

データ駆動科学の発展とインパクト

樋口知之

中央大学工学部 教授
AI・データサイエンスセンター 所長

 一般社団法人
データサイエンティスト協会 理事

(株)ブレインパッド 顧問
(株)ブリヂストン 特別技術顧問

新エネルギー・産業技術総合開発機構(NEDO)
ロボット・AI部 PL

産業技術研究所 人工知能研究センター 顧問

都市伝説「現場にはデータが大量にある」

良くある例:

業務(作業)報告書に数値を記載する欄

US: N/A (責任をとらされるので書かない)

Japan: 前回の記録を見ながら予測、空想で埋める
(空欄だと上司に怒られる)

結果: 意味の無い誤差だらけのデータがあるだけ!

この企業文化を変えることが重要な第一歩!

Science Paradigms: Four Kind of Methodology (Approach) of Science

演繹法

Deductive Approach

C: Computational

数値シミュレーション

Cyber space (Computer-Intensive)

Theory-driven method

First principle driven method

T: Theoretical

理論

帰納法

Inductive Approach

B: Big Data driven

機械学習

Data-driven method

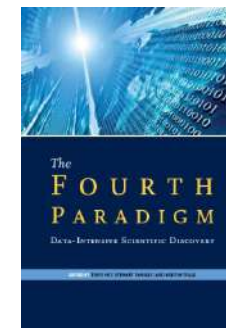
E: Empirical

Experimental

実験、経験

(Axle)
Data Assimilation

Driving Force of Science



By Tony Hey, Stewart Tansley, and Kristin Tolle, Eds. Microsoft Research, 2009.

