NTT

NTT Communication Science Laboratories

NII軽井沢土曜講話会 2011年11月4日

(於:軽井沢国際高等セミナーハウス)

統計的機械学習入門

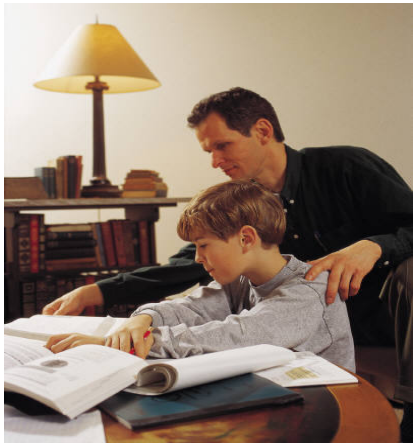
NTTコミュニケーション科学基礎研究所

上田 修功

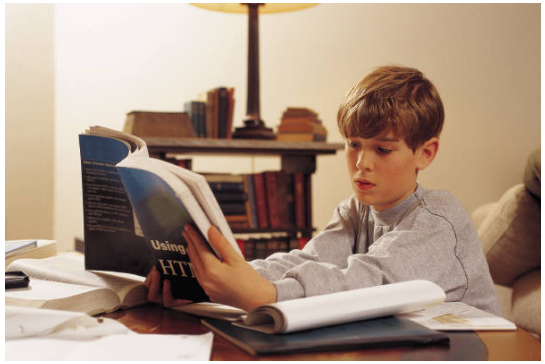

機械学習って何?

どんな研究？

人の学習に例えると…



教師あり学習
（先生に習う）



教師なし学習
（自習する）

半教師あり学習
（膨大な情報を
活用する）

アンサンブル学習
（皆で教え合う）

フォーマルには…

機械（情報処理システム）に学習能力を持たせる技術

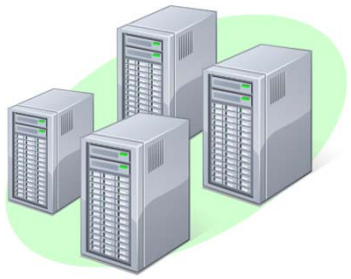
入力

文字認識 7210
画像

音声認識
音声信号

機械翻訳 This is a pen.
英語

情報処理システム



出力

7210
文字

こんにちは
文（文章）

これはペンです。
日本語

学習能力：所与のデータ（学習データ）だけでなく、未知のデータ（テストデータ）でも性能を発揮する汎化能力

汎化誤差の最小化が実用上重要

情報処理＝情報変換(価値創造)

$$y = f(x;\theta)$$

入力 x ⟹ y 出力(目標値)

汎化誤差
(平均損失)

$$E\{L\} = \int L(y, f(x))\, p(x, y)\, dxdy$$

損失関数(目標値とのずれ)

損失関数の典型例

$L(y, f(x)) = \| y - f(x) \|^2$ 自乗誤差 ➡ 最小自乗法

$L(y, f(x)) = -\log p(x, y)$ 対数損失 ➡ 最尤推定法

応用分野

パターン認識（文字認識、音声認識、テキスト分類 etc)

データマイニング(協調フィルタリング、スパムフィルター etc）

自然言語処理 (機械翻訳、固有表現抽出 etc）

バイオインフォマティックス・脳科学（DNA解析、BMI etc）

さらには、農業、経済、保健医療…

機械学習はICTの**汎用要素技術**

機械学習研究の動向

1980年代　**萌芽期**

多層ニューラルネット, SOM
（誤差逆伝搬法）

計算論的学習理論
Probably Approximately Correct (PAC)learnability

1990年代　**成長期**

潜在変数モデル
（階層、混合モデル、EM法）

汎化誤差理論
（モデル選択、アンサンブル学習）

カーネル理論

SVM（サポートベクトルマシン）

2000年代～　**発展期**

ベイズモデル
（変分ベイズ、Gibbsアルゴリズム）

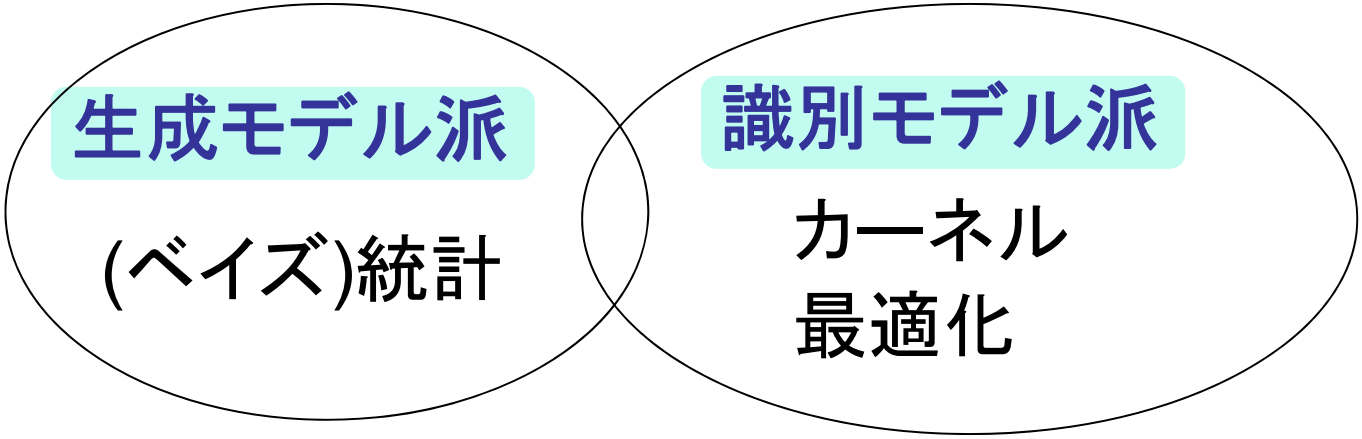
ベイズ理論
（ノンパラメトリックベイズ理論）

各種学習タスクの提案
（転移学習、半教師有り学習等）

(離散)最適化理論
（劣モジュラー関数）

機械学習の2大派閥？

確率分布の仮定の有無により
2つのアプローチに大別できる



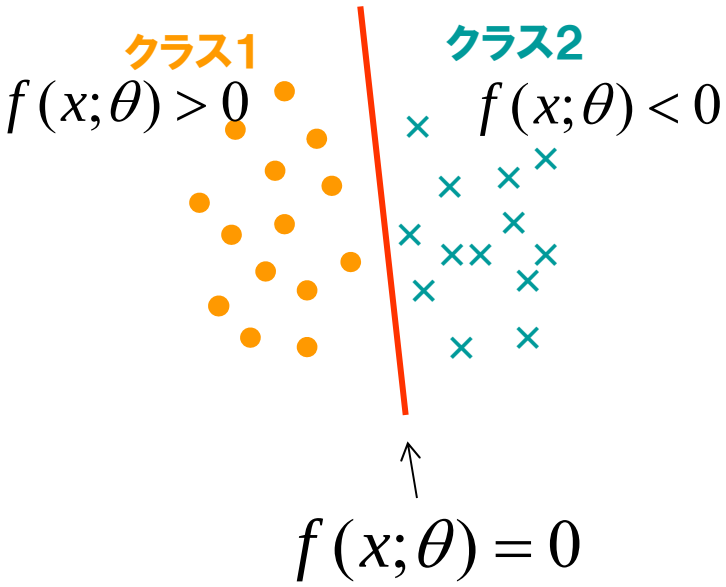
絶対的な優劣はなく、タスクとの
親和性と研究者の好みに依存する

識別モデル vs. 生成モデル

分類問題の場合

識別モデル

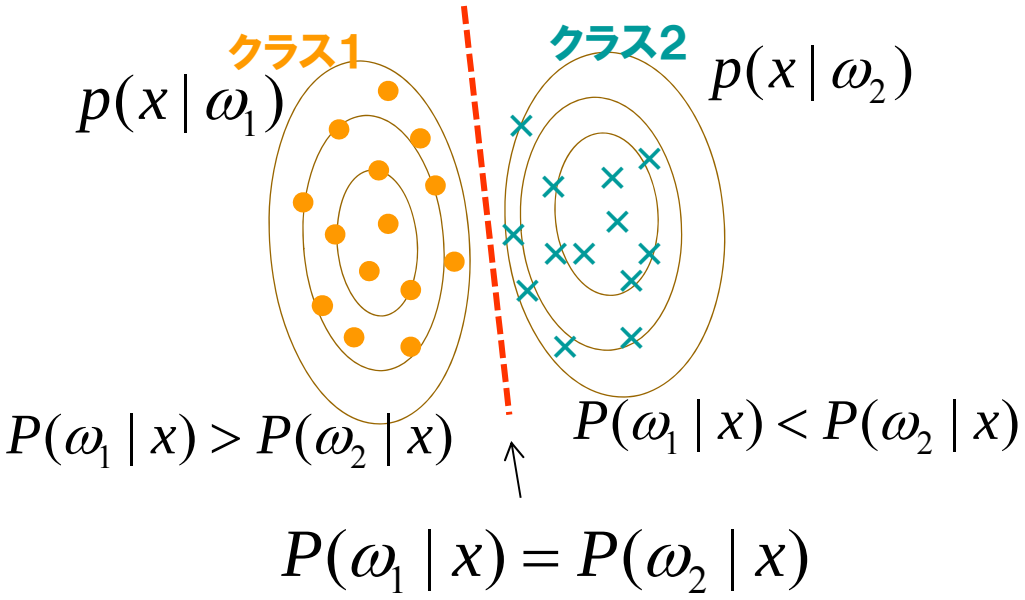
データのクラス境界
を直接学習



データの生成過程は考慮せず、
所与のデータで問題を直接解く

生成モデル

クラスごとの生成モデル
を確率分布として学習



クラス事後確率により識別
（クラス境界は結果として得られる）

識別モデルの代表例（SVM）

1998年にVapnik（ベル研）が考案した2分類器

どこが斬新？

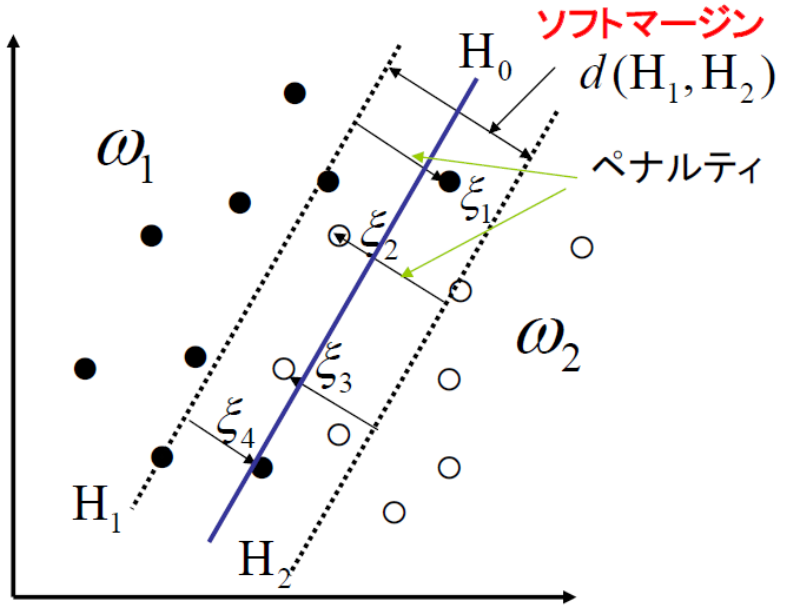
従来法：高次元の特徴ベクトルを**次元圧縮して，非線形識別関数**で識別

⇕

SVM：高次元の特徴ベクトルを**さらに高次元に非線形写像**し，**線形識別関数**を**新基準**で学習

マージン最大化基準

高次元空間で線形識別関数を学習



Note: カーネルトリック： $K(x_i, x_j) = \phi(x_i)^t \phi(x_j)$

講演内容

１．各種機械学習タスク

汎化誤差最小化の工夫は？
多様な実問題での学習タスクは？

２．生成モデルアプローチ

統計的機械学習の基礎（潜在変数
モデルとその周辺技術）

潜在的な情報をどうモデル化するか、また、
モデルの複雑さをどうコントロールするか?

Part I 各種機械学習タスク

アンサンブル学習

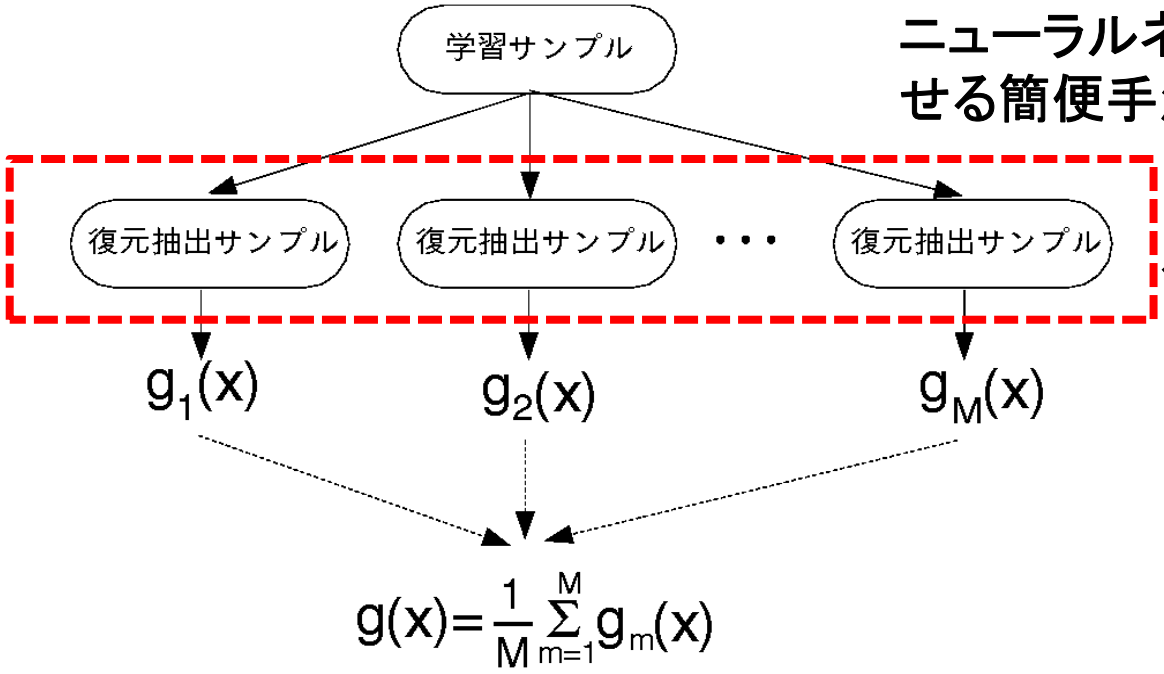
3人寄れば文殊の知恵？

アンサンブル学習の形態

（１）複雑な学習機械を独立に学習した後，
それらをアンサンブルする形態
（Perrone,1993; Breiman, 1996）

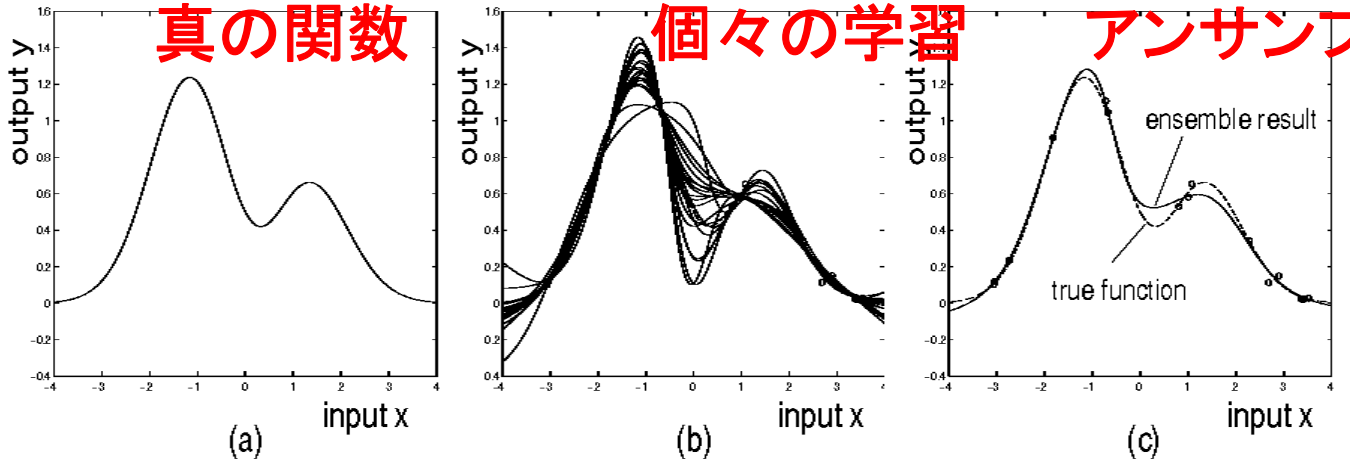
（２）単純な学習機械を逐次的に学習した後，
それらをアンサンブルする形態
（Freund & Shapire, 1997）

Bagging (Bootstrap Aggregating) 法 (Breiman, 1996)

