データサイエンスが誘う メゾスコピックHPC

樋口知之（中央大・理工学部）

新エネルギー・産業技術総合開発機構（NEDO）
ロボット・AI部　PM

産業技術研究所　人工知能研究センター　顧問

（株）ブレインパッド　顧問
（株）ブリヂストン　　特別技術顧問

一般社団法人
データサイエンティスト協会 理事

人間の近未来予測は当たらない

1982年→2019年 (37年先)

平成の３０年間

（図のソース： 2017/12/29日経Webニュース）

自己紹介（主にシミュレーションがらみ）

1989年3月
東京大学理学系研究科地球物理学専攻 博士課程修了
同時に 理学博士号取得

宇宙プラズマの乱流現象の大量の人工衛星観測データ（GOES, DMSP Voyager, ISEE）の解析に従事
- 東大大型計算機センターとの日々往復
- （株）ブラザーでワープロ開発のバイト

1989年4月
文部省統計数理研究所 予測制御研究系 予測理論研究部門助手に採用

1992年～1997年 通産省の大型プロジェクトReal World Computing（NEC理論チーム）に参加

1998年～2001年 特定領域研究「発見科学」（代表　有川先生、宮野先生の班）に参加

2002年7月
予測制御研究系 システム解析研究部門教授

2002年～2007年 文科省「人・自然・地球　共生プロジェクト」（代表　淡路先生の課題）に参加

2004年～2010年 CREST代表
研究領域「シミュレーション技術の革新と実用化基盤の構築」（総括　土井先生）
研究課題「先端的データ同化手法と適応型シミュレーション研究」

2004年4月～2010年3月
副所長（研究企画担当）

2006年～2014年 理研「次世代生命体統合シミュレーションソフトウェアの研究開発」プロジェクトに参加

2011年4月～2019年3月
統計数理研究所　所長

2019年4月～
中央大学理工学部

出版（2011）

1980年代数理・情報技術の再興

- ニューラルネットワーク→深層学習
- 情報統計力学
  - ✓ イジングモデル→量子アニーリング
  - ✓ シミュレイティドアニーリング
- 遺伝的アルゴリズム、遺伝的プログラミング→超並列計算による大規模離散最適化
- 連続最適化（降下法）→確率的勾配降下法、オンライン学習

最適化

- 制約付き最適化→スパースモデリング
- 次元圧縮→行列、テンソル分解

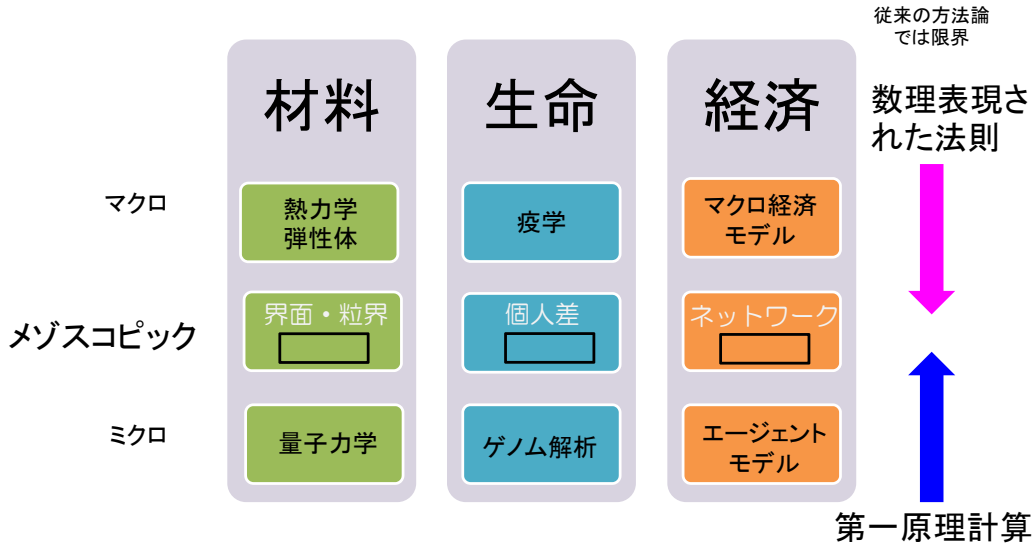
アウトライン

1. 演繹と帰納
2. データ同化
   a. 再訪
   b. ハミルトニアン・モンテカルロ法
3. 機械学習とシミュレーションの融合
4. 人材育成での課題

アウトライン

1. 演繹と帰納
2. データ同化
   a. 再訪
   b. ハミルトニアン・モンテカルロ法
3. 機械学習とシミュレーションの融合
4. 人材育成での課題

メゾスコピック・モデリング



異質・相反する思考法

演繹 ⟷ 帰納

機能のモデル化と真理の探求のバランス感覚

朝日新聞平成14年2月17日(日) 14版第2社会面

JR東日本が開発

実用化なら運休時間2、3割減

システムは、3分前時で記録した風速データを統計学の手法で解析した数式とあてはめ、その後の風速をコンピュータに計算させる。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以上が実際に吹いたら30分間列車を止め、この間に風速30以上にならない、予測できるようになるという。

