮野教授研究室



# 最尤法にもとづく構造学習



# “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

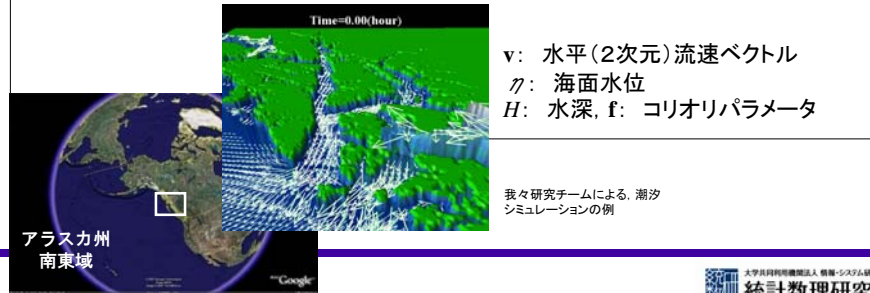
“個”によって異なる形状、形態情報をシミュレーションモデルに取り込む『メタシミュレーションモデル』

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

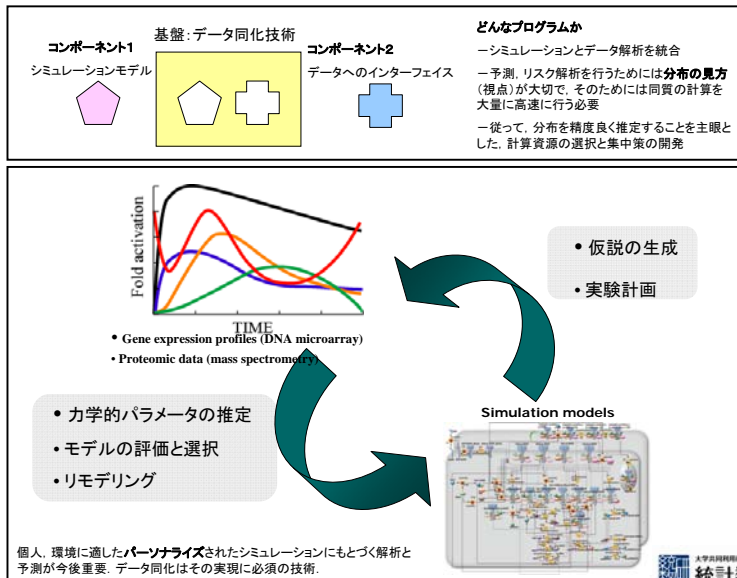
連続式: 
$$\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial x} (\mathbf{v} H) = 0$$

海底摩擦係数  $\gamma_b$ : 0.006

水深  $H$



# 知識循環と永続的なモデルの改良



# ご静聴ありがとうございました。



Email: [higuchi@ism.ac.jp](mailto:higuchi@ism.ac.jp)

Homepage:

<http://www.ism.ac.jp/~higuchi/>

<http://daweb.ism.ac.jp/>

参考文献:  
 北川・樋口、「予測とモデル」, 数理科学(1998)  
 樋口, 日本知能情報ファジー学会 (2008).  
 佐藤・樋口 日本統計学会誌 (2008)