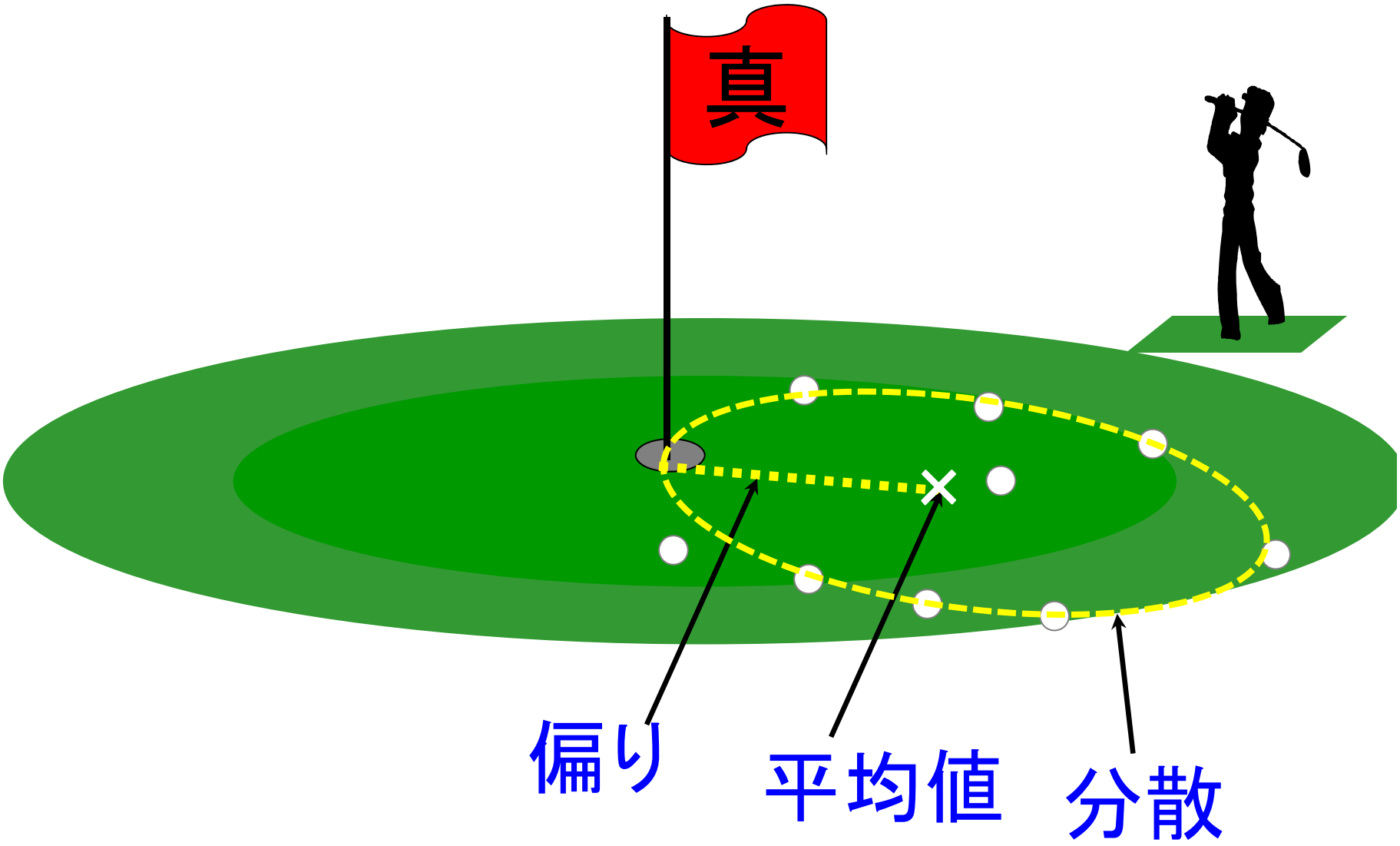


ニューラルネットの汎化誤差を減少させる簡便手法として注目された

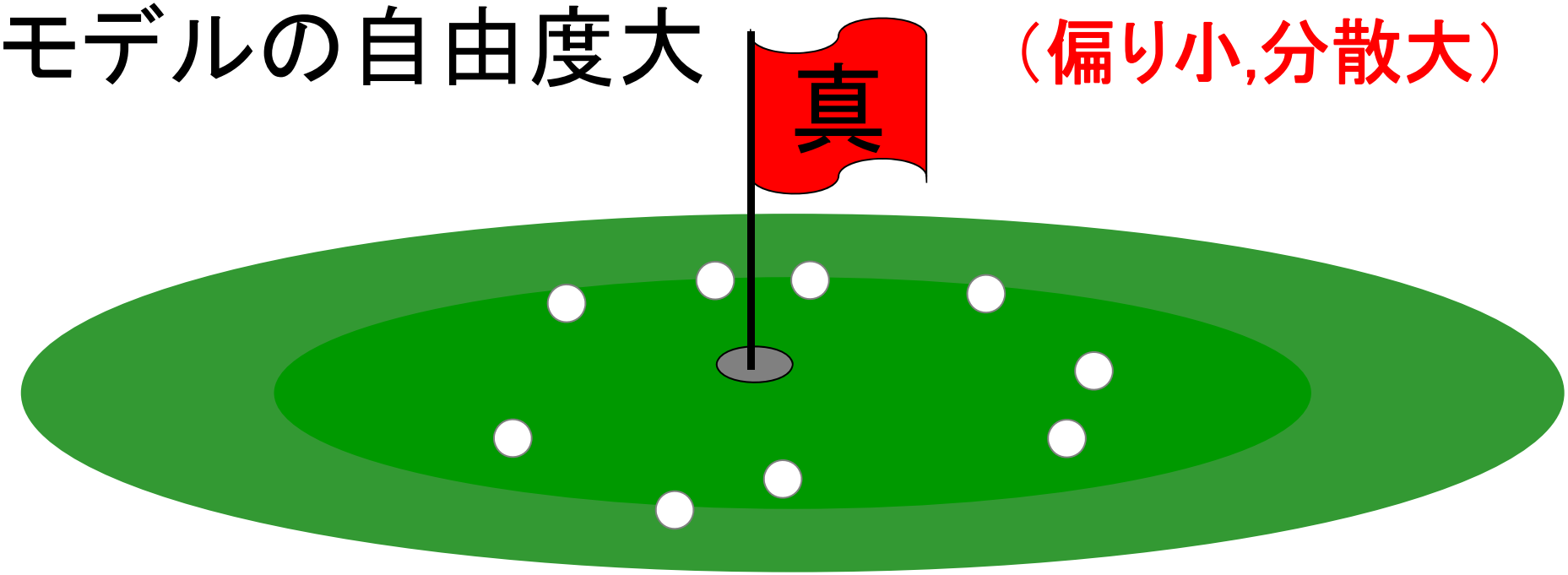
復元抽出によりサンプルセットを複数作成



最小自乗推定値の信頼性

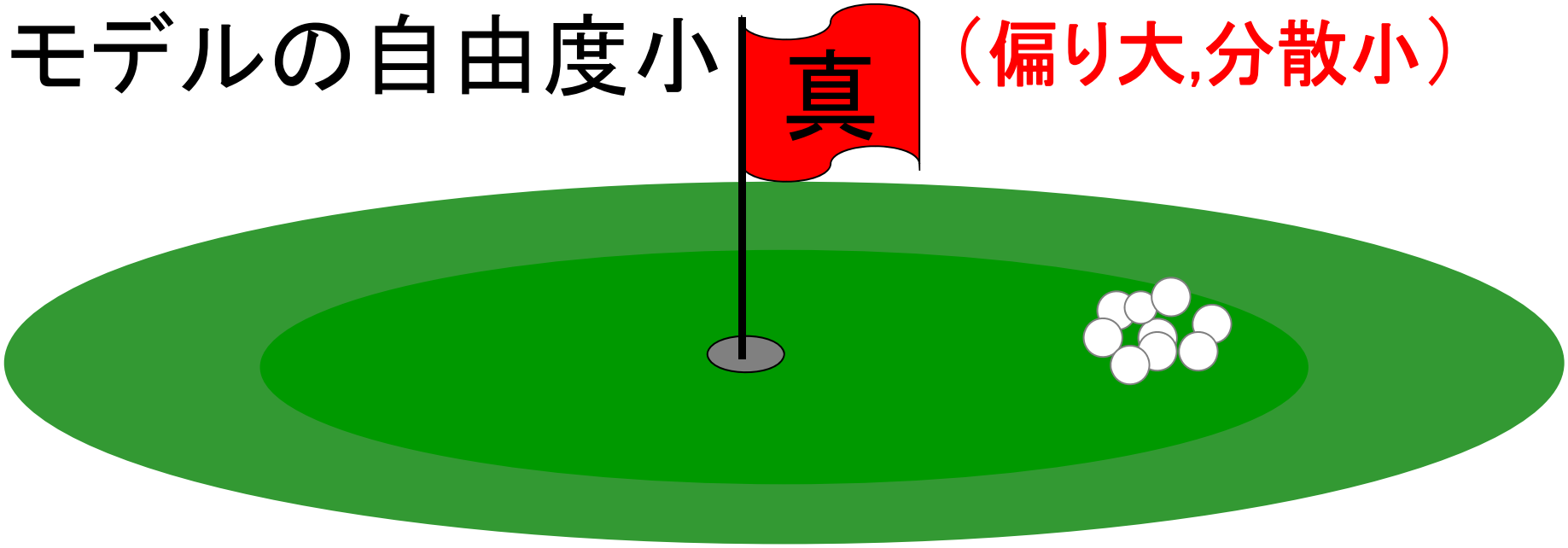


モデルの自由度大



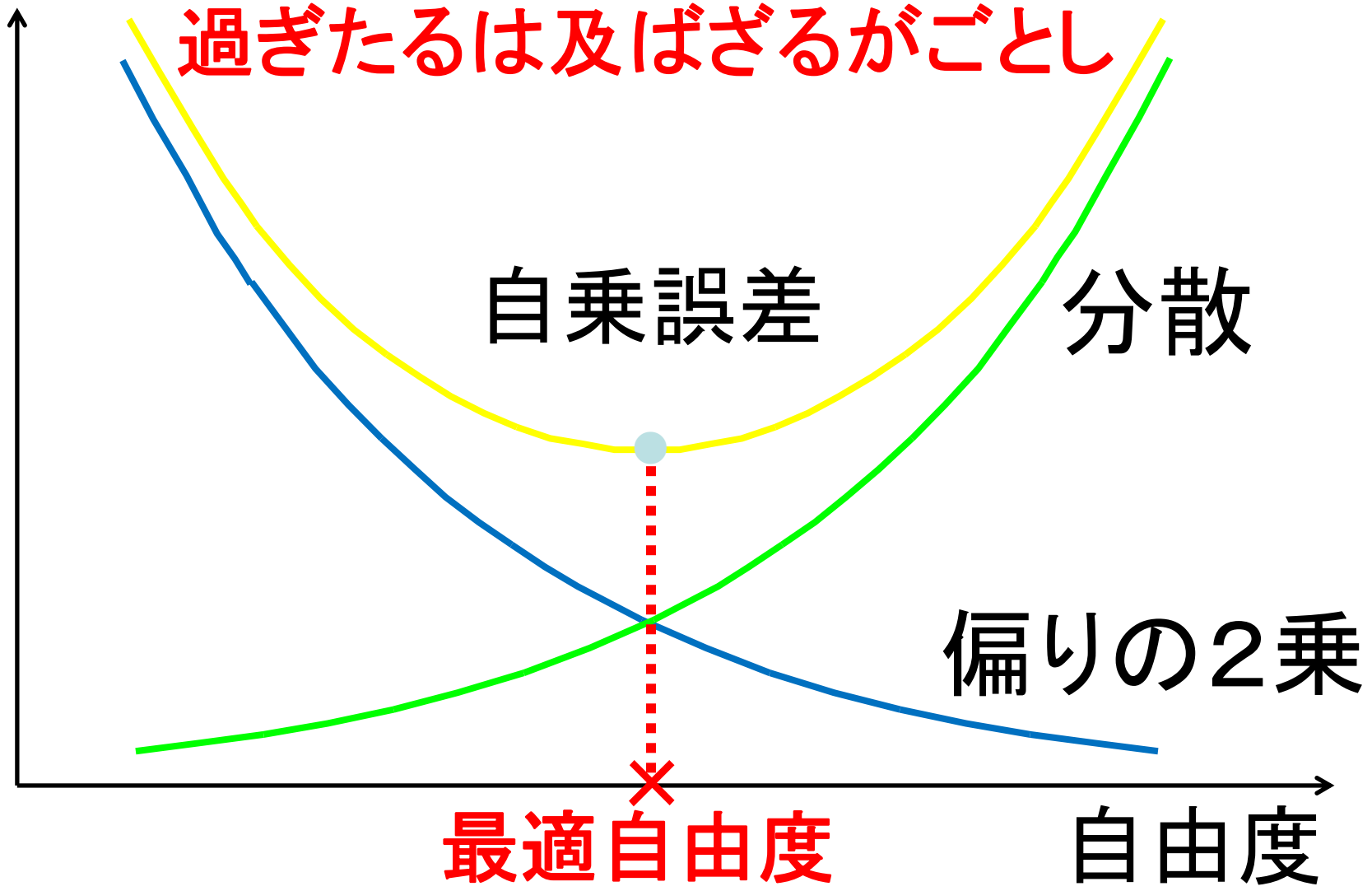
（偏り小,分散大）

モデルの自由度小



（偏り大,分散小）

偏りと分散のジレンマ



Note: 一般に偏りは未知（推定困難）

単純*M*アンサンブルの汎化誤差解析

(Ueda,1997)

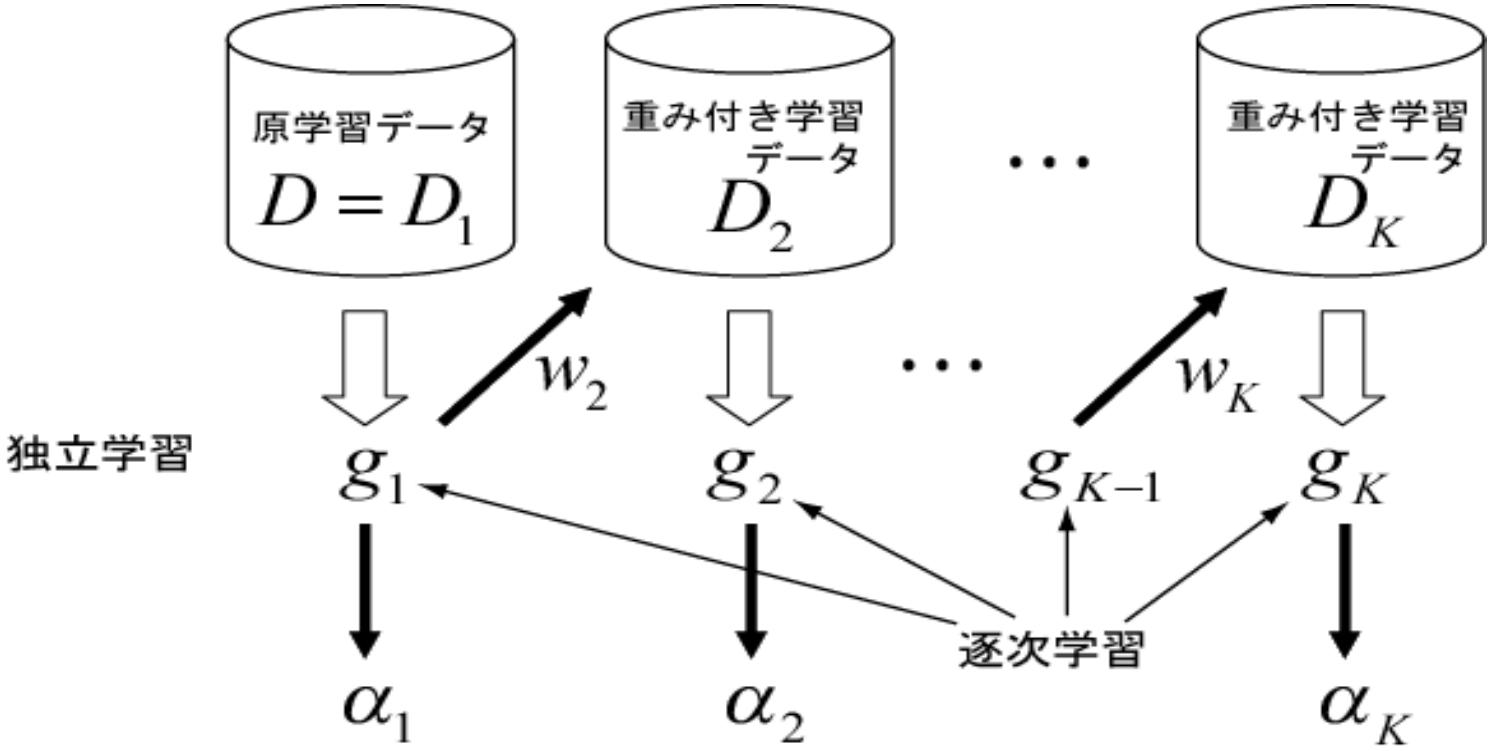
$$\mathrm{GErr}(f_{ens}^{M}) = E_{X_o}\left\{\frac{1}{M}\overline{\mathrm{Var}}(X_0) + \overline{\mathrm{Bias}}(X_0)^2 + \left(1-\frac{1}{M}\right)\overline{\mathrm{Cov}}(X_0)\right\} + \sigma^2$$

バリアンス（分散）　バイアス（偏り）　コバリアンス（共分散）

- 分散は1/Mに減少するが，**偏りは変化しない**
- 予測器の出力間に相関がある場合，**正の相関は、汎化誤差の減少を妨げる**

“3人寄れば文殊の知恵、**4人寄れば烏合の衆**”