実用化につなげた場合、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、関係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を検討していくという。

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

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

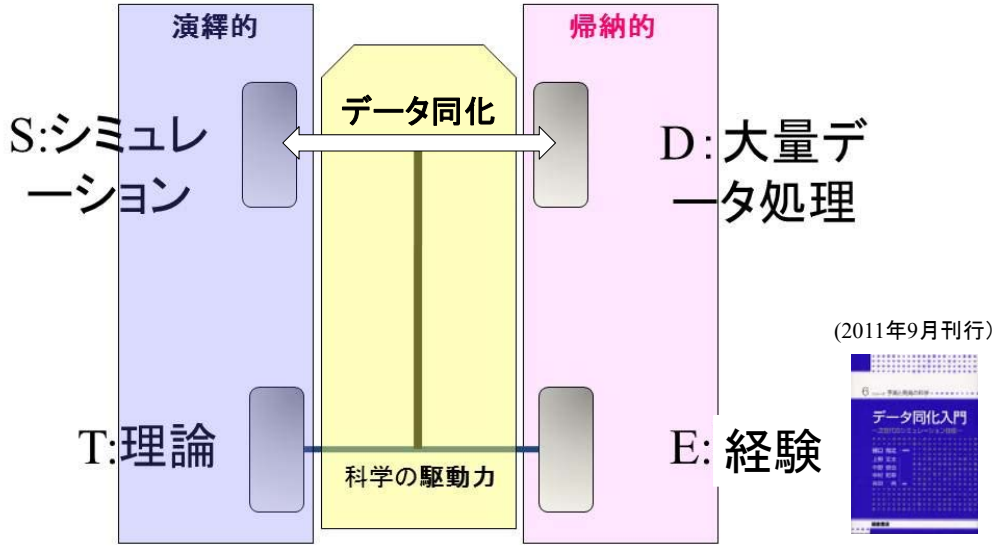
立ちはだかるは、物理帝国！壁は高い

- 帰納的推論と機能のモデル化
  - ○ 物理帝国主義観からの脱却
  - ○ ニュートンパラダイムからのシフト

アウトライン

1. 演繹と帰納
2. データ同化
   a. 再訪
   b. ハミルトニアン・モンテカルロ法
3. 機械学習とシミュレーションの融合
4. 人材育成での課題

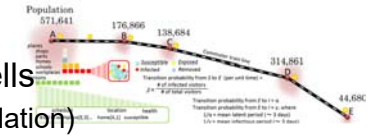
つなぐ：データ同化

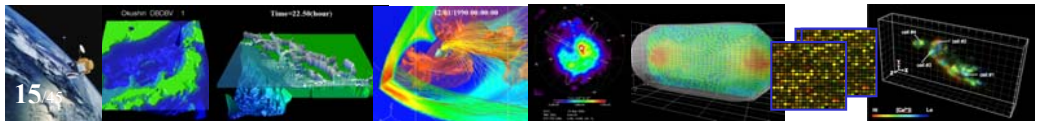


Application studies carried out by our group for **model** improvements

Research Projects:

- Typhoon trajectory
- Tsunami, Ocean tide
- Auroral phenomena
- 3D structure of ring current
- Acoustic waves
- Intercellular fluid dynamics
- Genome informatics
- Drug response prediction
- Neuronal circuits of whole nerve cells
- Influenza Pandemic (Multi agent simulation)





数理的課題としてデータ同化の枠組み

- 動的潜在変数 → 状態ベクトル $\boldsymbol{x}_t$
  - ✓ 直接は観測できない
  - ✓ 計測できてもノイズが存在する
- 時間発展方程式はシミュレーショモデルで規定 $\boldsymbol{x}_t = f(\boldsymbol{x}_{t-1}|\boldsymbol{\theta})$
  - ✓ 通常は時間発展方程式をとりあつかうが、構造シミュレーション（有限要素法、境界要素法）も取り扱える
- 観測（計測）データ $\boldsymbol{y}_{1:T} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T]$

Challenging problem: Huge dimension and inversion

- Data Assimilation = Estimation problem of state vector $\boldsymbol{x}_t$:

Simulation model

(system model) $\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}, \boldsymbol{v}_t | \boldsymbol{x}_0)$

(observation model) $\boldsymbol{y}_t = H_t \boldsymbol{x}_t + \boldsymbol{w}_t$

or

$\boldsymbol{y}_t = h_t(\boldsymbol{x}_t, \boldsymbol{w}_t)$

$\boldsymbol{y}_t \sim p(\boldsymbol{y}_t \mid \boldsymbol{x}_t, \boldsymbol{\theta})$