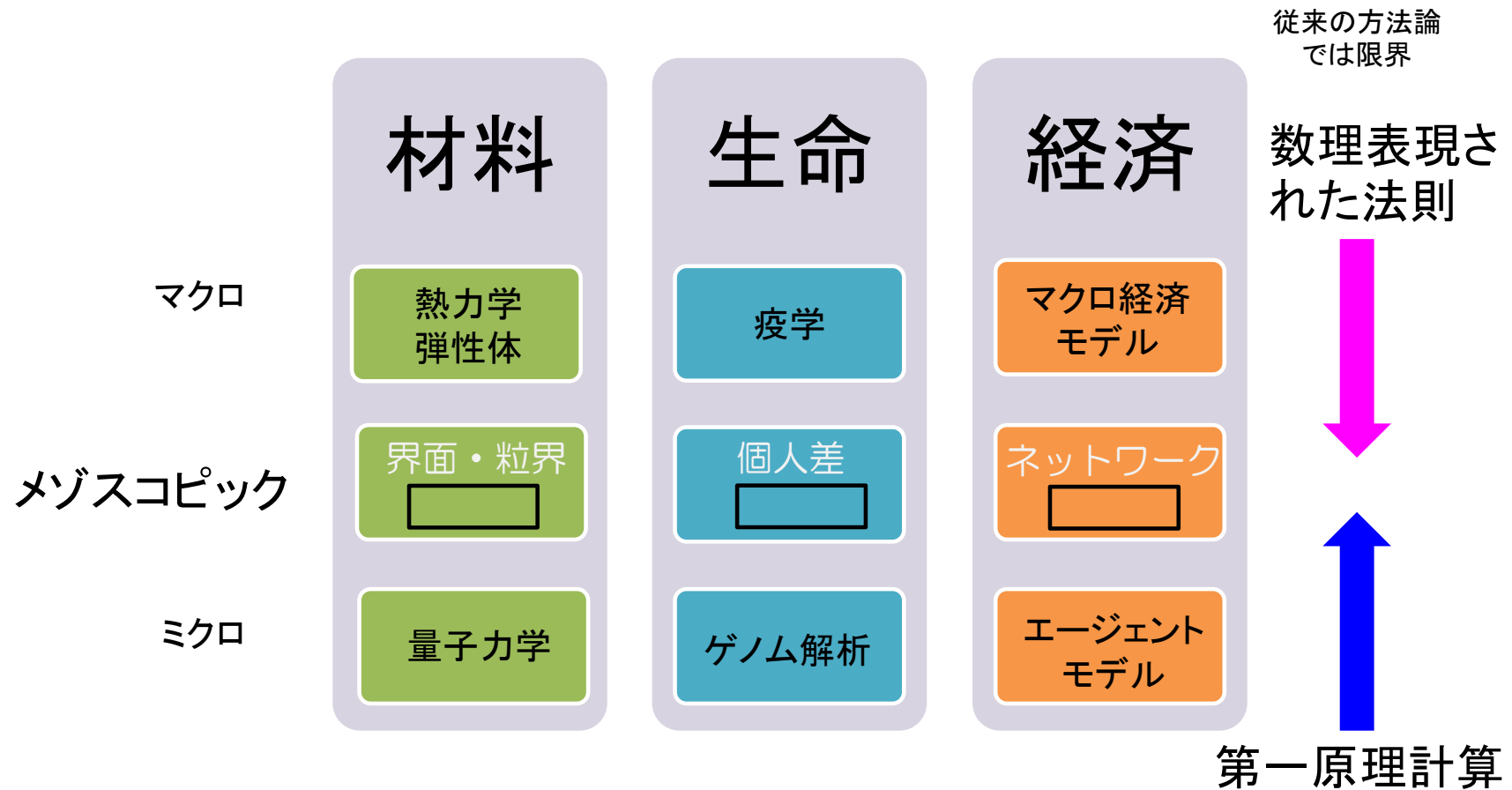
Science Paradigms

- Thousand years ago: science was **empirical** describing natural phenomena
- Last few hundred years: **theoretical** branch using models, generalizations
- Last few decades: a **computational** branch simulating complex phenomena
- Today: **data exploration** (eScience) unify theory, experiment, and simulation