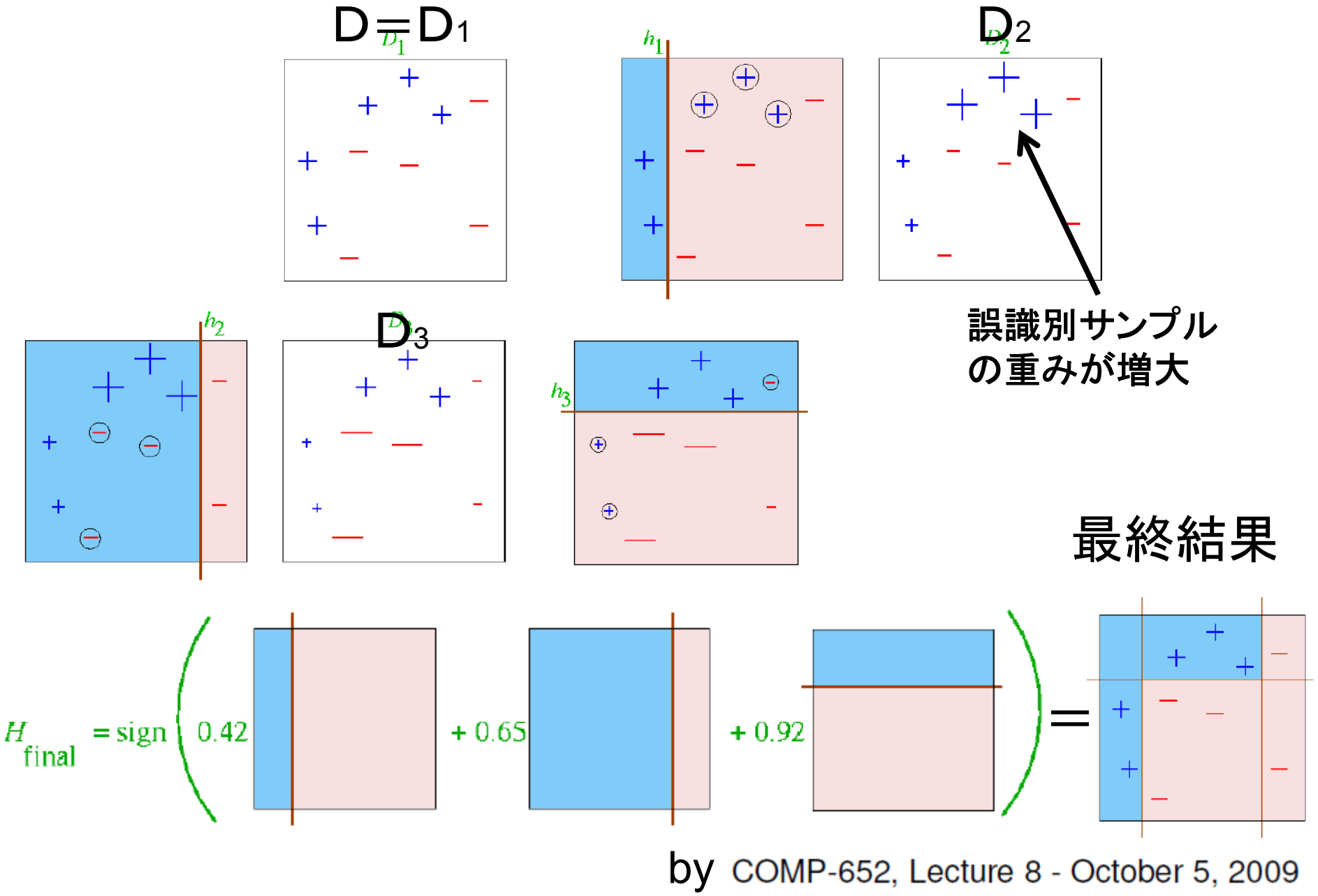
ブースティング法



$$g(\boldsymbol{x}) = \sum_{i=1}^{K} \alpha_i g_i(\boldsymbol{x})$$

識別関数クラス判定 $g(\boldsymbol{x}) \gtrless_{\omega_2}^{\omega_1} 0$

ブースティング法の処理の流れ

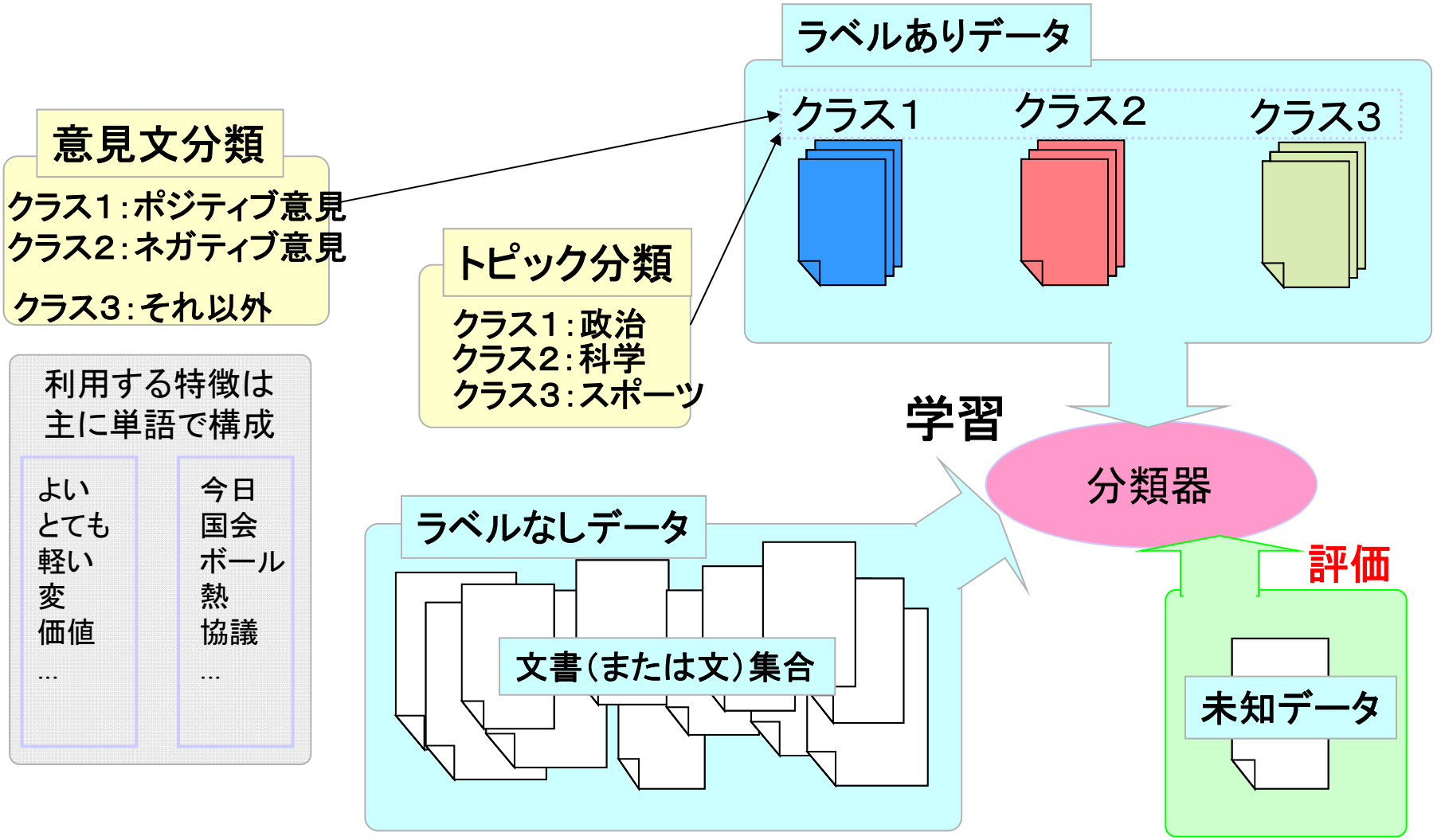


by COMP-652, Lecture 8 - October 5, 2009

半教師有り学習

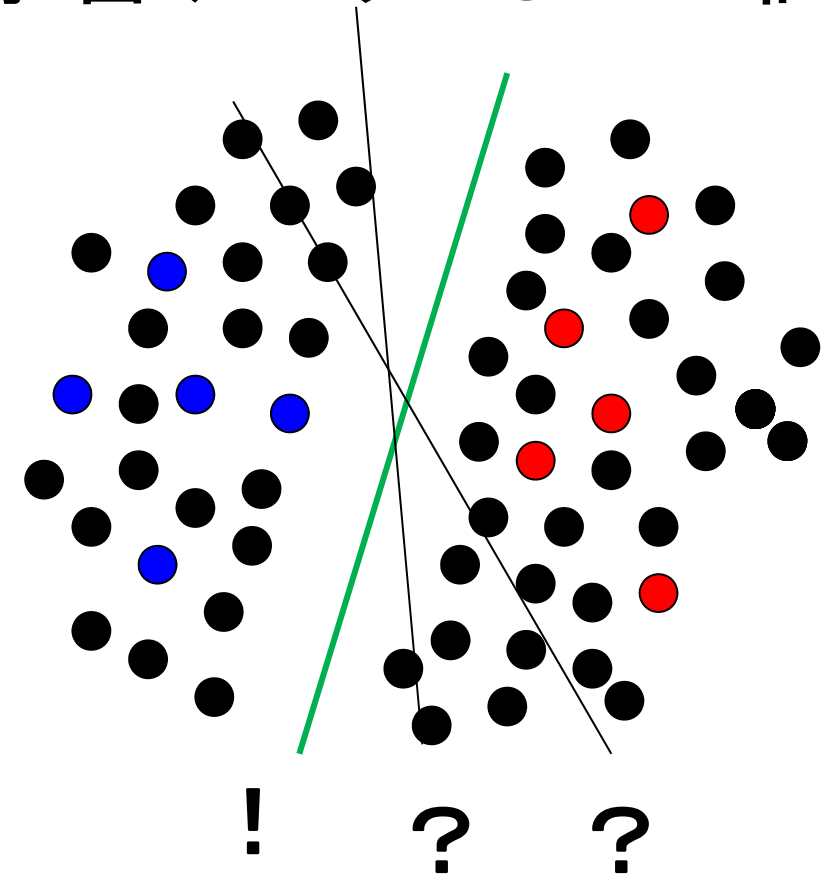
タダより安い?ものはない!

テキスト分類問題の場合



ラベル無しデータの有用性

- 収集が低コスト(almost free)
- 学習データとしての価値は？



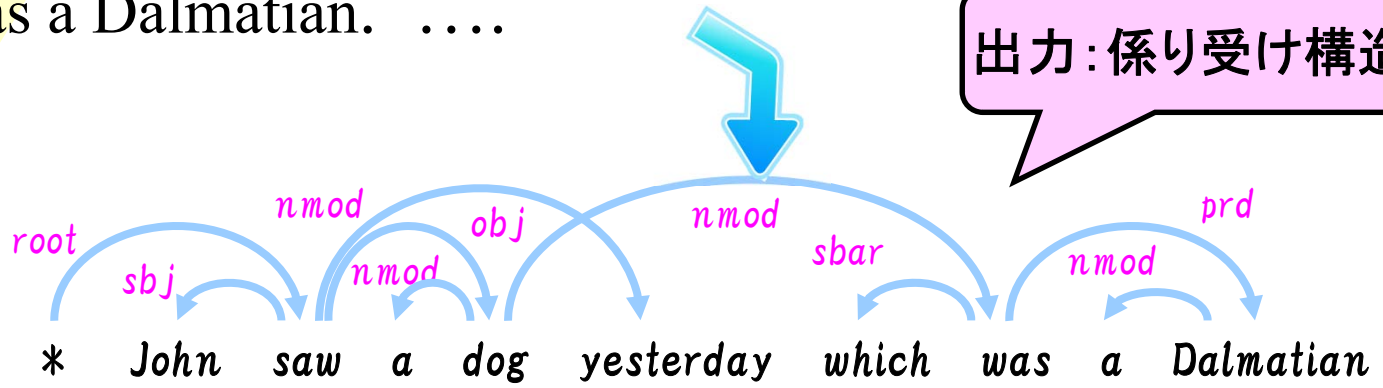
クラスラベルが
なくても十分有用！

自然言語処理での応用（係り受け解析問題）

…. John saw a dog yesterday which was a Dalmatian. ….

入力：文章

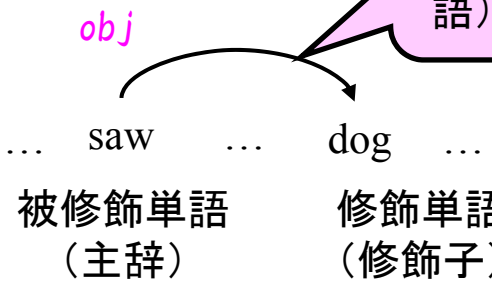
出力：係り受け構造



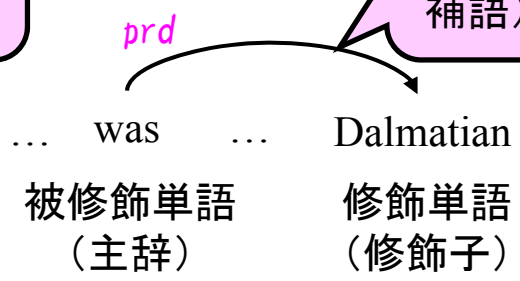
- 文内の二単語間の修飾・被修飾関係を推定する問題

例

「dog」が「saw」を係り受け関係「obj（目的語）」で修飾



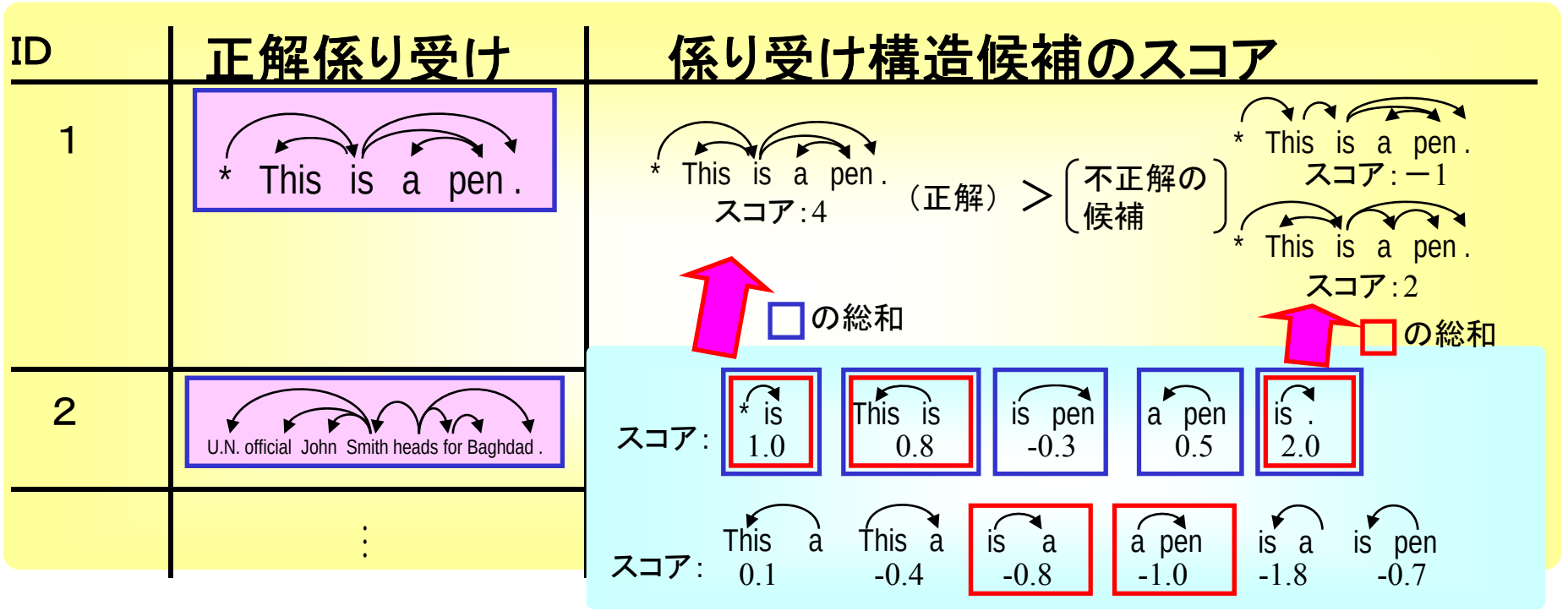
「Dalmatian」が「was」を係り受け関係「prd（叙述的補語）」で修飾



係り受け解析モデルの学習

J. Suzuki, 2009

- 係り受け構造へのスコアリング
  - 入力文に対する係り受け構造の候補の中で正解係り受け構造のスコアが最も高くなるようにスコアを与える
    - 係り受け構造のスコア　＝　実際には二単語間の係り受け関係のスコアの総和
  - 正解データ内の全ての入力文で上の条件が満たされるように実施



係り受け解析モデル学習時に大量かつ容易に獲得可能なラベルなしデータを利用

	構成内容	獲得コスト	量
正解データ	文 ＋ 係り受け構造	高コスト 専門家が係り受け構造を人手で付与	少量 （新聞記事数日分）
ラベルなしデータ	文 （普通のテキスト）	低コスト 電子化文書、webテキストから獲得	大量 （新聞記事数十年 正解データの数千倍以上

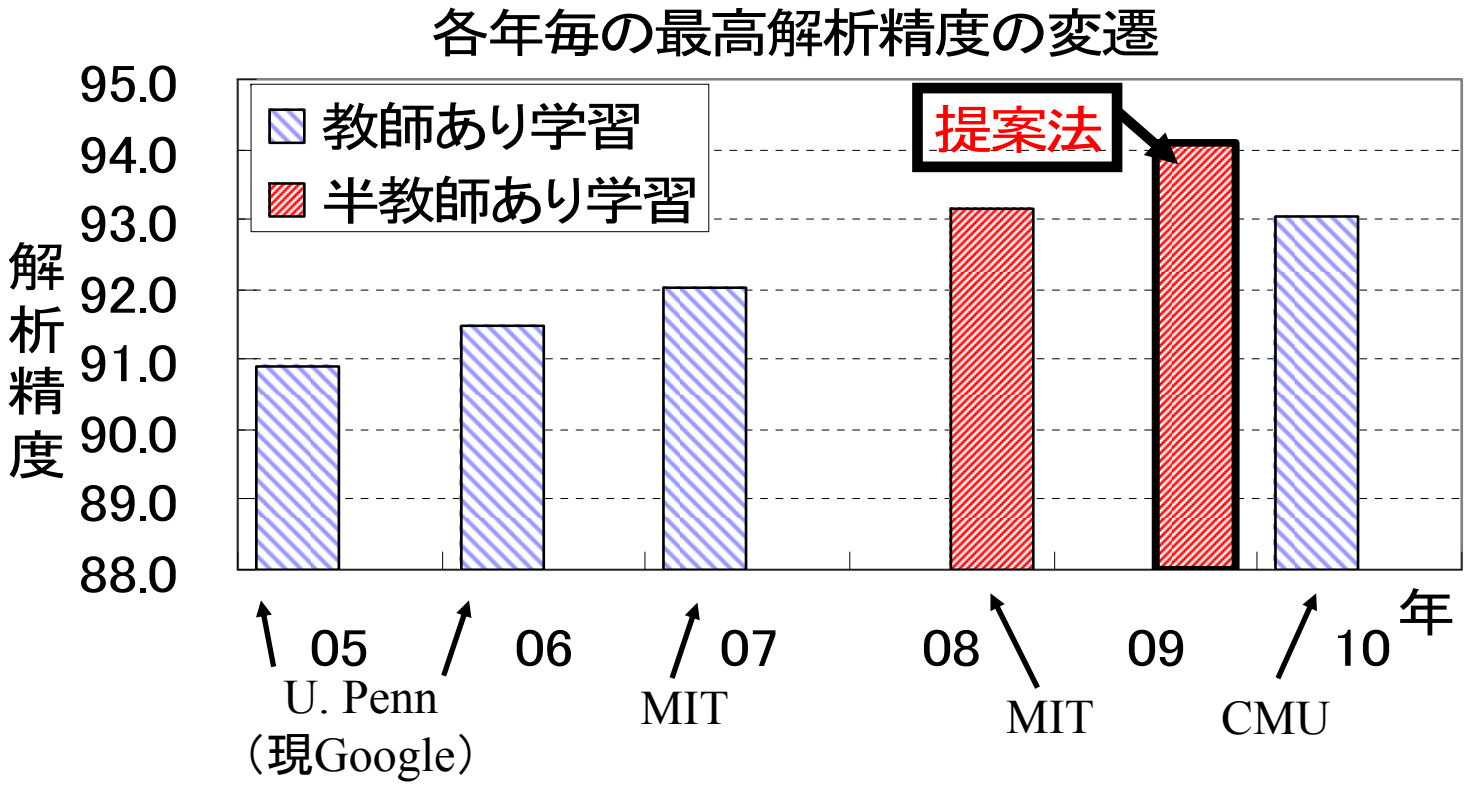
どのように利用するか

1. ラベルなしデータに含まれる全ての二単語間の係り受け確率を推定

2. 求めた二単語間の係り受け確率を特徴量とし係り受け解析モデルを学習

- 係り受け解析タスクの国際標準ベンチマークテストデータ（英語、チェコ語）でトップの成績（2011年1月現在）

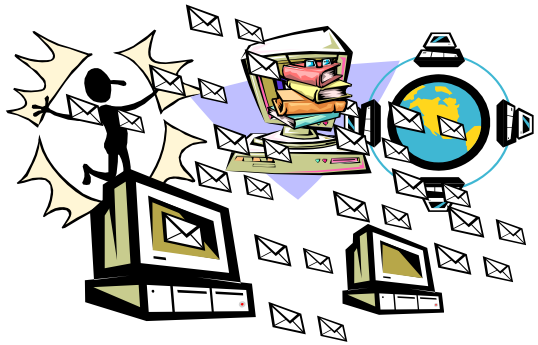
  - 正解データ量　　約100万単語（新聞記事数日分）
  - ラベルなしデータ量　　約370000万単語（新聞記事60年分:
    正解データの3700倍）



マルチラベル学習

テキストの多重分類モデル, Ueda et al., 2002

膨大な文書



インターネット，電子図書館，電子メール等

多重トピック分類

何故難しいのか?

(1) 不定形，高次元で粗なデータ

⇨ 従来の学習手法が苦手なデータ

(2) 多重トピック (多クラスかつ**多重ラベルデータ**)

⇨ 排他的分類を行う伝統的なパターン認識問題とは異なる

テキストの多重トピック学習問題

語彙が膨大故，ニューラルネット(NN)の様な関数近似法では学習困難

学習データ

	単語頻度 x	トピックス y
	野球 選手 試合 …	野球 選手 試合 …
文書1	2 0 0 1 0 0 0 •••	0 0 1 1 0 0 1 •••
文書2	0 1 1 0 0 0 0 •••	1 1 0 0 0 0 0 •••
文書3	2 0 0 1 0 0 0 •••	0 0 1 0 0 0 0 •••
⋮	⋮	⋮
文書n	2 1 0 1 0 3 0 •••	1 1 0 0 0 0 0 •••
⋮	⋮	⋮