- $\boldsymbol{x}_t$: All variables in simulation model
- $\boldsymbol{y}_t$: All observed variables
- $\boldsymbol{v}_t$: Stochastic part to represent uncertainty of model (boundary condition, …)
- $\boldsymbol{w}_t$: Observation error
- $\boldsymbol{v}_t, \boldsymbol{w}_t$: Normally Gaussian $\boldsymbol{x}_0$: Initial condition

dimension $\boldsymbol{x}_t$:10^4~10^7 $\boldsymbol{y}_t$:10^2~10^5 $\underbrace{\dim(\boldsymbol{x}_t)}_{m} >> \underbrace{\dim(\boldsymbol{y}_t)}_{n}$

データ同化で推定したい変数

- 時間に依存しないパラメータ

$\boldsymbol{\theta}$: モデルパラメータ（次元は < 20程度）

$\boldsymbol{x}_0$: 初期状態ベクトル（初期条件）

$\dim(\boldsymbol{x}_0) = 10^6 \sim 10^8$ Adjoint法

- 状態ベクトルの列

$$\boldsymbol{x}_{1:T} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]$$

事後分布 $p(\bullet \mid \boldsymbol{y}_{1:T})$

ベイズの反転公式

$\boldsymbol{x}$: 興味のある対象

$\boldsymbol{y}$: データ

順解析 : シミュレーション等

ベイズの反転公式

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

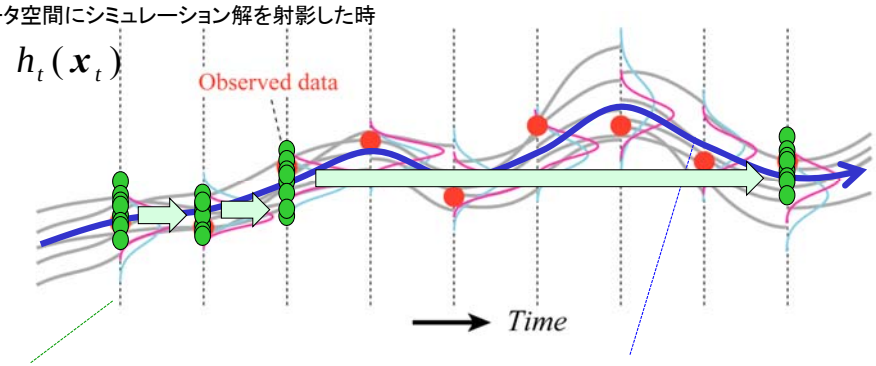
逆解析

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

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）

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	MPF　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。	・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

・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。

気象庁 → 全世界的にこちらにシフト
東大・大気海洋研
理研

アンサンブル数は数十から百のオーダー

非線形度が高い
小規模問題

逐次データ同化

Kalman Filter (KF) は1960年代から

- PF (Particle Filter) 1996 年
- EnKF (Ensemble KF) 1994
  - EnTKF (En transform KF) 2001
  - LETKF (Local EnTKF) 2007
- MPF (Merging PF) 2007
- pMCMC (particle MCMC) 2010

アウトライン

1. 演繹と帰納
2. データ同化
   a. 再訪
   b. ハミルトニアン・モンテカルロ法
3. 機械学習とシミュレーションの融合
4. 人材育成での課題

システムノイズがないとき

Prediction

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

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t-1}) = \int \boxed{\delta(\boldsymbol{x}_t - f(\boldsymbol{x}_{t-1}))}p(\boldsymbol{x}_{t-1}|\boldsymbol{y}_{1:t-1})d\boldsymbol{x}_{t-1}$$

システムノイズがない場合

普通にシミュレーションを実行

$$\{\boldsymbol{x}_t\}_{t=1}^{T}$$

システムノイズがないとき
時間に依存しないパラメータ

- 尤度関数が簡単に計算

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

$$p(\boldsymbol{y}_t|\boldsymbol{\theta}) \to \frac{1}{C}\exp\left\{-\frac{(\boldsymbol{y}_t - H\boldsymbol{x}_t)^2}{2\sigma_t^2}\right\}$$

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

- 事後分布をMCMCで計算

最近の動向：ハイブリッドMCの研究が盛ん

$$p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T}) = \frac{p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y}_{1:T})} \propto p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

プロポーザル分布

M-H（Metropolis-Hastings）法 $r = \dfrac{p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\boldsymbol{y}_{1:T}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}$

HMC：ハミルトニアン・モンテカルロ法

$$p(\boldsymbol{\theta},\boldsymbol{p}|\boldsymbol{y}_{1:T}) = p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T})p(\boldsymbol{p}) \propto p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T})\exp(-\frac{\boldsymbol{p}^2}{2\sigma^2})$$