 - Data captured by instruments or generated by simulator
 - Processed by software
 - Information/knowledge stored in computer
 - Scientist analyzes database/files using data management and statistics

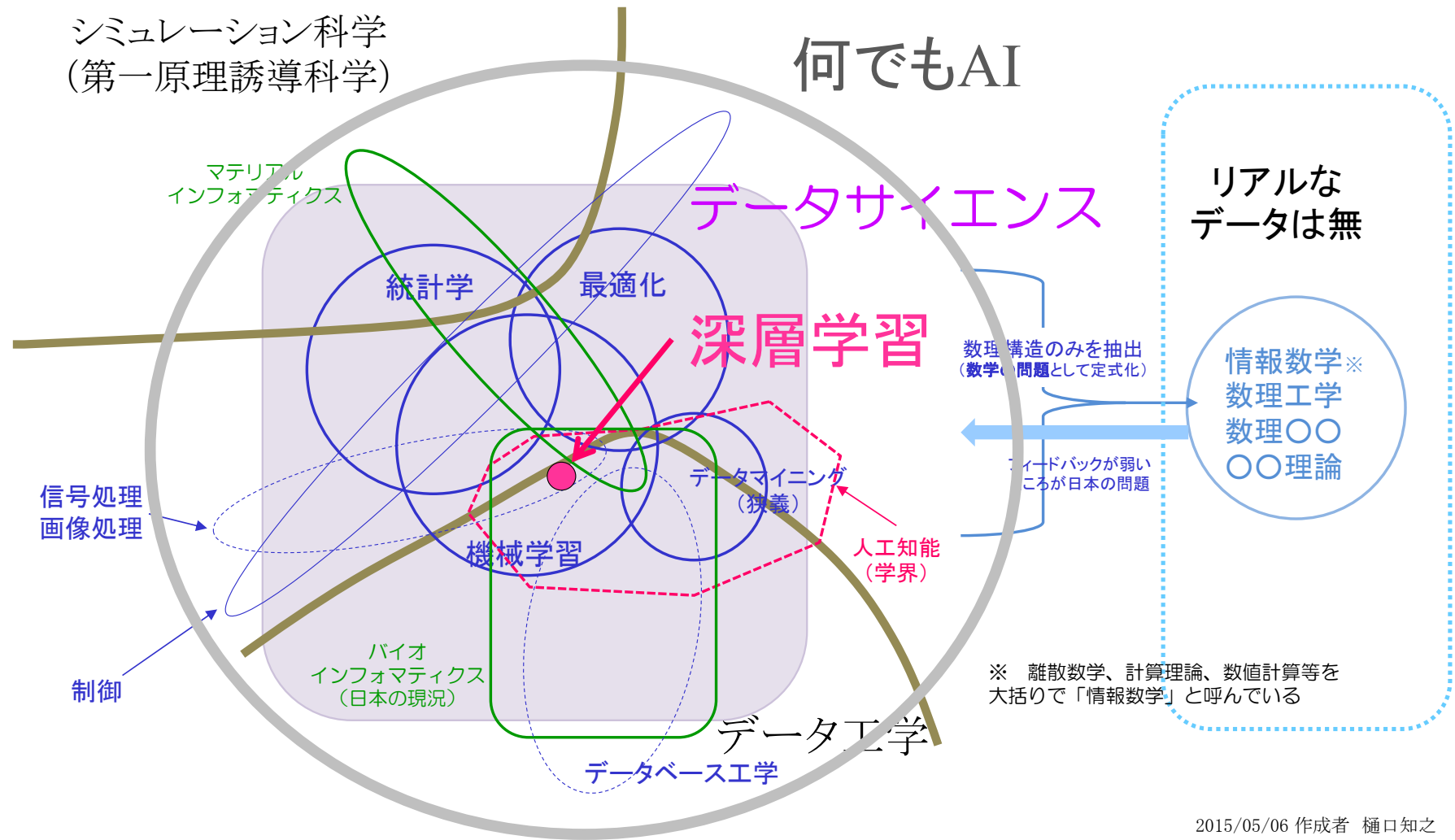
$$\left(\frac{a}{a'}\right)^2 = \frac{4\pi G\rho}{3} - K\frac{c^2}{a'^2}$$