正当率ではな抽出率（F値）で評価

同一のトピックスでも単語の分布は異なるため，k近傍法（kNN）の様なメモリベースの方法では限界

各トピックで見ると殆ど零故，従来の2分類手法（Naive Bayes: NB法や Support Vector Machine: SVM法）では限界

学習：(x, y) から $f : x \to y$ を推定

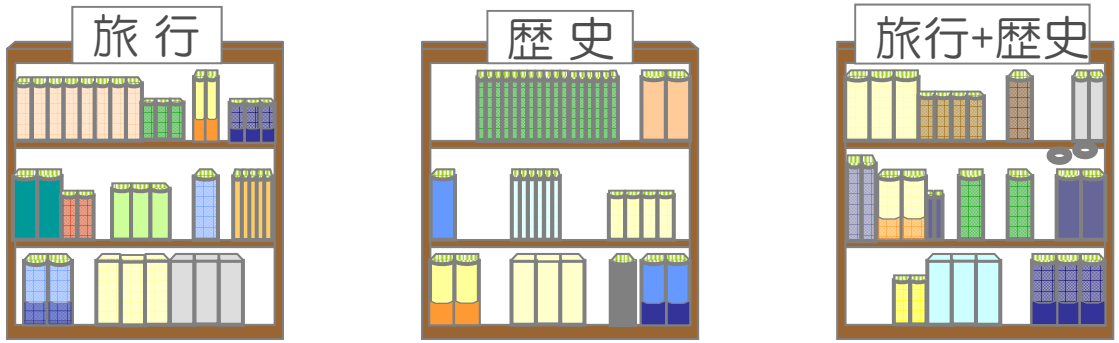
予測：　新たな x に対し，$\hat{y} = f(x)$ を求める

いま何故多重分類か？

図書館の例

従来の図書館では，実体として一冊しかない本は本棚のどこか一箇所に収納して分類しなければならない
⇒　本質的に単一分類

これから主流になる電子図書館では，本が電子化されているため，リンク等を駆使することによって，
何箇所にも同時に分類できる　⇒　多重分類が主流に



図書館の係りの人が・・・

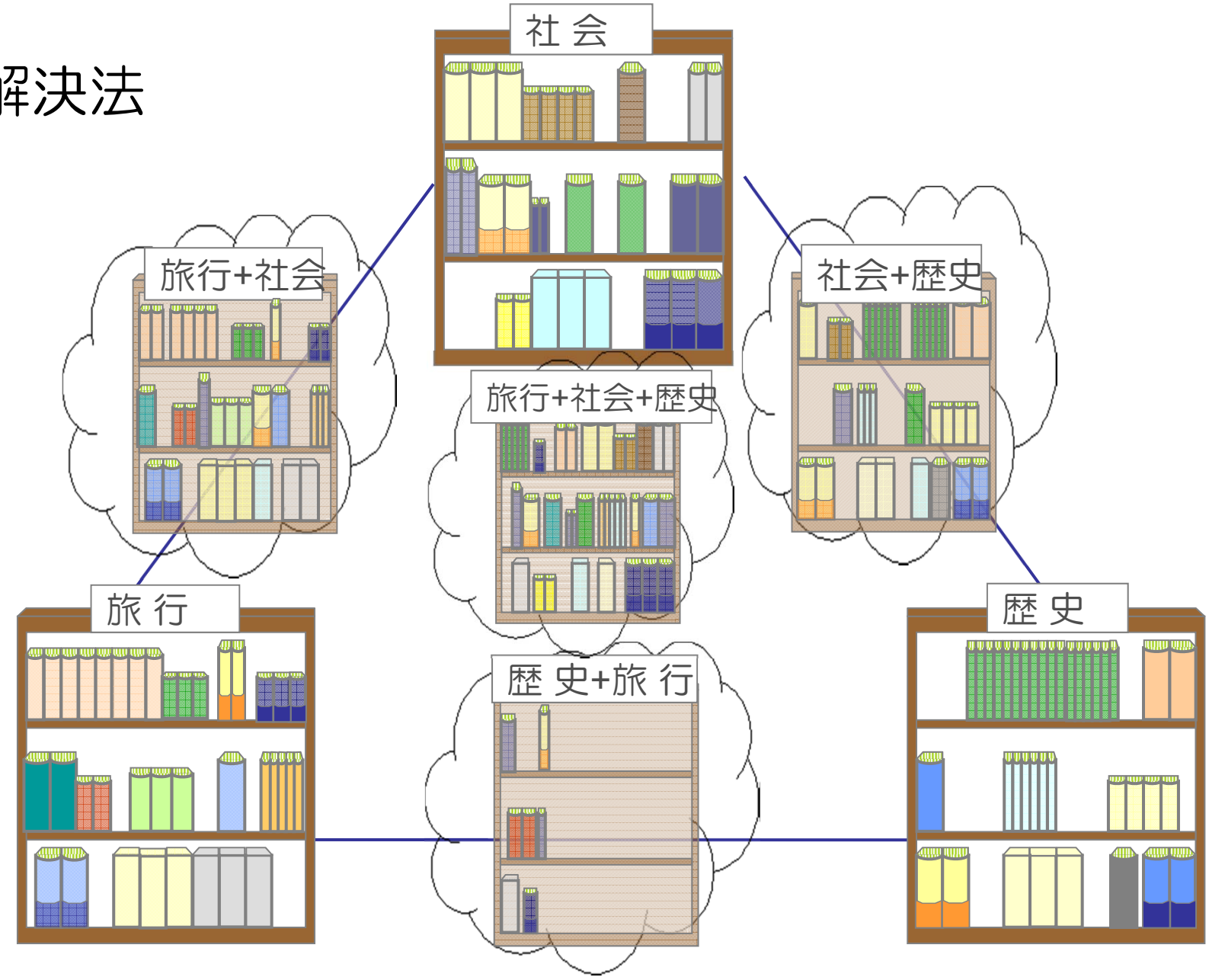
情報（書籍・ニュース・ホームページ等）は複数のジャンルにまたがる!!

物理的な本では不可能
（書籍が複数必要・書棚の限界）

デジタル情報では可能？

しかし、ジャンルが50あるとすると、組合せジャンルは
約1000兆個にもなる・・・

解決法



組合せジャンルを仮想的に生成

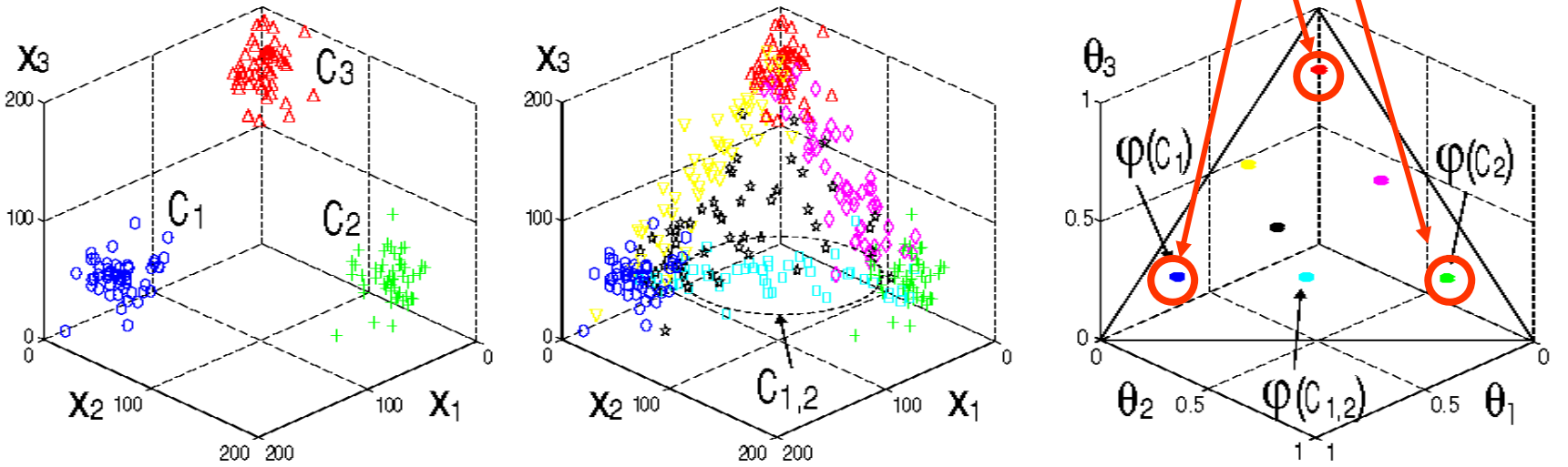
Conventional mixture models VS. PMMs

$$\alpha_1 M(\theta_1) + \alpha_2 M(\theta_2) \qquad M(\alpha_1\theta_1 + \alpha_2\theta_2)$$

Exhaustive models VS. PMMs

$$M(\theta_1), ..., M(\theta_{2^L-1}) \qquad M(\varphi(y\ ; \theta_1, ..., \theta_L))$$

V=3



(a) Single topic (b) Multi-topic (c) PMM

解決法

単一ジャンルの最適組合せを推定

新撰組
旅行
京都
池田屋
沖田総司
幕末
旅館
名所
ツアー
訪ねる
・
・

→

旅行

参加する
海外
ツアー
飛行機
見る
旅館
名物
費用
添乗員
楽しい
・
・
・

＋

歴史

戦争
ローマ
支配する
明治維新
憲法
ピラミッド
ヒロシマ
ナイル川
司馬遼太郎
幕府
・
・
・

＋

社会

テロ
学校
差別
ルール
福祉
運動
選挙
改革する
自治会
産業
・
・
・

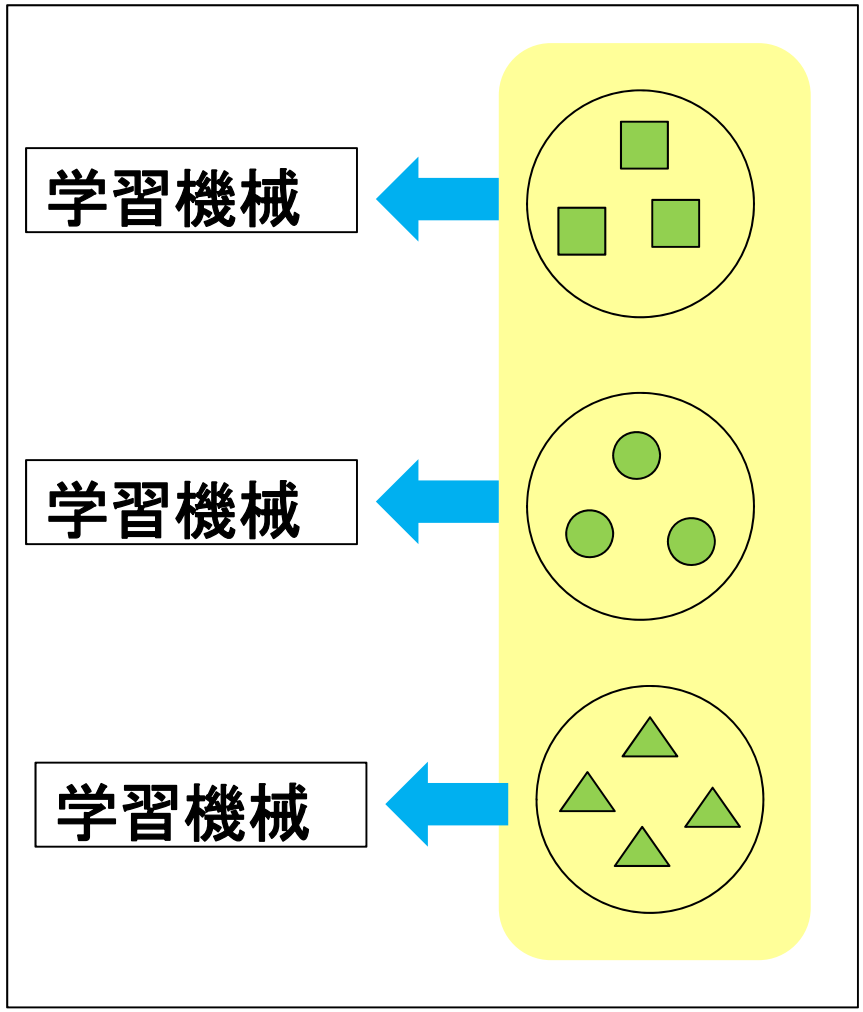
多重分類を高速・高精度に実現する パラメトリック混合モデル(ＰＭＭ)[注]の考案

- 分類済みデータから多重トピック文書中の単語の確率分布を高速に学習
  （１万文書をパソコンで約１分）
- 推定分布の数学的最適性を理論保証
- ｗｅｂデータを用いた多重トピック分類実験において世界一の分類性能を確認
  人手による作業 → 完全自動化

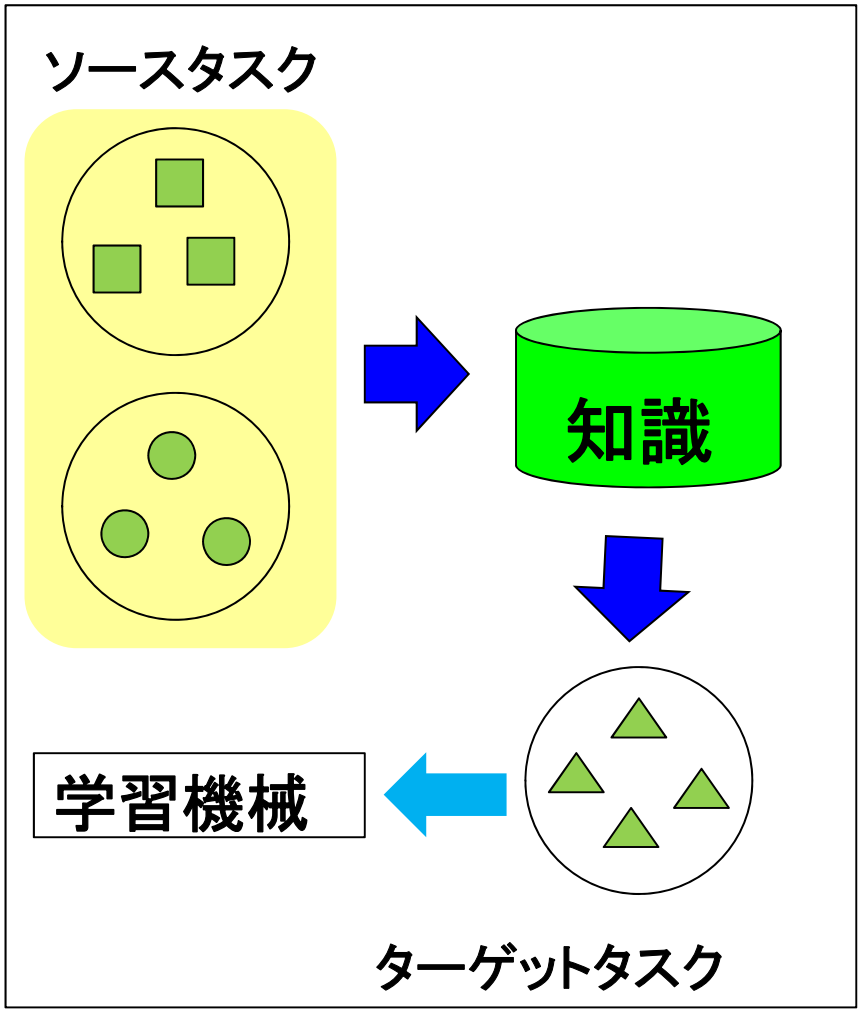
（注）PMM：Parametric Mixture Models

マルチタスク学習
転移学習

異なる類似タスク
を独立に学習



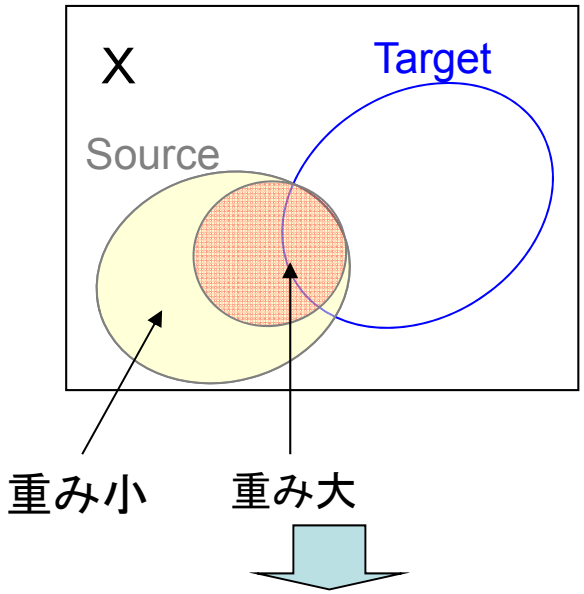
類似タスクの
知識を伝搬



転移学習の各種アプローチ

学習データ:
元ドメインのラベルありデータ＆目標ドメインのラベルなしデータ

- **データの重み付け**
  - ラベルありデータ:**密度比推定**
    (Sugiyama+(2005) etc.)
  - ラベルなしデータ:Bootstrapping
    (Jiang+(2007), Wu+(2009))

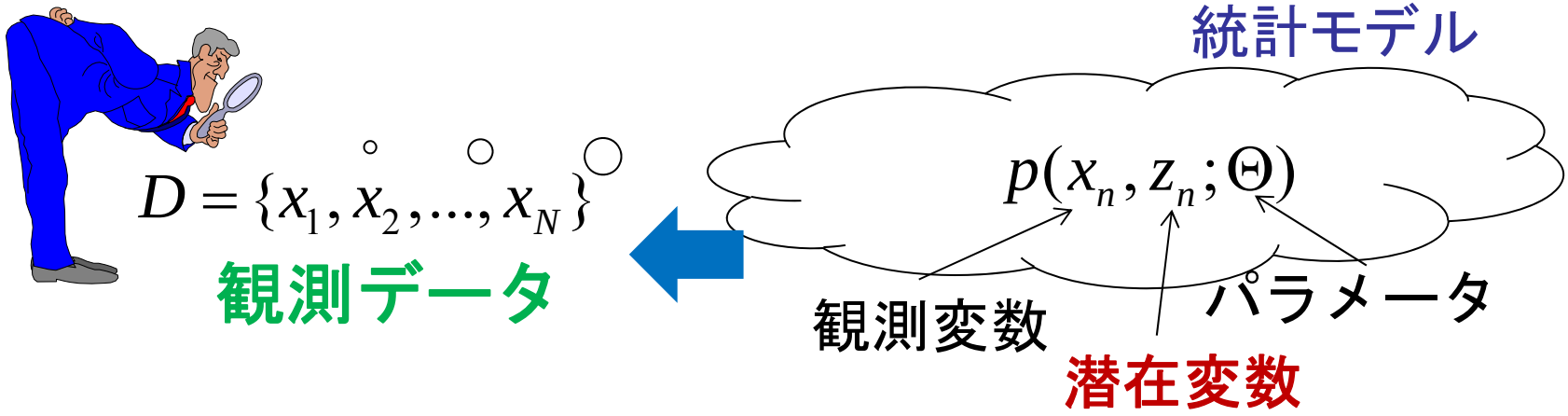


- **特徴空間の変換**
  - 特徴選択(Blitzer+(2006) etc.)
  - グラフの利用(Pan+(2009) etc.)

既存の教師あり・半教師あり学習
アルゴリズムと組み合わせて実行

Part II 生成モデルアプローチ

生成モデルの構成要素



生成モデル(generative model) :
観測データがどのような確率モデル(の系列)として
生成されたかの仮説

潜在変数(latent variable(s)) :
本来観測されない隠れ変数

パラメータ(parameter(s)) :
確率分布(統計モデル)を特徴づける変数

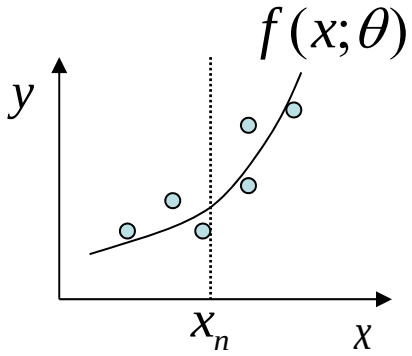
観測変数のみの統計モデルの例

回帰モデル $y = f(x;\theta) + \varepsilon$ ← 付加正規ノイズ

真値

平均0,分散 σ^2

観測データ $D = \{x_n, y_n\}$



統計モデル： 真値を平均値とする1次元正規分布

$$p(y_n \mid x_n;\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2\sigma^2}(y_n - f(x_n;\theta))^2\}$$

最尤推定

$$\sum_{n=1}^{N} L(y_n, f(x_n;\theta)) = -\sum_{n=1}^{N} \log p(y_n \mid x_n;\theta) \equiv \sum_{n=1}^{N} (y_n, f(x_n;\theta))^2 \to \text{Min}_{\theta}$$