- 仮想的時間と運動量（補助変数）を導入：$\boldsymbol{p}_\tau$
- 事後分布を位置エネルギーに変換：$p(\boldsymbol{\theta}|\boldsymbol{y}_{1:T}) = \frac{1}{C}\exp\{-\beta H(\boldsymbol{\theta})\}$
- エネルギーを定義：$\mathcal{H}(\boldsymbol{p},\boldsymbol{\theta}) = \frac{\boldsymbol{p}^2}{2} + \beta H(\boldsymbol{\theta})$

- 運動方程式を数値的に解く：

$$\begin{cases} \frac{d\boldsymbol{\theta}}{d\tau} = \boldsymbol{p} \\ \frac{d\boldsymbol{p}}{d\tau} = -\beta\frac{\partial H(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \end{cases}$$

$$\boldsymbol{p}^{[i-1]} \sim N(0,\sigma^2)$$

- 棄却率を計算し、MHを実行
（数値誤差による不具合を回避するため）

$$\Delta\mathcal{H} = \mathcal{H}(\boldsymbol{\theta}^*,\boldsymbol{p}^*) - \mathcal{H}(\boldsymbol{\theta}^{[i-1)]},\boldsymbol{p}^{[i-1]})$$

$$r = \min\{1,\exp(-\beta\Delta\mathcal{H})\}$$

NUTS(No U-turn sampler)

アウトライン

1. 演繹と帰納
2. データ同化
   a. 再訪
   b. ハミルトニアン・モンテカルロ法
3. 機械学習とシミュレーションの融合
4. 人材育成での課題

生成モデルと識別モデル（関数）
Generative Model vs. Discriminative Model

生成モデル（広義）

$G(z)$: $p(\boldsymbol{y}|\boldsymbol{\theta},z)$

データ　パラメータ　説明変数（アスペクト）

クラス分類問題　例：C=1 or 0

生成モデルからベイズの定理を経由して条件付き確率を計算　$p(C=C_i|\boldsymbol{y},\boldsymbol{\theta}) = \dfrac{p(\boldsymbol{y},C=C_i|\boldsymbol{\theta})}{\sum_i p(\boldsymbol{y},C=C_i|\boldsymbol{\theta})}$

生成モデル（狭義）

$p(\boldsymbol{y},C|\boldsymbol{\theta})$　同時分布

データ　クラス

$p(\boldsymbol{y}|C,\boldsymbol{\theta}_2)\cdot p(C|\boldsymbol{\theta}_1)$

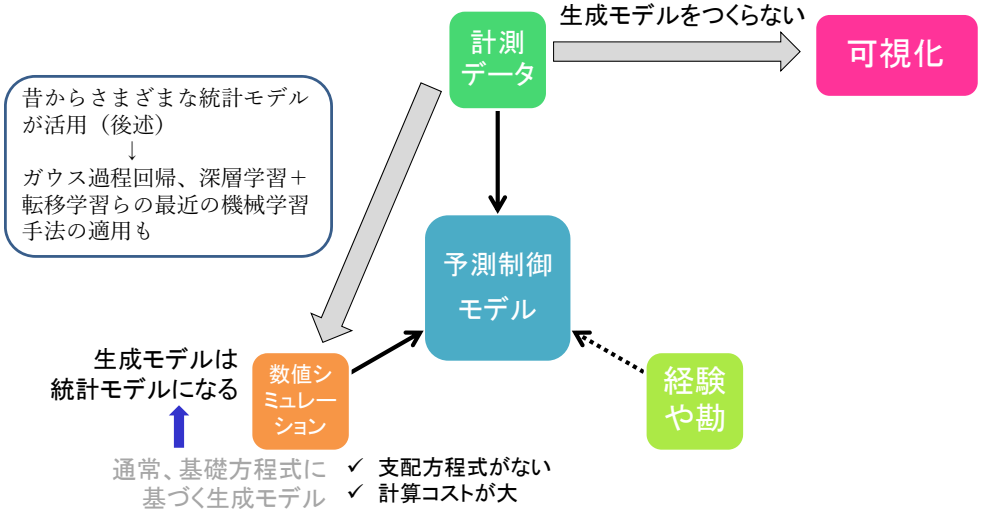
識別モデル

$D(\boldsymbol{y})$: $p(C|\boldsymbol{y},\boldsymbol{\theta})$　条件付き分布

識別関数：　$\boldsymbol{y}$ を C に写像する関数

計測データ駆動型生成モデル:　予測モデル

従来のプロセス制御

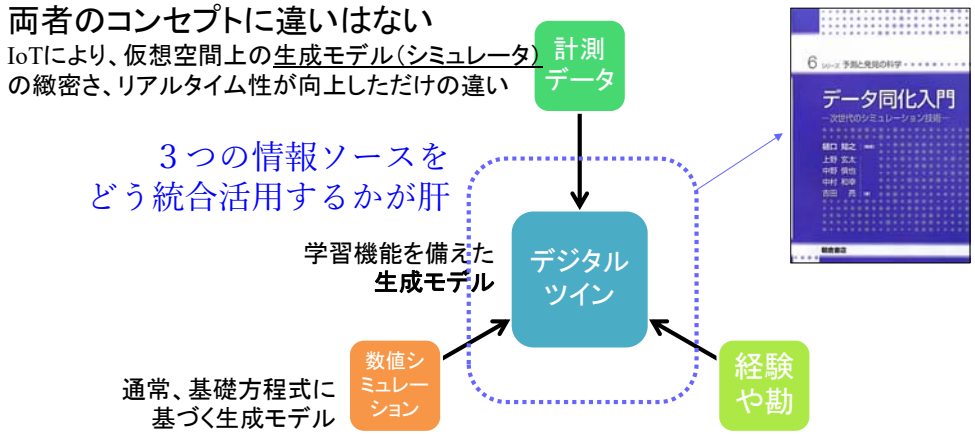


デジタルツインとCPS(Cyber Physical System)

物理世界の出来事を、そっくりそのままデジタル上にリアルタイムに再現する

両者のコンセプトに違いはない

IoTにより、仮想空間上の生成モデル（シミュレータ）の緻密さ、リアルタイム性が向上しただけの違い



Emulation

気候変動の確率地図

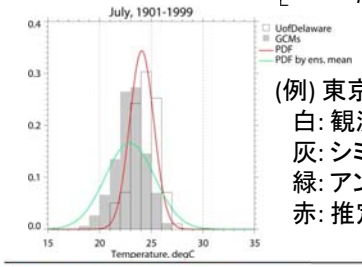
（上野玄太）
文科省創生プログラム「気候変動予測データの統計学的解析手法の開発」

複数の全球気候モデルのシミュレーション結果を用い、気候変動を確率的に評価する手法を開発し、東アジア及び日本における気温の将来変化の確率地図を初めて作成した。

観測値　シミュレーション

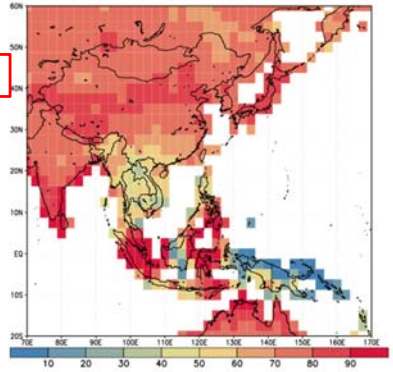
スパース回帰モデル

$$J\left(\beta_0,\beta_1,\cdots,\beta_N\right)=\sum_{t=1}^{T}\left(y_t-\beta_0-\sum_{n=1}^{N}\beta_n x_t^{(n)}\right)^2$$
$$+\gamma\left[(1-\alpha)\sum_{n=0}^{N}\beta_n^2+\alpha\sum_{n=0}^{N}\left|\beta_n\right|\right]$$



(例) 東京の気温
白: 観測値
灰: シミュレーション
緑: アンサンブル平均
赤: 推定した確率分布

21世紀後半の気温上昇の確率地図



2℃以上昇温する確率

東アジア及び日本の多くの地域で2度以上月平均気温が上昇する確率が70～80%以上。

SOUSEI Program for Risk Information on Climate Change 気候変動リスク情報創生プログラム

利用したデータセットは、国家基幹技術「海洋地球観測探査システム」：データ統合・解析システム(DIAS) の枠組みの下で収集・提供されたものである。

エミュレーション、仮想計測

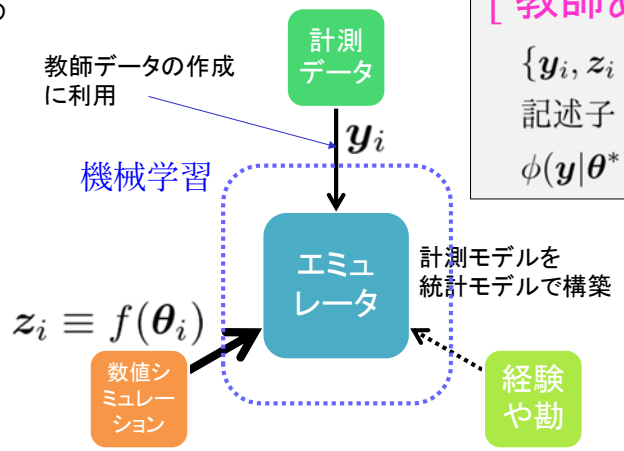
マテリアルズインフォマティクスの典型例の一つ

[教師あり学習]

$\{\boldsymbol{y}_i, \boldsymbol{z}_i \equiv f(\boldsymbol{\theta}_i)\}_{i=1}^{N}$