Note: 正規ノイズモデルでは、最尤推定は最小自乗推定と等価

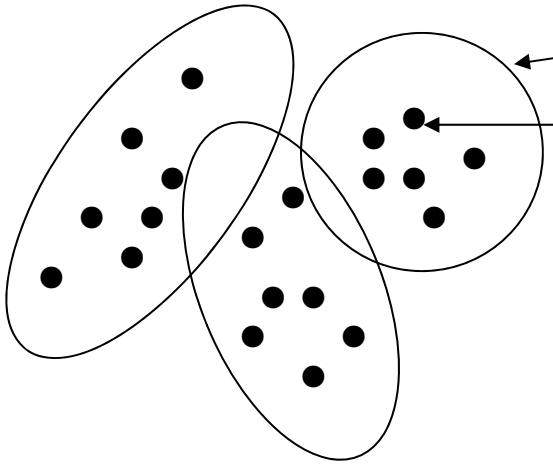
潜在変数を含む統計モデルの例

混合分布モデル

Mixture models

$$p(x;\Theta) = \sum_{i=1}^{m} \alpha_i p_i(x;\theta_i)$$

混合比　　要素分布



x_nはどの要素分布から生成されたかは観測不可能

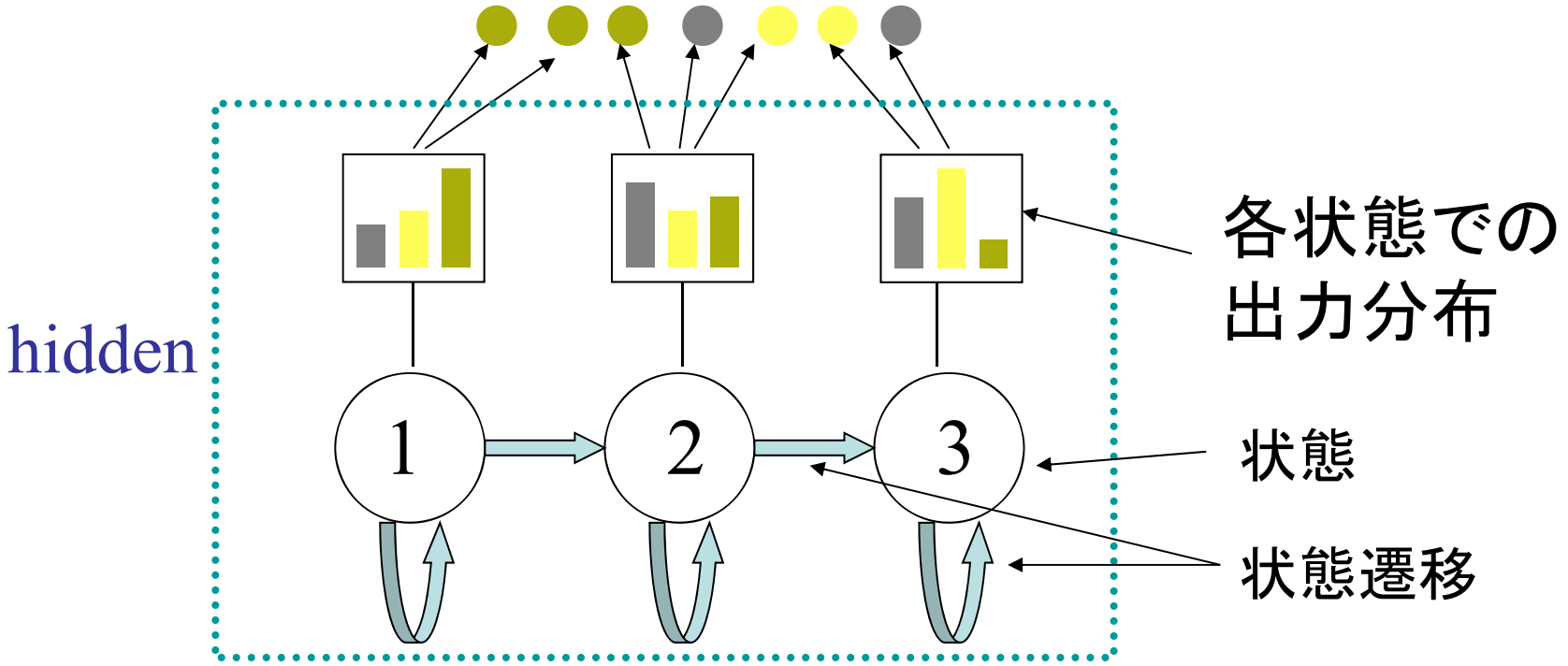
$z_n \in \{1,\cdots,m\}$: **潜在変数**

x_nを生成した要素分布のインデックス

潜在変数を含む統計モデルの例

隠れマルコフモデル (Hidden Markov Models)

観測データ: $x_1, x_2, x_3, \cdots, x_T$



x_t が**どの状態**から生成されたかは観測不可能

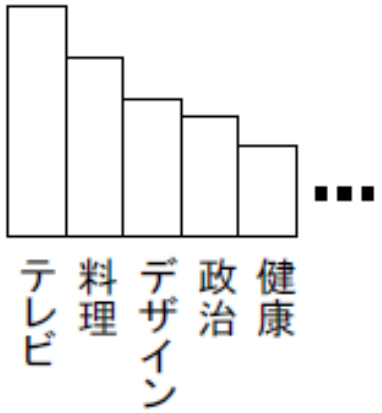
潜在変数：$z_t \in \{1,2,3\}$

潜在変数を含む統計モデルの例

トピックモデル（データマイニングの主要技術）

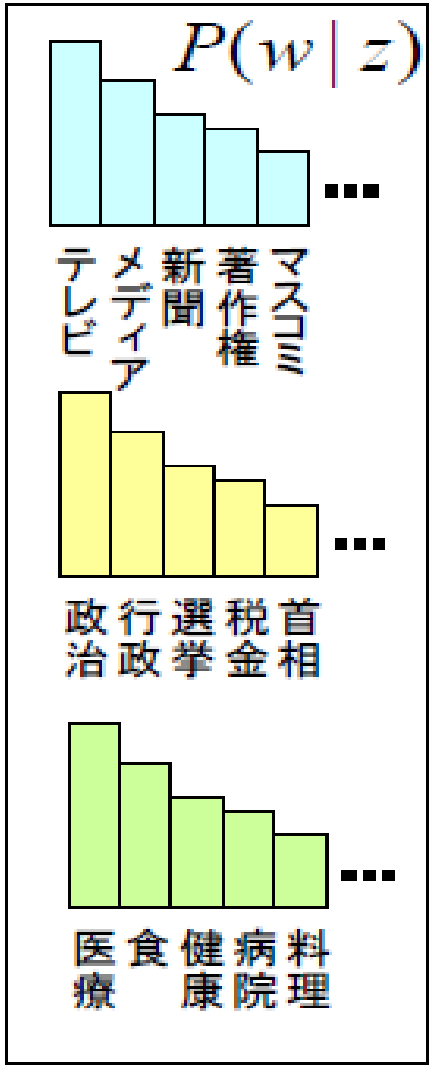
テキスト中の
単語頻度分布

$P(w)$



潜在クラス
の導入

潜在トピック毎
に単語頻度分布
を推定



代表的なトピックモデル

PLSA(=PLSI): Probabilistic Latent Semantic Analysis

$$P(\mathbf{w}) = \prod_{n=1}^{N} \sum_{z} P(z \mid \theta) P(w_n \mid z)$$

LDA: Latent Dirichlet Allocation

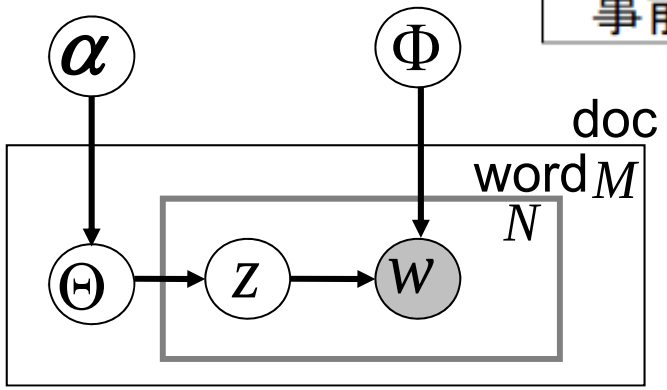
– PLSA+トピック比率にディリクレ事前分布を仮定

$$P(\mathbf{w}) = \int P(\theta) \prod_{n=1}^{N} \sum_{z} P(z \mid \theta) P(w_n \mid z) d\theta$$

ディリクレ事前分布

トピック比率

単語分布



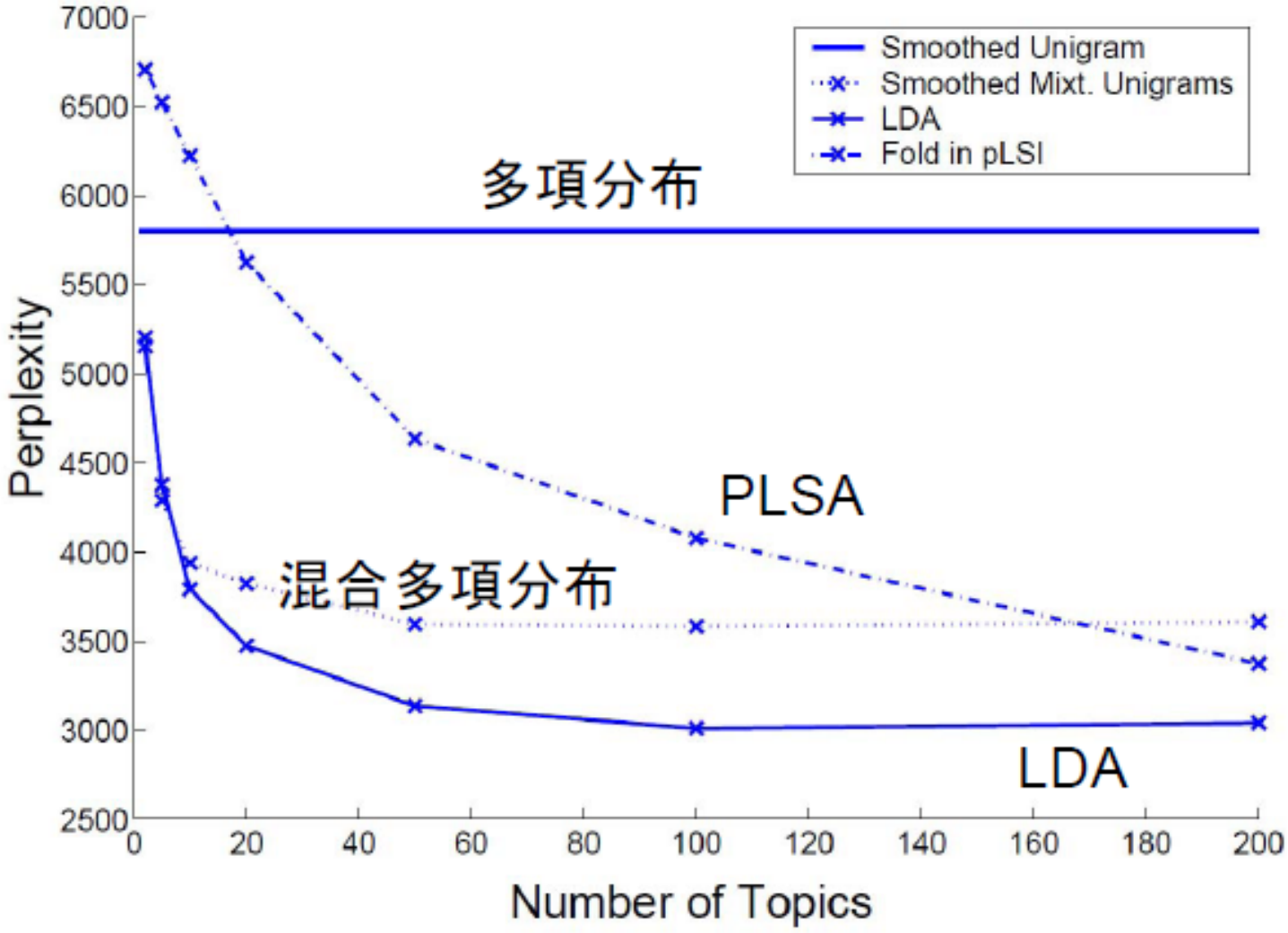
例えば(テキストマイニング応用)
文書集合から潜在トピックの抽出

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003

実データでの比較



D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation, JMLR2003

協調フィルタリングへの応用

テレビ番組　映画　CD / DVD　VIDEO / DVD

協調フィルタリング（推薦システム）

既に分かっている評点を元に，　部分を予測する．

アイテム（ 数千個 ～ ）

ユーザ（数万人～）

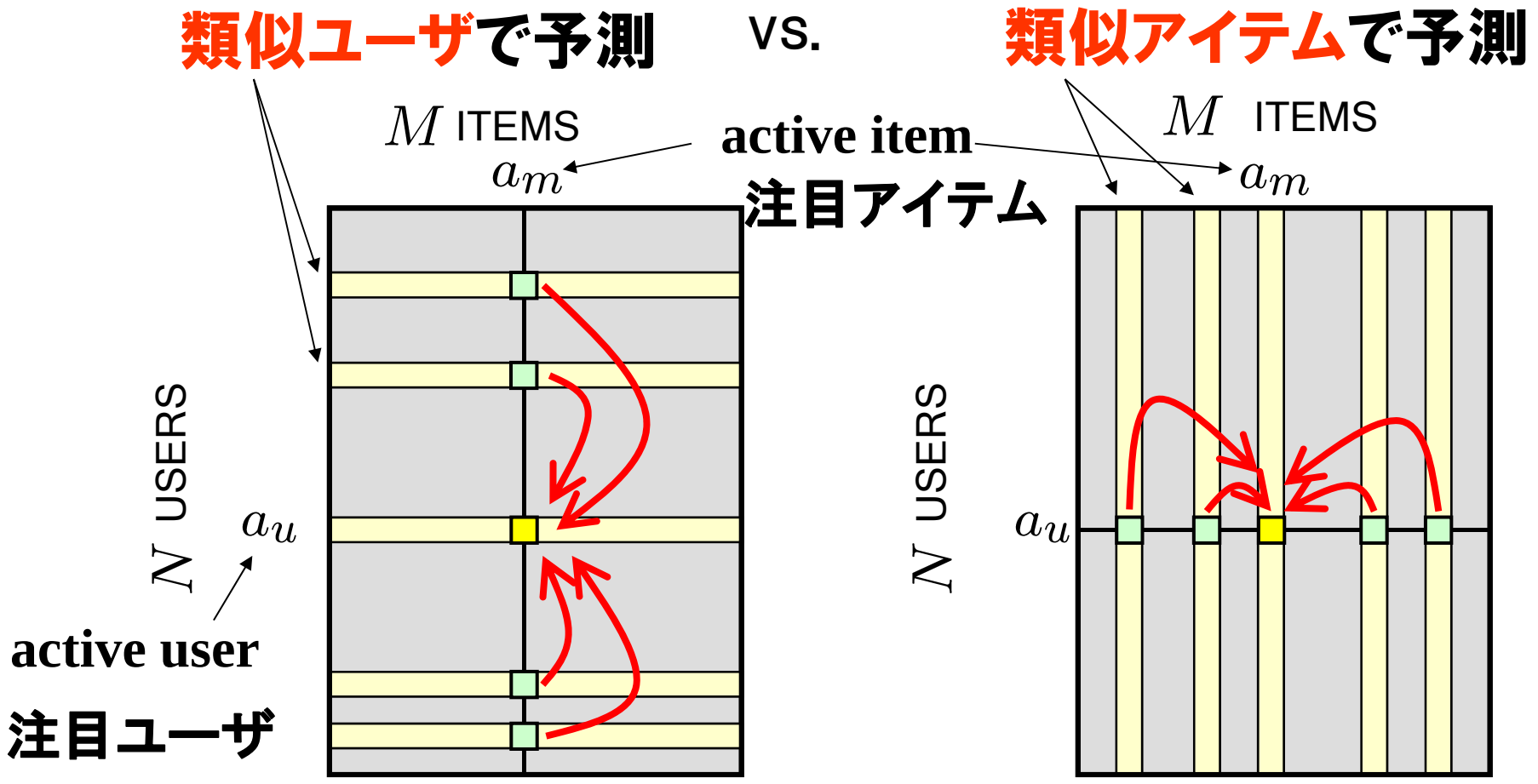
	ランボー	男はつらいよ	釣りバカ日誌	リング
A		4		3
B	3		2	5
C				2
D	4	4	2	
E			3	4
F	4		2	
G	4			2

１：嫌い　←→　５：好き

ただし，表の90％以上は欠損

古典的手法

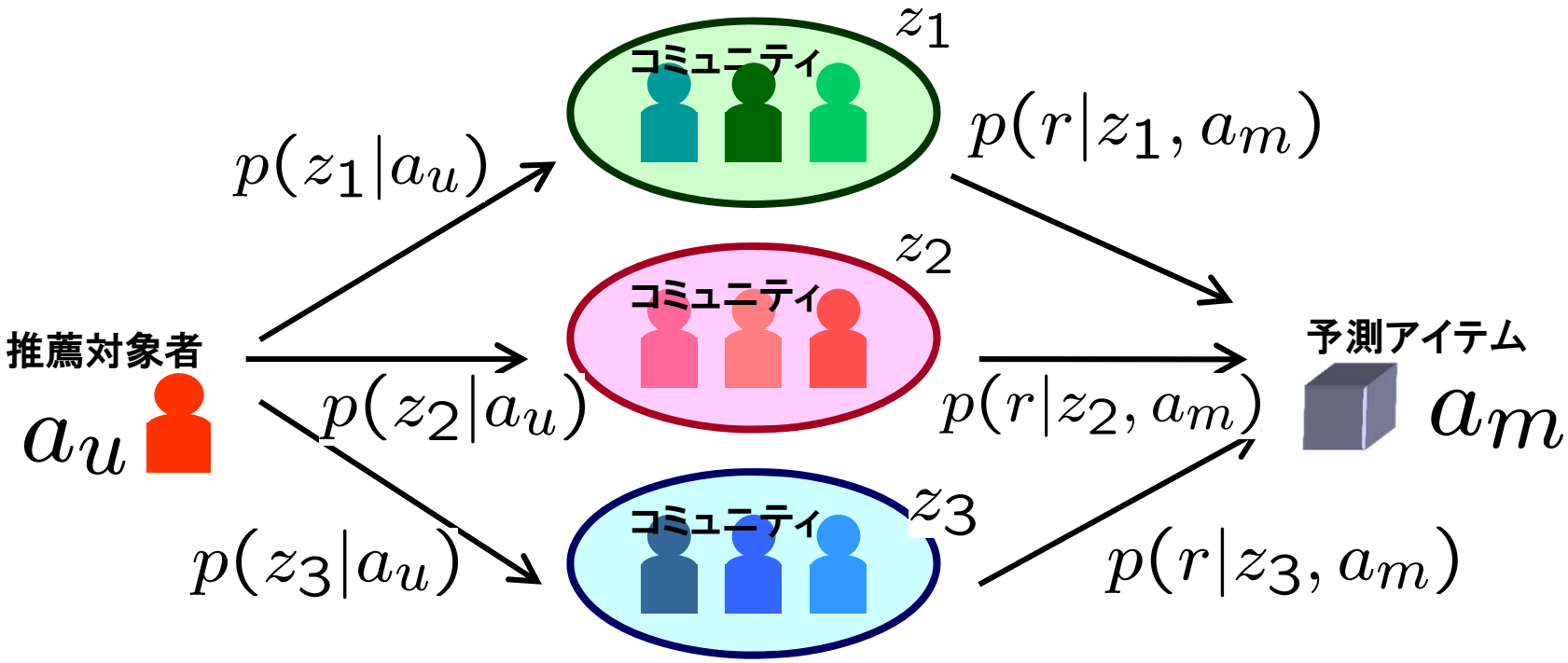
K類似ユーザ，もしくは，K類似アイテムの評点を用いて予測



生成モデル（統計モデル）

probabilistic Latent Semantic Analysis（pLSA）

嗜好が似通った 潜在コミュニティ（潜在クラス）の存在を仮定



$$\text{評点} = \sum_{i=1}^{3} p(z_i|a_u)\, p(r|z_i, a_m)$$

帰属度　　評点分布（正規分布を仮定）

ベイズ学習

ベイズの定理

Thomas Bayes

1702-1761

posterior　　prior

$$P(H \mid E, c) = \frac{P(E \mid H, c)P(H \mid c)}{P(E \mid c)}$$

hypothesis　effect　context

ベイズの定理は以下の積則から導出される

$$P(A, B \mid I) = P(A \mid B, I)P(B \mid I) = P(B \mid A, I)P(A \mid I)$$

条件付き確率：

$$P(A \mid B) = \frac{P(A, B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)}$$

機械学習的には、

観測データ　潜在変数

$$P(\theta \mid D, Z) = \frac{P(D \mid \theta, Z)P(\theta \mid Z)}{P(D \mid Z)}$$