記述子 ⇓

$\phi(\boldsymbol{y}|\boldsymbol{\theta}^*)$

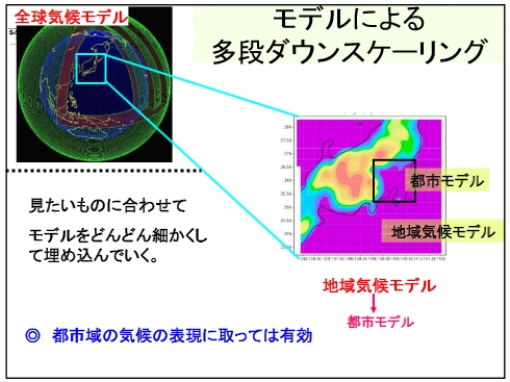


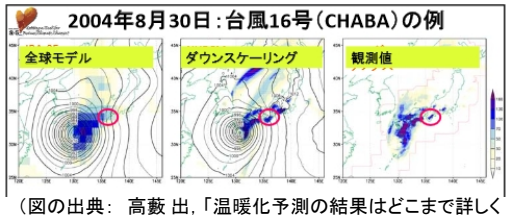
基礎方程式に基づく生成モデルでビッグデータを産出

ダウンスケーリング＆サロゲートシミュレーター

時間・空間高解像度の予測計算サービス

・力学的
・統計的





（図の出典： 高藪 出，「温暖化予測の結果はどこまで詳しく出来るか？」，10/14/11 気候シナリオ「実感」プロジェクト S-5 シンポジウム スライドより）

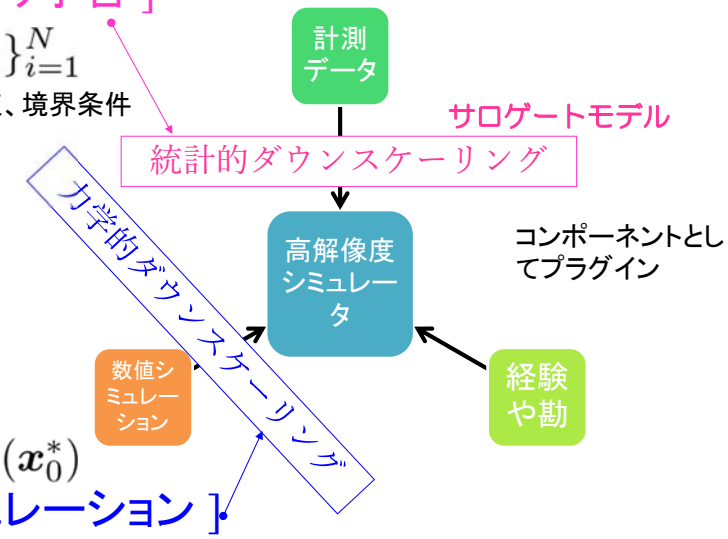
ダウンスケーリング：超解像

[教師あり学習]

$\{\boldsymbol{y}_i, \boldsymbol{x}_{0,i}\}_{i=1}^{N}$

⇓ 初期値、境界条件

$\phi(\boldsymbol{y}|\boldsymbol{x}_0^*)$



$\boldsymbol{y} \equiv g(\boldsymbol{x}_0^*)$

[シミュレーション]

Deep Convolutional Generative Adversarial Network (DCGAN)

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks
Alec Radford, Luke Metz, Soumith Chintala,