モデル
パラメータ

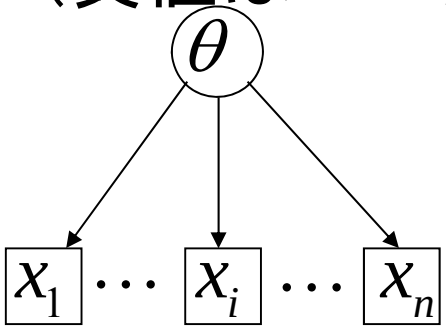
朝日新聞・別紙「be」 3面（2007年11月24日(土)）の記事

300年後に脚光ベイズの定理
迷惑メール対策や人工知能・新薬開発・・・

ベイズモデル

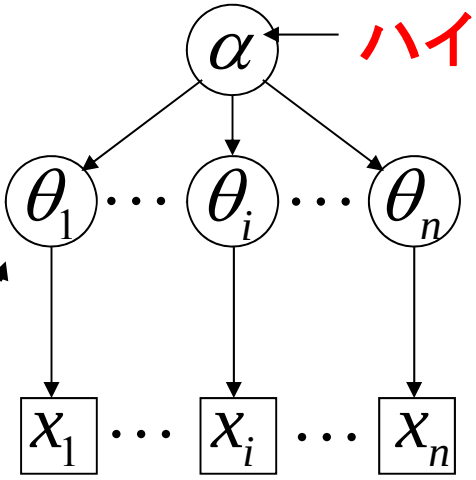
非ベイズ

パラメータは数学的な変数
（真値は一つ）



$p(x;\theta)$

ベイズモデル



ハイパーパラメータ

$p(\theta;\alpha)$

$p(x \mid \theta)$

パラメータは事前分布の一実現値

$\theta_i \in \{\theta_{(1)},\ldots,\theta_{(K)}\}$

K(分割)の事前分布は？

⇩

ノンパラメトリックベイズモデル (Dirichlet Process)

最尤法 vs. ベイズ法

混合モデルの例 $p(x;\theta)=\sum_{k=1}^{K}\pi_k\, p(x;\ \theta_{(k)})\ \ (\pi_k>0, \sum_k \pi_k=1)$

	データ生成（順過程）	学習（逆問題）
最尤法	$z_i \mid \pi \sim \mathrm{Discrete}(z_i;\ \pi)$, $\pi=(\pi_1,\ldots,\pi_K)$ $x_i \mid \theta_{(z_i)} \sim p(x;\ \theta_{(z_i)})$ パラメータは数学的変数	$D=\{x_1,\ldots,x_N\}$ $\Theta=\{\theta_{(1)},\ldots,\theta_{(K)}\}$ $\Theta_{ML}=\arg\max_{\Theta} P(D;\Theta)$ 尤度最大化による点推定 $p(x;\widehat{\Theta}_{ML})$ plug-in estimate
ベイズ法	$\pi=(\pi_1,\ldots,\pi_K)\sim \mathrm{Dirichlet}(\pi;\ \phi_1,\ldots,\phi_K)$ $\theta_{(k)} \sim G_0$ ← 事前分布 $z_i \mid \pi \sim \mathrm{Discrete}(z_i;\ \pi)$ $x_i \mid \theta_{(z_i)} \sim p(x \mid \theta_{(z_i)})$ パラメータも確率変数	ベイズ則による事後分布推定 $p(\Theta \mid D)=\frac{p(D \mid \Theta)p(\Theta)}{p(D)}$ 事後予測分布 $p(x \mid D)=\mathrm{E}_{p(\Theta \mid D)}\{p(x \mid \Theta)\}$ $=\int p(x \mid \Theta)p(\Theta \mid D)d\Theta$

簡単な具体例

２値の独立事象(1/0)の系列 1010011000 を
観測した際、次以降の系列は？

1. 1が生起する確率を θ とすると、n回試行して、1がx回生起する確率は２項分布：

$$P(x\,|\,\theta) = {}_nC_x\,\theta^x(1-\theta)^{n-x}$$ として計算できる.

2. 2項分布のパラメータ θ の共役事前分布としてベータ分布を導入する.

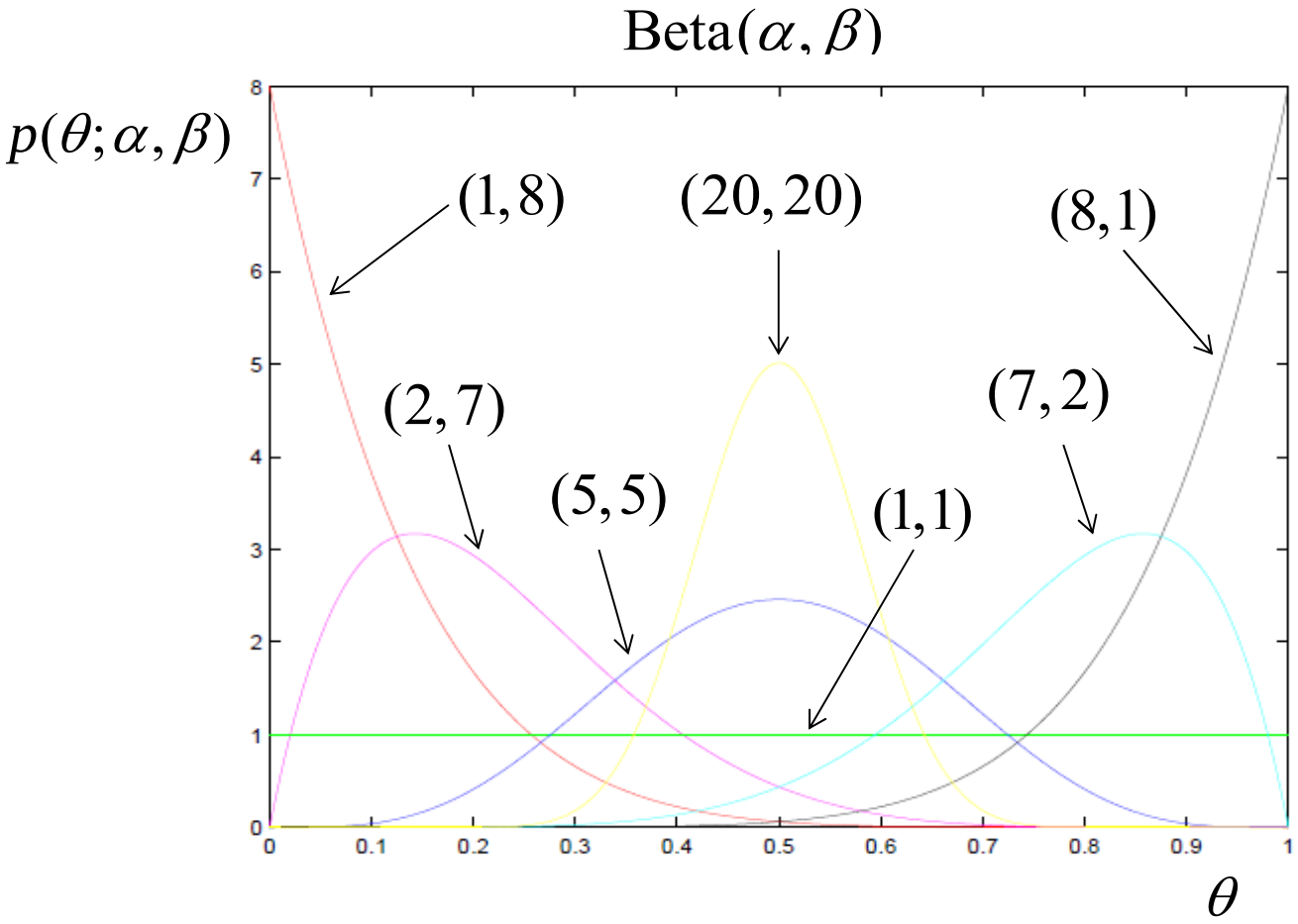
$$p(\theta) = \mathrm{Beta}(\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \longleftarrow \int p(\theta)d\theta = 1$$

とするための正規化項

但し，ベータ関数 $B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$

Note：確率と確率密度関数の表記の違い（大文字・小文字）に注意。

（続き）ベータ分布



ハイパーパラメータ(α, β)を変えることで
様々な分布形が得られる.

Note: [0,1]一様分布はベータ分布の特殊ケース

（続き）

3. $D=\{1010011000\}$ を観測後の θ の事後分布は

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$$

$$= {}_{10}C_4 \theta^4 (1-\theta)^6 \times \theta^{\alpha-1}(1-\theta)^{\beta-1} / B(\alpha,\beta)$$

$$\propto \theta^{4+\alpha-1}(1-\theta)^{6+\beta-1} \equiv \mathrm{Beta}(\alpha+4, \beta+6)$$

Note: 事後分布は正確には，ベイズの定理によりp(D)も考慮して計算すべきであるが、共役事前分布の性質（事後分布もベータ分布となることは既知）を利用して、θ に関連する項だけに注目して変形すれば良い。

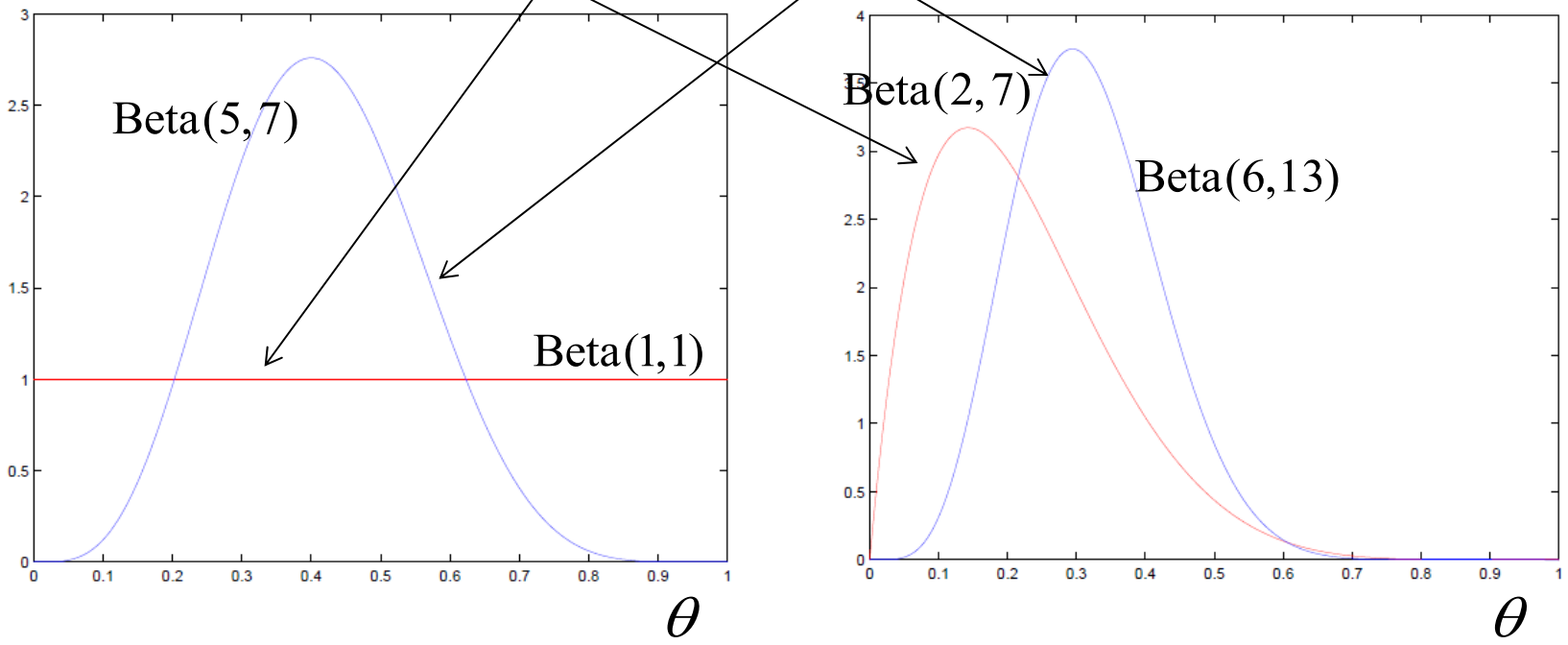
1の生起回数　0の生起回数

$\mathrm{Beta}(\alpha,\beta)$ ➡ Dを観測 ➡ $\mathrm{Beta}(\alpha+4, \beta+6)$

事前分布　　事後分布

（続き）事前分布と事後分布

$\mathrm{Beta}(\alpha, \beta)$　$\mathrm{Beta}(\alpha + 4, \beta + 6)$



事前知識がない場合（一様分布）

事前情報がないため、事後分布のモード（確率密度のピーク）は、1の生起比率(4/10=0.4)に一致している.

事前知識がある場合

事前には0が非常に出やすいとしていたが、実際の観測により、若干修正され（モードがより1.0に移動）ていることが分かる.

（続き）

4. 11回目以降，m回試行して、1が x 回生起する確率は以下の様に計算できる.

$$P(x \mid D) = \int P(x \mid \theta)\, p(\theta \mid D) d\theta$$

$$= \int_0^1 {}_m C_x \theta^x (1-\theta^{m-x}) \times \theta^{\alpha+3}(1-\theta^{\beta+5}) / B(\alpha+4, \beta+6) d\theta$$

$$= \frac{{}_m C_x B(x+\alpha+4, m-x+\beta+6)}{B(\alpha+4, \beta+6)}$$

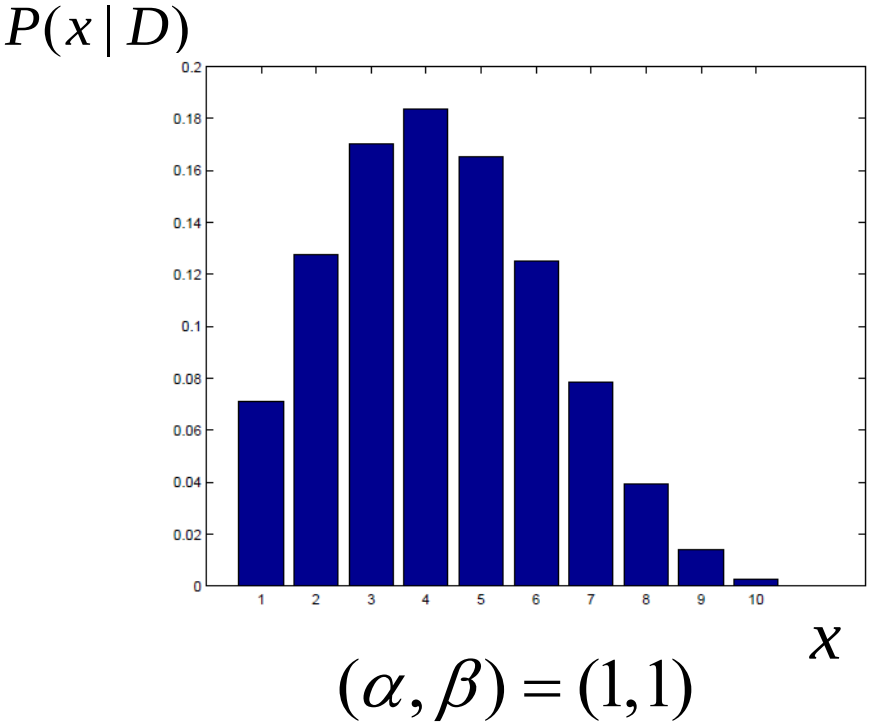
Note：ベータ関数はガンマ関数を用いて以下のように計算でき、
さらに、自然数***n***に対し、$\Gamma(n) = (n-1)!$ が成り立つ。

$$B(k,l) = \frac{\Gamma(k)\Gamma(l)}{\Gamma(k+l)}$$

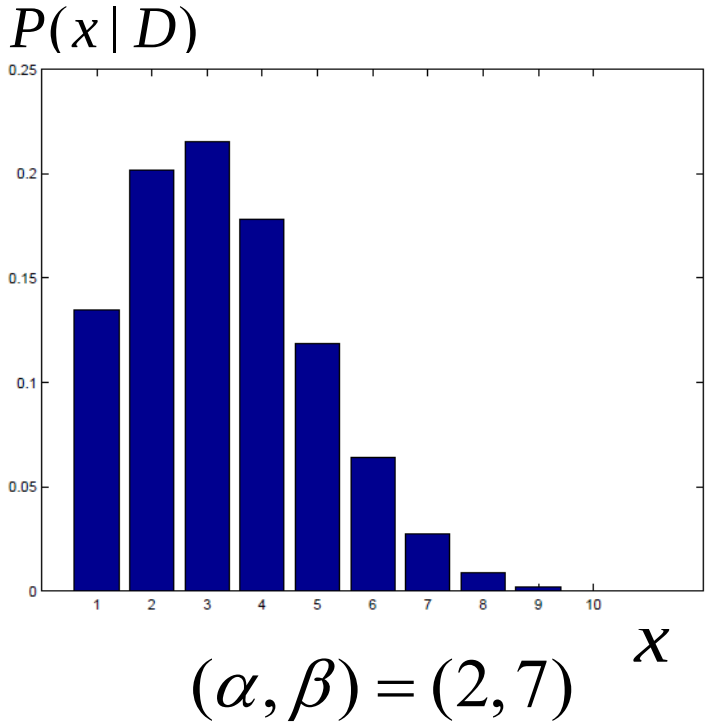
（続き）予測分布の数値例

$$P(x \mid D) = \frac{{}_m C_x B(x+\alpha+4, m-x+\beta+6)}{B(\alpha+4, \beta+6)}$$

m=10の場合



$(\alpha, \beta) = (1,1)$



$(\alpha, \beta) = (2,7)$

Dを観測した後の10回の試行では，左図では事前情報がないため、Dを反映した予測分布（モードはx=4)となっているが，右図では事前情報も考慮して，左図に比べると，より0の回数が多い予測（モードはx=3)となっている．

パラメトリックベイズモデルの課題

（1）パラメトリック事前分布（共役事前分布）の妥当性

共役事前分布は数学的な取り扱いの良さから多用されるが，
実応用でのその妥当性は疑問

（2）モデル選択の問題

例えば，混合モデルの場合，適切な混合数をどう決めるか

モデル選択基準(AIC,BIC etc)は必ずしも適切ではない．また
候補数が多い場合，学習後に基準を評価するのは非効率

（1），（2）の問題を同時に解決するモデリングが
ノンパラメトリックベイズモデル（Dirichlet Process)

ノンパラメトリックベイズ

統計学での起源は古いが，統計学のみならず，統計的学習，
自然言語処理(NLP)の分野で最近流行り出している

1973年　T.S. Fergusonがディリクレ過程(Dirichlet Process)を定義

“A Bayesian analysis of some nonparametric problems.”
Annals of Statistics, vol.1, pp.209-230, 1973

統計的学習の分野では…

・90年台半ばに R. Nealが無限中間層のNNを提案
　この頃には周りがその基礎理論を認識していなかった

・2000年頃にNIPS業界でDPMが流行

　この頃は変分ベイズやMCMC法が浸透していたので
　DPMの流行は自然

近年，機械学習の分野で手法として確立.