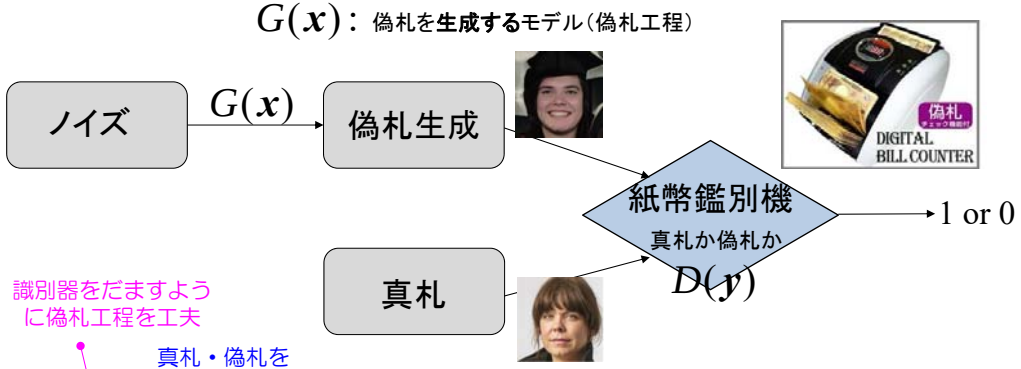
2016年

http://www.whichfaceisreal.com/index.php

いたちごっこ：GANの基本アルゴリズムの比喩的解説

$D(\boldsymbol{y})$：真札である確率

$G(\boldsymbol{x})$：偽札を生成するモデル（偽札工程）



識別器をだますように偽札工程を工夫

真札・偽札を正しく識別

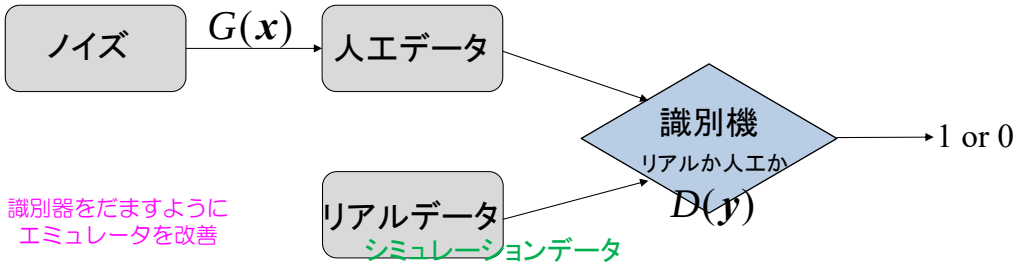
学習アルゴリズム

$$\min_G \max_D \mathrm{E}_{\boldsymbol{y}\sim p_{\mathrm{data}}(\boldsymbol{y})}\left[\ln D(\boldsymbol{y})\right] + \mathrm{E}_{\boldsymbol{x}\sim p_{\mathrm{x}}(\boldsymbol{x})}\left[\ln\left(1 - D(G(\boldsymbol{x}))\right)\right]$$

GANをつかってエミュレータをつくる

$D(\boldsymbol{y})$：リアルデータである確率

$G(\boldsymbol{x})$：エミュレータ



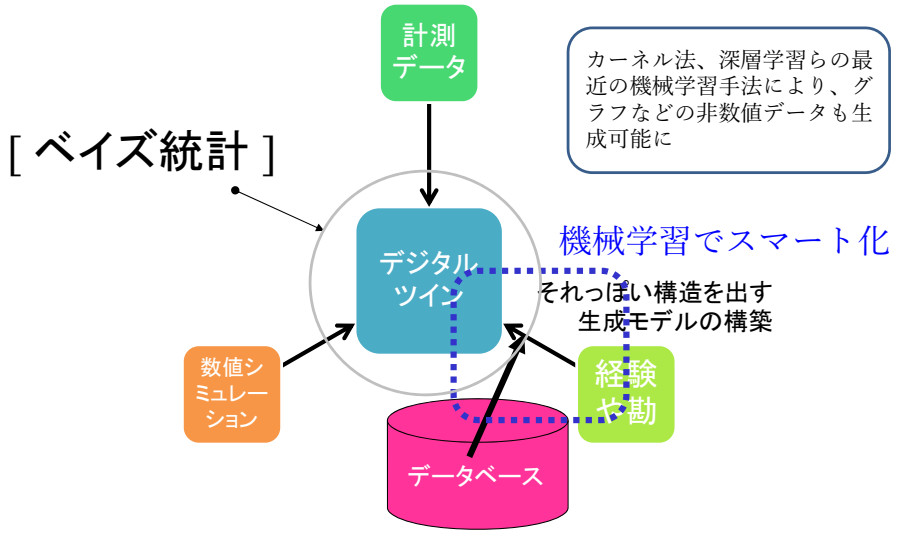
識別器をだますように
エミュレータを改善

リアルデータと人工
データを正しく識別

学習アルゴリズム

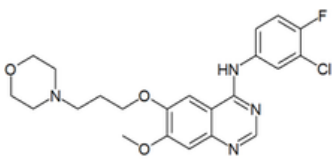
$$\min_G \max_D \mathrm{E}_{\boldsymbol{y}\sim p_{\mathrm{data}}(\boldsymbol{y})}\left[\ln D(\boldsymbol{y})\right]+\mathrm{E}_{\boldsymbol{x}\sim p_{\mathrm{x}}(\boldsymbol{x})}\left[\ln\left(1-D(G(\boldsymbol{x}))\right)\right]$$

暗黙知（経験と勘）生成モデル

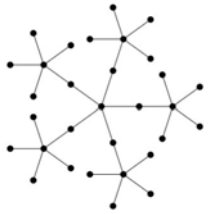


CHEMICALLY FAVORABLE STRUCTURE?