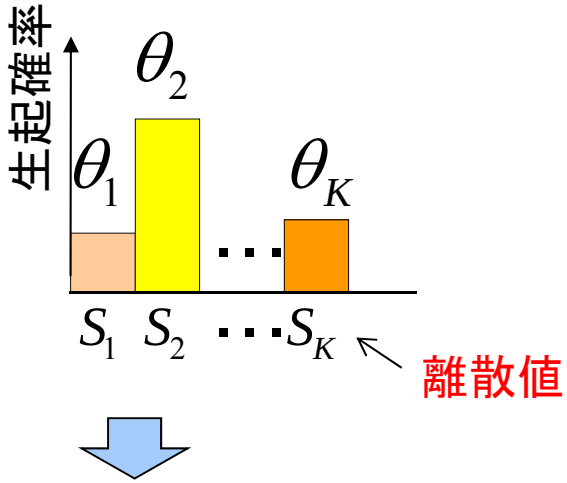
ディリクレ分布 vs. ディリクレ過程

ディリクレ分布：

$\theta \sim \mathrm{Dirichlet}(\alpha)$

α ：ハイパーパラメータ

$\theta = (\theta_1, \ldots, \theta_K)$ 有限次元



有限(K)種類の
離散変数を生成

ディリクレ過程：

$G \sim \mathrm{DP}(\gamma, G_0)$

γ ：ハイパーパラメータ

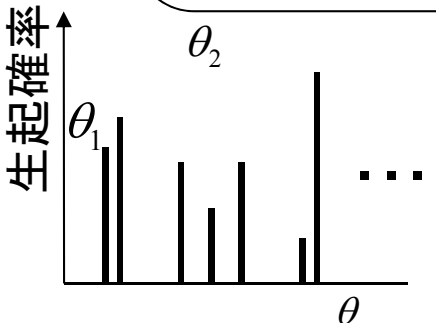
G_0：基底測度(base measure)
基底分布

$G = (\theta_1, \theta_2, \theta_3 \ldots.)$ 無限次元

G_0のドメインに応じてGの要素のドメインが変わる

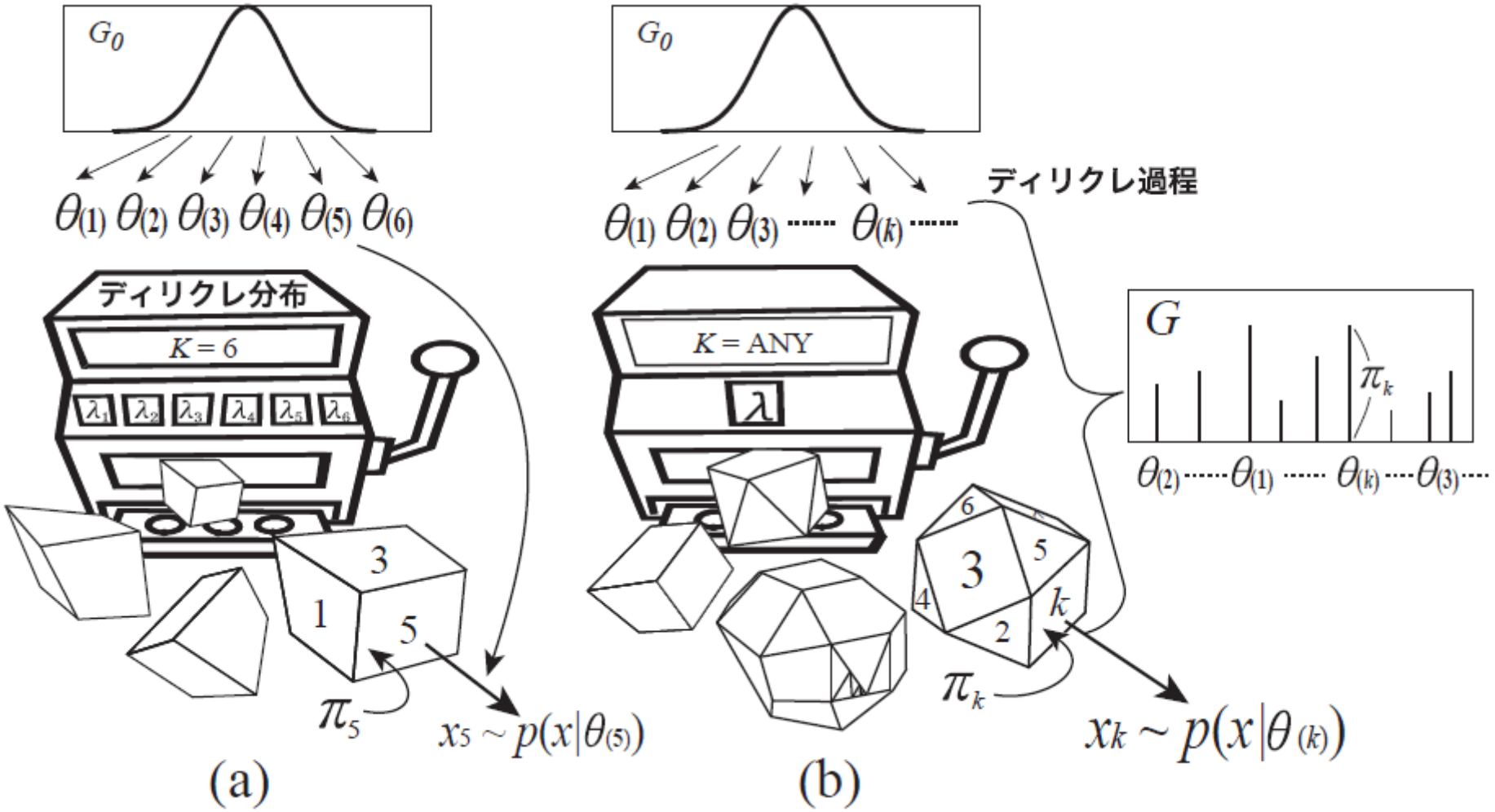
G_0が実数上の連続分布の時、Gの要素は**実数値の離散変数**となる.

G_0が離散分布の時、Gの要素は**離散値の離散変数**となる.



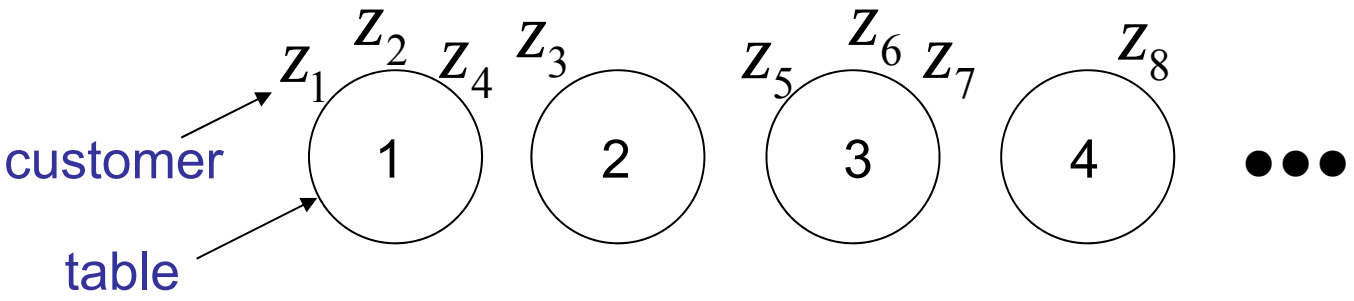
無限種類の
離散変数を生成

有限ディリクレ　vs. 無限ディリクレ



Chinese Restaurant Process (Aldous, 1985)

中国のレストランでは客は混んでいるテーブルを好む？



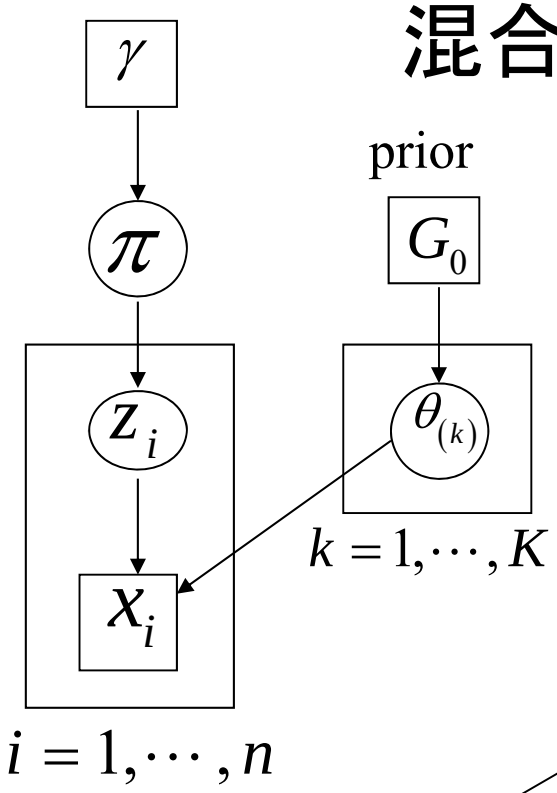
第i客が第kテーブルに座る確率：

$$P(z_i = k \mid z_{1:i-1}, \gamma) = \begin{cases} \dfrac{m_k}{i-1+\gamma} & \text{if } k \text{ is a old table} \\ \dfrac{\gamma}{i-1+\gamma} & \text{if } k \text{ is a new table} \end{cases}$$

m_k ← # of customers at table k

Hoppe's urn schemeと同じ

パラメトリックベイズでの混合分布の生成モデル



混合比

$$p\left(x\middle|\Theta\right) = \sum_{k=1}^{K} \pi_k \, p\left(x\middle|k, \theta_{(k)}\right)$$

$$\pi_k \geq 0, \ \sum_{k=1}^{K} \pi_k = 1$$

要素分布

モデル $p(x \mid \theta_{(k)})$ の共役事前分布

K混合モデルでのデータ生成過程

(1) $\theta_{(k)} \sim G_0$, for $k = 1, \ldots, K$

ディリクレ分布のパラメータ

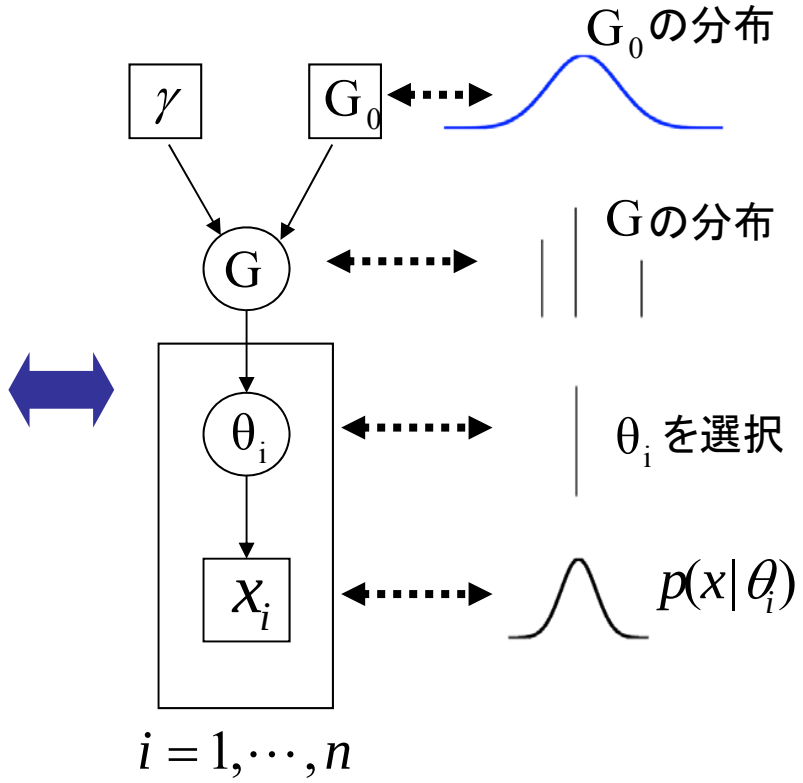
(2) $\pi = (\pi_1, \ldots, \pi_K) \sim \text{Dirichlet}(\pi;\ \gamma_1, \ldots, \gamma_K)$

(3) $z_i \mid \pi \sim$ Discrete$(z;\ \pi)$, and $x_i \mid \theta_{(z_i)} \sim p(x \mid \theta_{(z_i)})$, for $i = 1, \ldots, n$

DPMモデル（一般形）

DPMでのデータ生成過程

$$G: \ \mathrm{DP}(\gamma, \mathrm{G}_0)$$
$$\theta_i | G: \ G, \ \text{for } i = 1,\ldots,n$$
$$x_i: \ p(x|\theta_i), \ \text{for } i = 1,\ldots,n$$



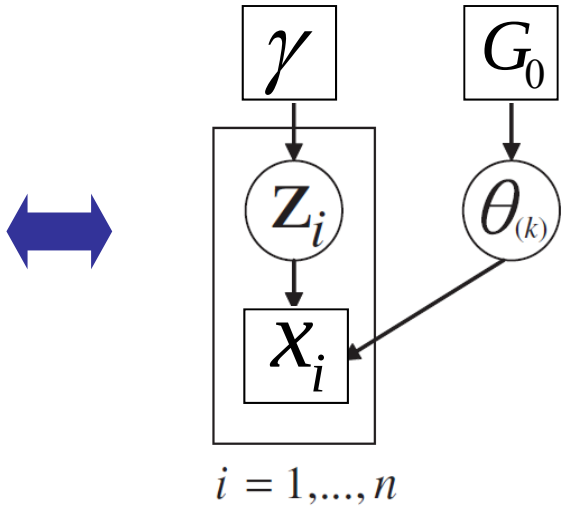
Note: 各 x_i 毎に θ_i が用いられるが，DPのクラスタリング効果により，パラメータ値がクラスタリングされる事に注意

$$P(\theta_i | \theta_{1:i-1}) = \frac{\gamma}{i-1+\gamma} G_0(\theta_i) + \frac{1}{i-1+\gamma} \sum_{k=1}^{K} m_k \delta_{\theta_{(k)}}(\theta_i)$$

DPMモデル（CRP表現）

この表現が最も分かりやすく多用される

$$Z: \ \mathrm{CRP}(\gamma)$$
$$\theta_{(k)} \mid G_0: \ G_0$$
$$x_i: \ p(x \mid \theta_{(z_i)}), \ \text{for } i = 1,\ldots,n$$



$$P(z_i = k \mid z_{1:i-1}) = \begin{cases} \dfrac{m_k}{i-1+\gamma} & \text{既存クラスの選択確率} \quad m_k > 0 \\ \dfrac{\gamma}{i-1+\gamma} & \text{新規クラスの生成確率} \end{cases}$$

Note: $P(z_{1:n} \mid \gamma) = \dfrac{\gamma^K \prod_{k=1}^{K} (m_k - 1)!}{\gamma(\gamma+1)\ldots(\gamma+n-1)}$

DPMモデル（SBP表現）

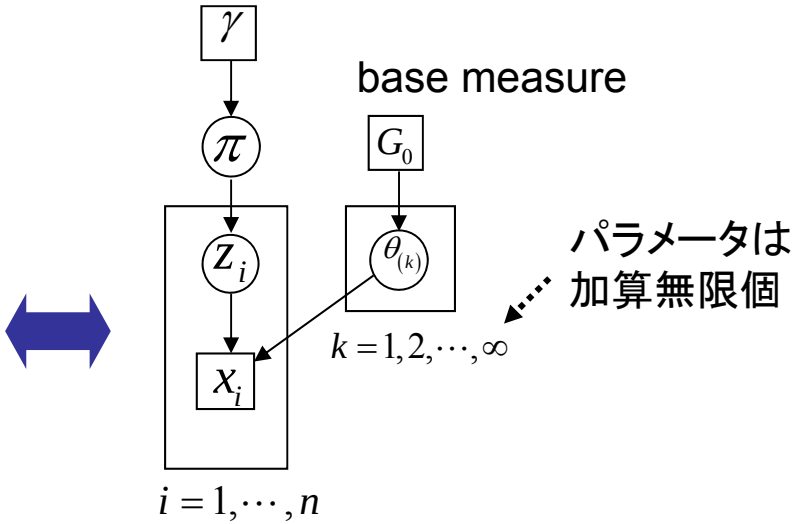
無限混合モデル（SBP-DPM)

$$\pi \sim \mathrm{Stick}(\gamma) \leftarrow \left[v_j \sim \mathrm{Beta}(1,\alpha),\ \pi_k = v_k \prod_{j=1}^{k-1}(1-v_j) \right]$$

$$\theta_{(k)} \mid G_0 \sim G_0$$

$$z_i \mid \pi \sim \mathrm{Discrete}(z_i;\pi)$$

$$x_i : \quad p(x \mid \theta_{(z_i)}),\ \text{for } i = 1,\ldots,n$$



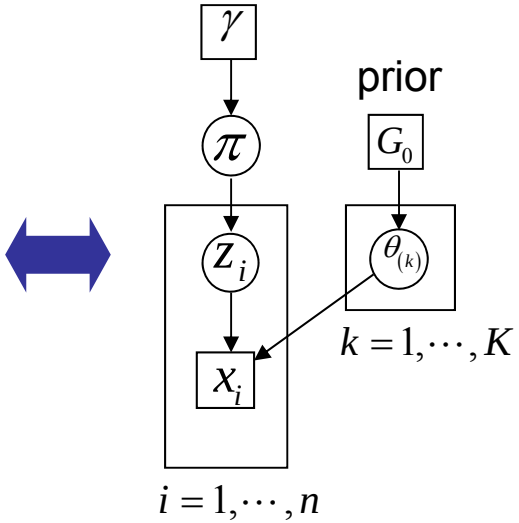
Cf. 有限混合モデル

$$\pi \sim \mathrm{Dirichlet}(\pi;\gamma)$$

$$\theta_{(k)} \mid G_0 \sim G_0$$

$$z_i \mid \pi \sim \mathrm{Discrete}(z_i;\pi)$$

$$x_i : \quad p(x \mid \theta_{(z_i)}),\ \text{for } i = 1,\ldots,n$$



ベイズモデルの学習法

ベイズ法での学習＝事後分布の学習

基本的には，以下の2つのアプローチがある

- MCMC法（Gibbsサンプリング）
  サンプリングに基づく
- 変分ベイズ(Variational Bayes)法
  EM法のベイズ拡張

IRM： Infinite Relational Models

Kemp et al., 2003

関係データマイニングへの応用

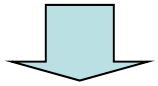
Goal: 教師無し学習による関係(relation)の発見

“domain theory” (E. Davis, 1990)の一実現手法

entity(object)間の潜在関係構造を発見

例） cancer, diabetes → “disorders”

asbests, arsenic → “chemicals”



“chemicals” cause “disorders”