Molecular graph



General graph



$p(\underline{\boldsymbol{x}})$

prior knowledge

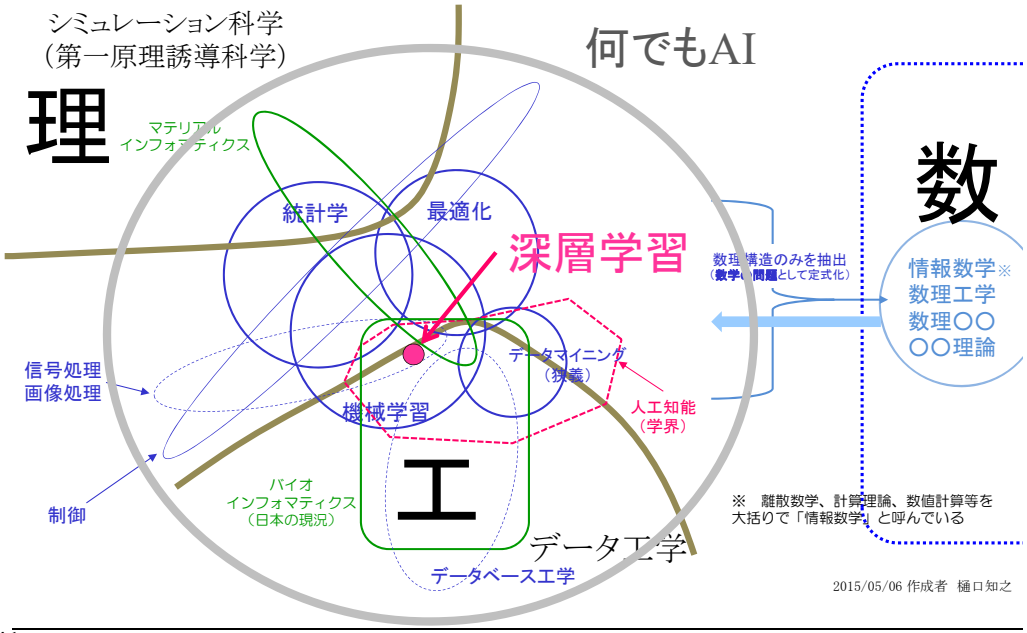
- Geometry
- Exclusion of chemically unfavorable bonds, *e.g.* F-N, C=O=C
- Chemical stability
- Drug-likeliness
- Synthesizability

（統数研・吉田准教授のスライドを借用）

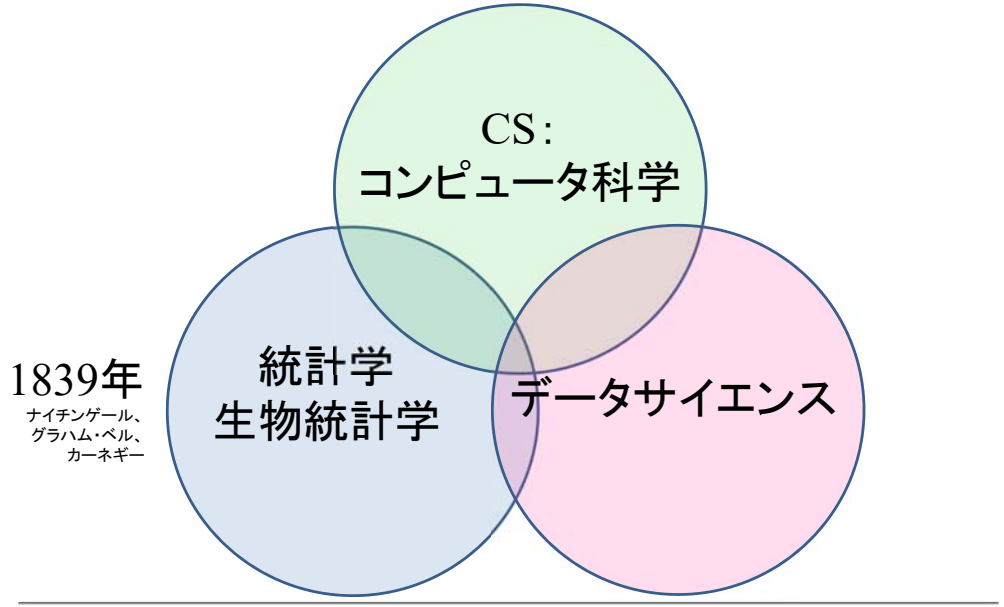
アウトライン

1. 演繹と帰納
2. データ同化
   a. 再訪
   b. ハミルトニアン・モンテカルロ法
3. 機械学習とシミュレーションの融合
4. 人材育成での課題

データに関連した数理分野の俯瞰図



学部、研究科：世界と日本の歴史の違い



シミュレーション科学の人材

- 数学基礎力がある **「地頭（じあたま）」が良い**
  - ✓ 量子力学を修得
  - ✓ データ解析でも、回折や分光の数理的枠組みの習得が必須
- モデリングの素養あり
  対象の時間・空間変動スケールにあわせた、さまざまな近似計算のやり方も知っている

- 物理帝国主義に染まっている（逆解析が弱い）
- 最近の帰納法技術の発展をあ[illegible]ない
- コト作り（サービス創発）のセンスがない

PDCAをまわすのが苦手

何が肝か？：データサイエンスはネットワーク型知識構造

薄く幅広く習得するのがよい。引き出しをたくさん持っておくこと。

理由： Webでいくらでも自学自習できる。実務でもしっかり体得できる。
「厳密・唯一」から、「近似・アンサンブル」の世界観を身につけさせる。

- 高次元
- 超大サンプル数 （オンライン）
- 非線形
- 最適化
- シミュレーション
- プログラミング

- ビッグデータをベースに、従来、分断されていた教育内容を整理・融合
- 学部前期教育もそれに向けた数学の基礎とデータリテラシーに