関係

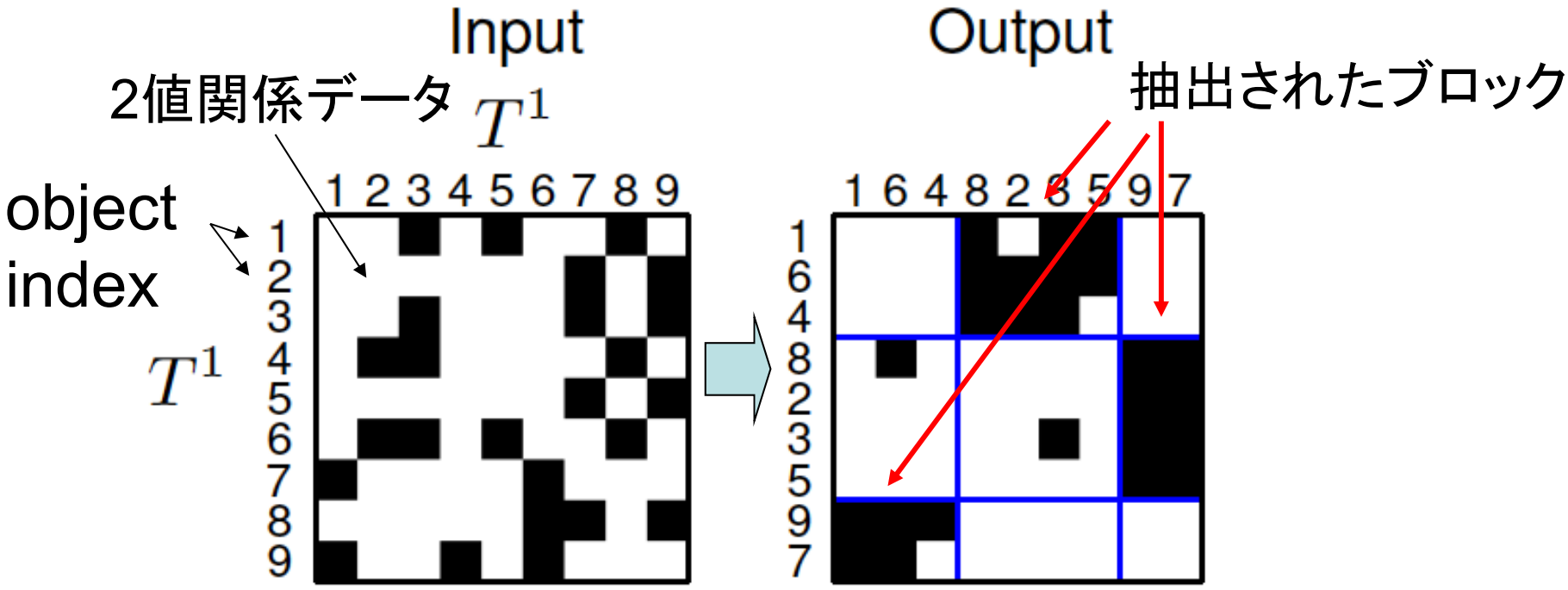
Given: m個の関係　$R^l,\ l=1,...,m$

n個のタイプ　$T^j,\ j=1,...,n$

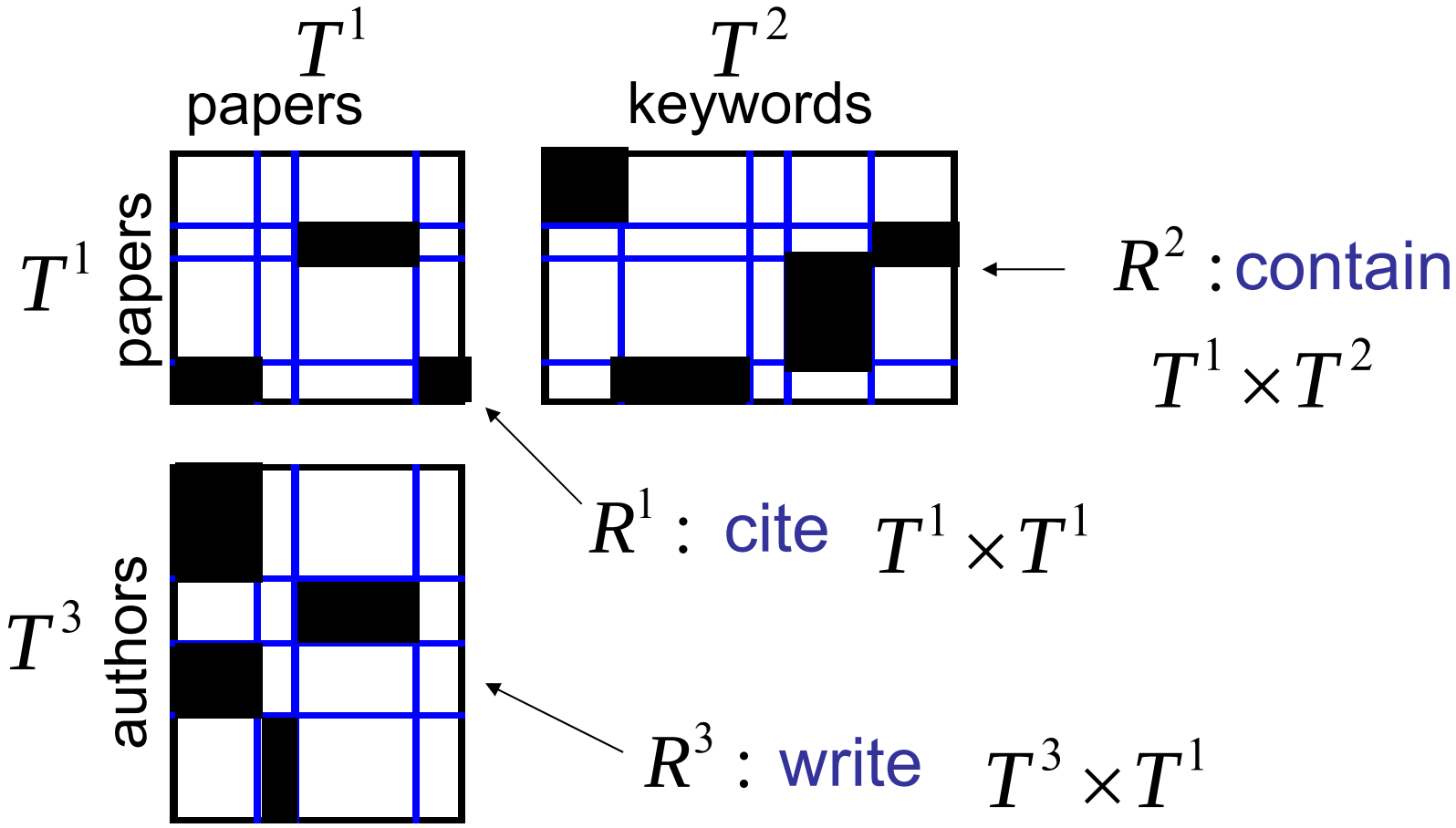
$$R^l : T^i \times T^j \rightarrow \{0,1\}$$

Goal: 潜在クラス（グループ）間の関係の発見

$m=1, n=1$の単純な場合



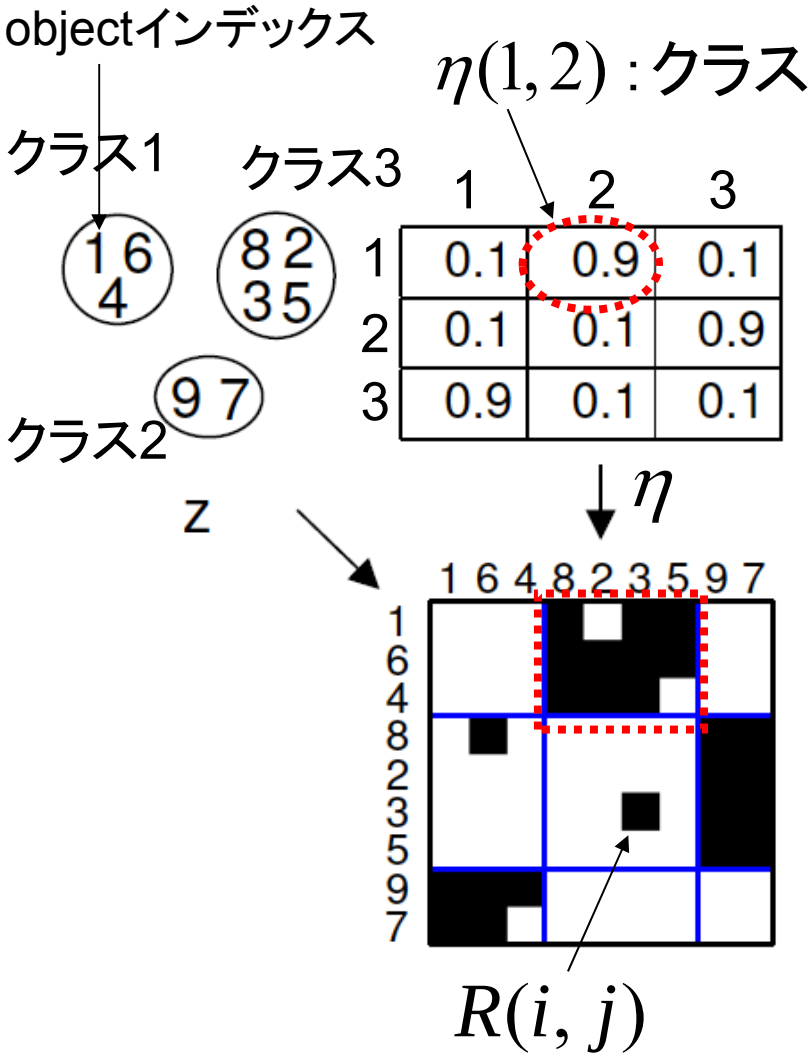
ex) 論文引用データ解析 (*n*=3,*m*=3)



3重の同時クラスタリング

IRMによるデータ生成過程

$R: \quad T \times T \rightarrow \{0,1\}$ の場合



$$
\begin{aligned}
z \mid \gamma &\sim \mathrm{CRP}(\gamma) \\
\eta(a,b) \mid \beta &\sim \mathrm{Beta}(1, \beta) \\
R(i,j) \mid z, \eta &\sim \mathrm{Bernoulli}(\eta(z_i, z_j))
\end{aligned}
$$

object i のクラス

object j のクラス

Note: a, b クラスインデックス

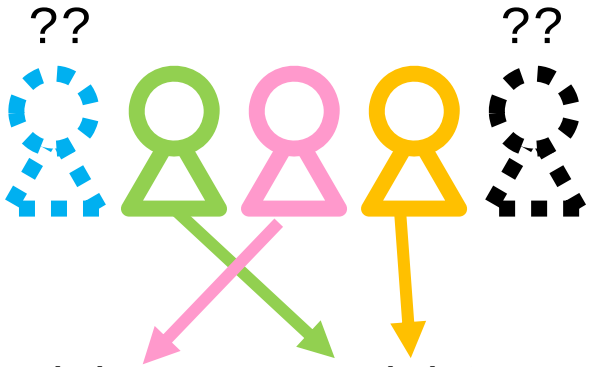
i, j objectインデックス

無限状態カルマンフィルター

Ishiguro et al., 2008

動画像データ

誰が画面からいなくなった？
誰が画面に入ってきた？



対象数の変化を推定

Model 1

$$\mathbf{x}(t) = f(\mathbf{x}(t-1);\xi)$$

$$\mathbf{y}(t) = h(\mathbf{x}(t);\psi)$$

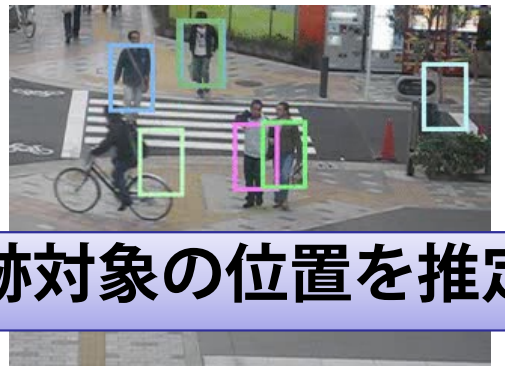
Model 2

$$\mathbf{x}(t) = \hat{f}(\mathbf{x}(t-1);\hat{\xi})$$

$$\mathbf{y}(t) = \hat{h}(\mathbf{x}(t);\hat{\psi})$$

対象の運動モデルを推定

どのモデルが適切？
モデルのパラメータは？

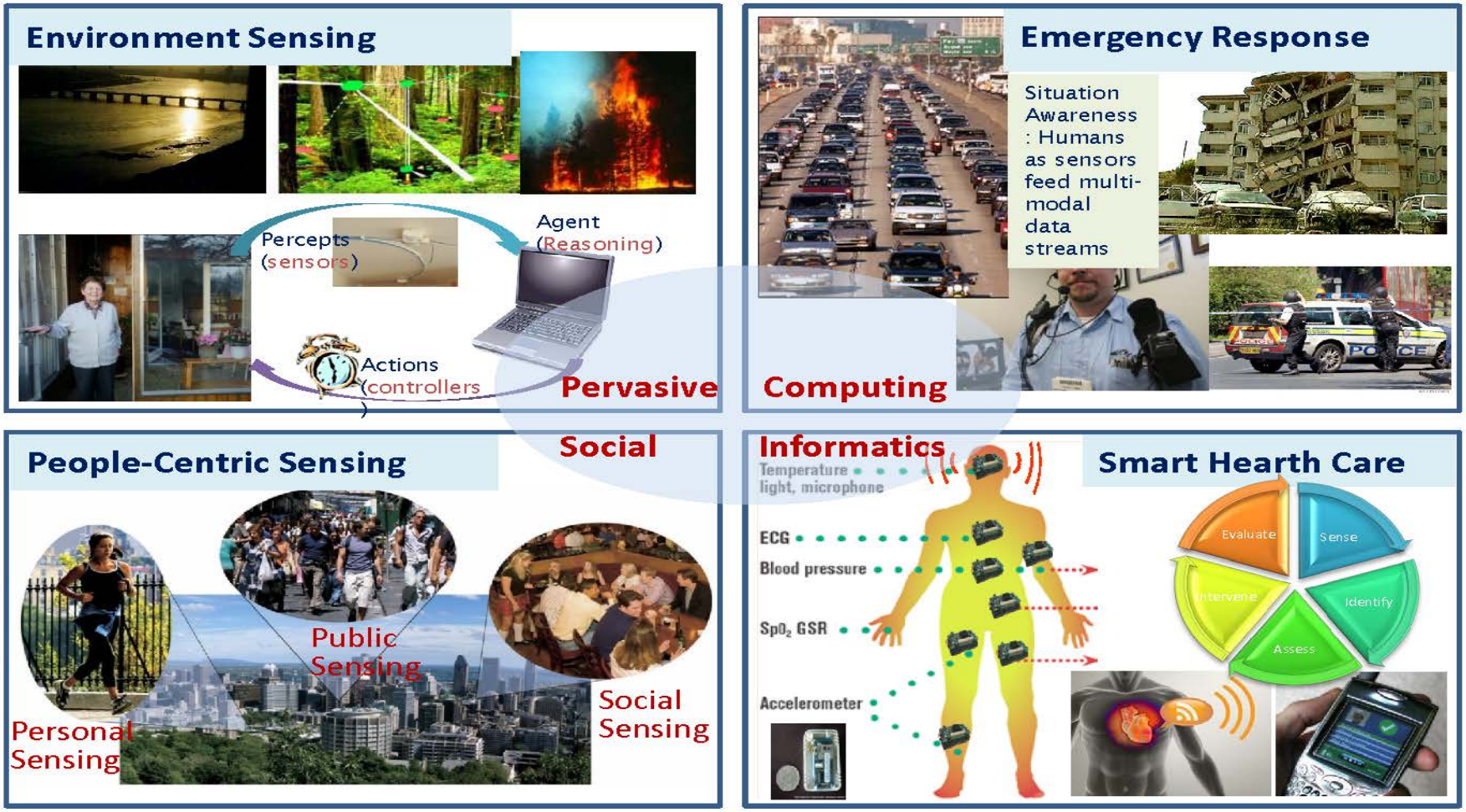


追跡対象の位置を推定

Big Data & CPS時代における機械学習研究

Big Data & Cyber Physical System(CPS)の時代

The Age of Observation: Smart Sensing, Reasoning and Decision



Source: Sajal Das, Keith Marzullo

Credit: Image courtesy of University of Florida

内閣府：最先端研究開発支援プログラム(FIRST)

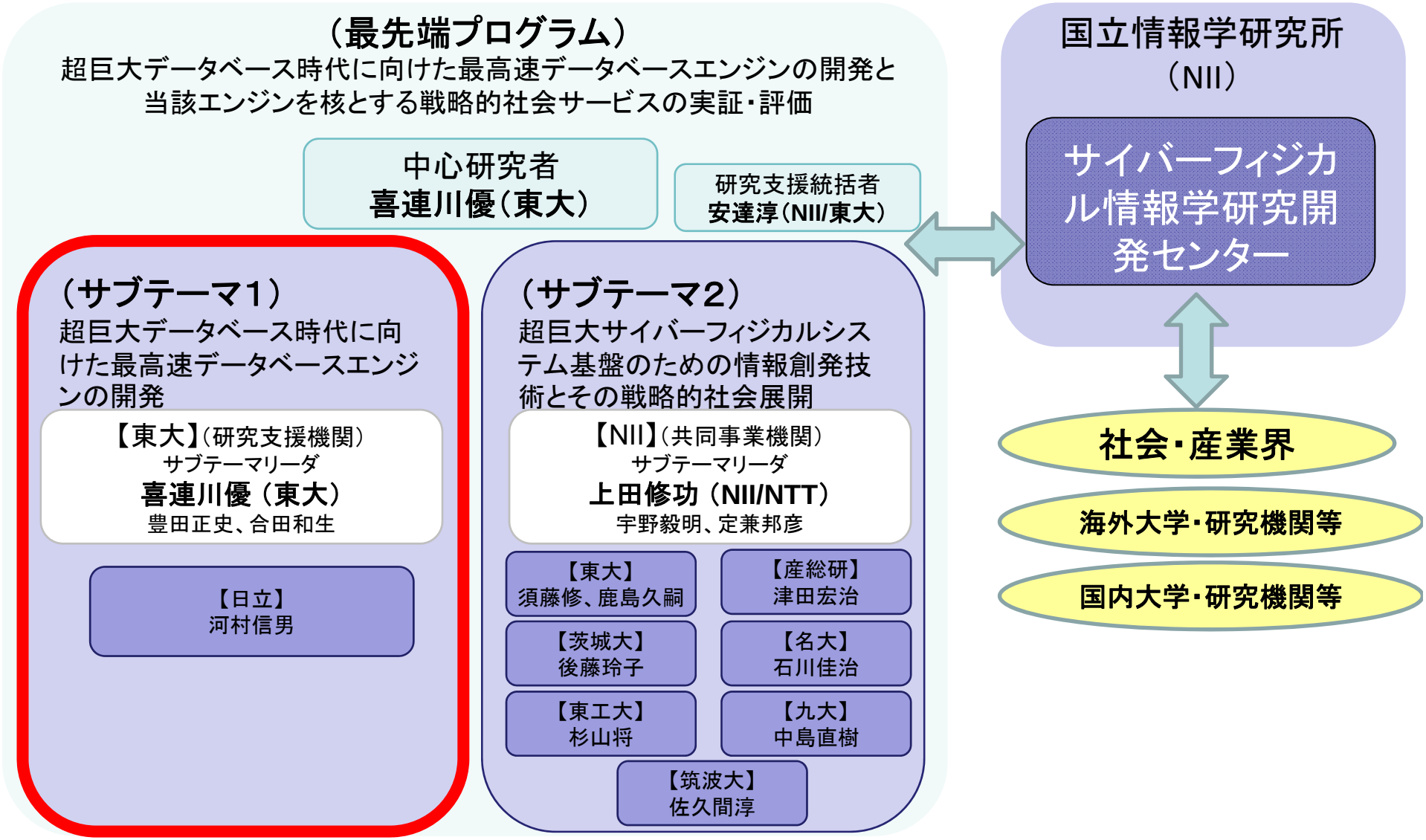
H22年3月～H25年3月（代表研究者：喜連川 優（東大））

サブテーマ1： **超巨大データベース時代に向けた最高速データベースエンジンの開発**
（サブテーマリーダ：喜連川優）

サブテーマ2： **超巨大サイバーフィジカルシステム基盤のための情報創発技術とその戦略的社会展開（サブテーマリーダ：上田修功（NTT，NII））**

サイバー情報（web）とセンサー情報（実世界）との融合による新価値の創造

実施体制


（最先端プログラム）
超巨大データベース時代に向けた最高速データベースエンジンの開発と
当該エンジンを核とする戦略的社会サービスの実証・評価
中心研究者
喜連川優（東大）
研究支援統括者
安達淳（NII/東大）
（サブテーマ1）
超巨大データベース時代に向けた最高速データベースエンジンの開発
【東大】（研究支援機関）
サブテーマリーダ
喜連川優（東大）
豊田正史、合田和生
【日立】
河村信男
（サブテーマ2）
超巨大サイバーフィジカルシステム基盤のための情報創発技術とその戦略的社会展開
【NII】（共同事業機関）
サブテーマリーダ
上田修功（NII/NTT）
宇野毅明、定兼邦彦
【東大】
須藤修、鹿島久嗣
【産総研】
津田宏治
【茨城大】
後藤玲子
【名大】
石川佳治
【東工大】
杉山将
【九大】
中島直樹
【筑波大】
佐久間淳
国立情報学研究所
（NII）
サイバーフィジカル情報学研究開発センター
社会・産業界
海外大学・研究機関等
国内大学・研究機関等

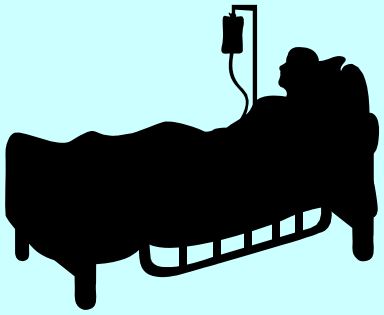
例えば：保健医療CPSの実現

（生活習慣病の合併症管理や高齢者見守り）

Aging independently in place:
Move ¼ of institutional care to the home in 10 years

病院でのコストの高い治療
「非日常空間」における治療

外来



入院

在宅治療の安全化・精緻化・低コスト化
生活習慣病の「日常空間」における治療

データセンタでパスや診療情報と融合しながら
遠隔で多くの患者（高齢者）を管理

操作のいらない
センサネットワーク

生活上ストレスに
ならないセンサ類
（バーチャルナース）

酸素濃度計

脈拍計

行動計

血糖計

血圧計

体重計

情報薬配信
（日常・緊急指示）
- コール
- メール
- 携帯Web

スマート
フォン

テレビ（地デジ）など
インターネット端末

高齢者でも使いやすい
インターネット機器

まとめ

統計的機械学習技術の特長

1. 学習データさえあれば、大半の問題に適用可能な汎用技術

2. 未知データの確率的予測が可能

3. 欠損やノイズを含むデータにも対処可能