メゾスコピック・モデリング



データに関連した数理分野の俯瞰図



2015/05/06 作成者 樋口知之

統計的機械学習を規定する *Triplet*

課題発見、問題設定

目的

Supervised learning

判別、分類

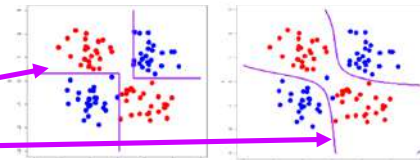
回帰

Unsupervised learning

クラスタリング、
次元削減

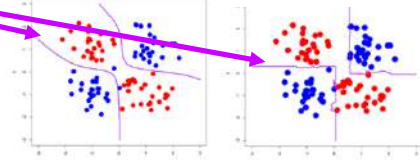
決定木

SVM+Gaussian Kernel



Neural Network

ランダムフォレスト



Original Source
<http://tjo.hatenablog.com/entry/2014/01/06/190456>

非線形関数

Data

Latent variable

Parameter

$$f(|\{y_i\}, \{x_j\}, \{a_k\})$$

機械学習

(目的 | モデル | 学習アルゴリズム)

$$w_{i,j}^{[k+1]}$$

深層学習

入力データ



結果

Higuchi

State Space Model, Gaussian Process,

Support Vector Machine, Neural Network

DNN

DL is one of the technique in statistical Machine Learning

内挿と外挿

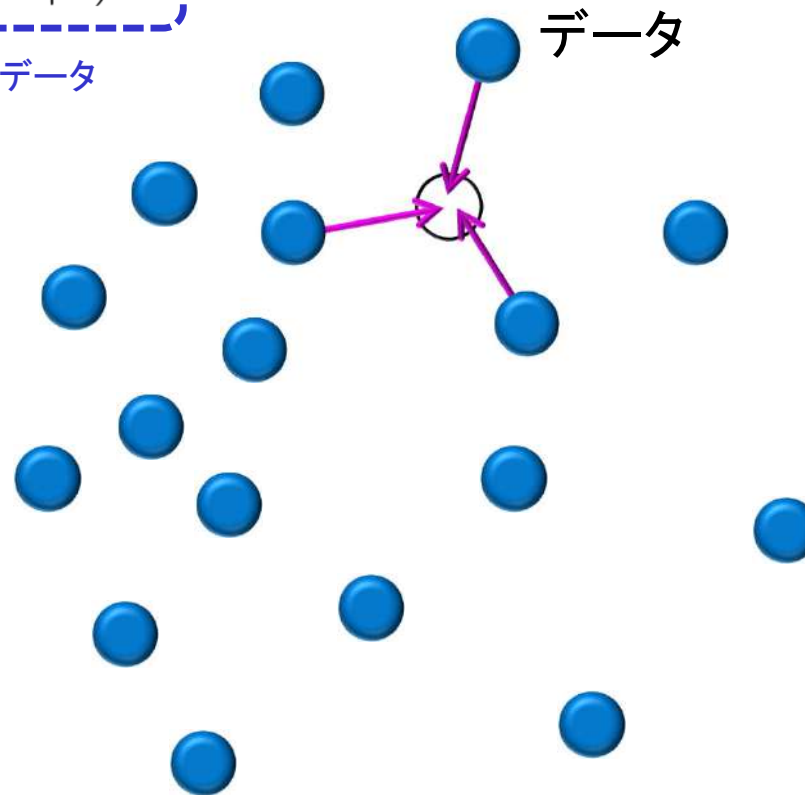
教師あり機械学習

$$\{x_i, y_i\}_{i=1}^N \longrightarrow y_* = f(x_* | \hat{\theta})$$

相互に独立

未来の入力データ

高次元サンプルを
2次元に射影
(適切な射影空間を求める
こと自体が難問)



データ無しの領域

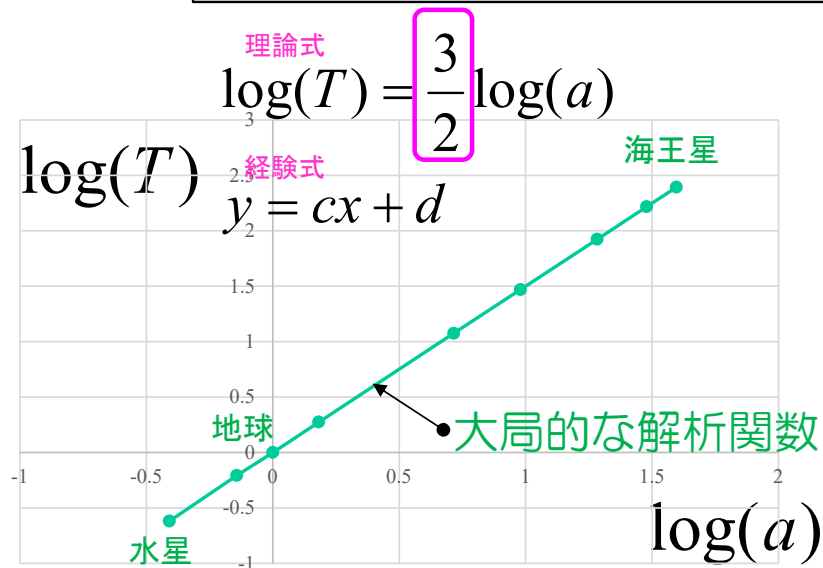
内挿型研究はパワーゲーム！

概括：帰納法とデータ今昔

ケプラーの第三法則

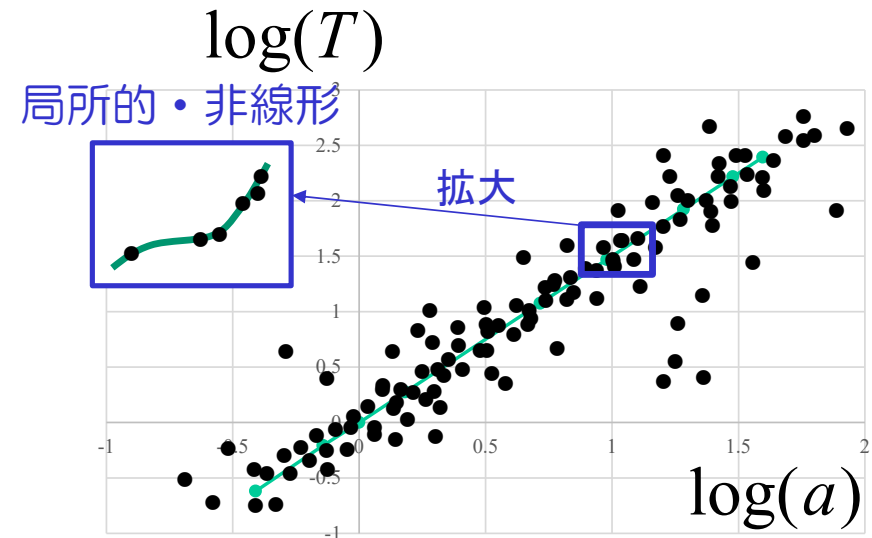
師匠であったティコ・ブラーエの観測記録から推定し定式化

惑星の公転周期 T の2乗は、楕円軌道の半長軸 a の3乗に比例する



ティコ・ブラーエの観測データ

狙って観測、帰納推論から経験則、そして一般則(万有引力の法則)を導出



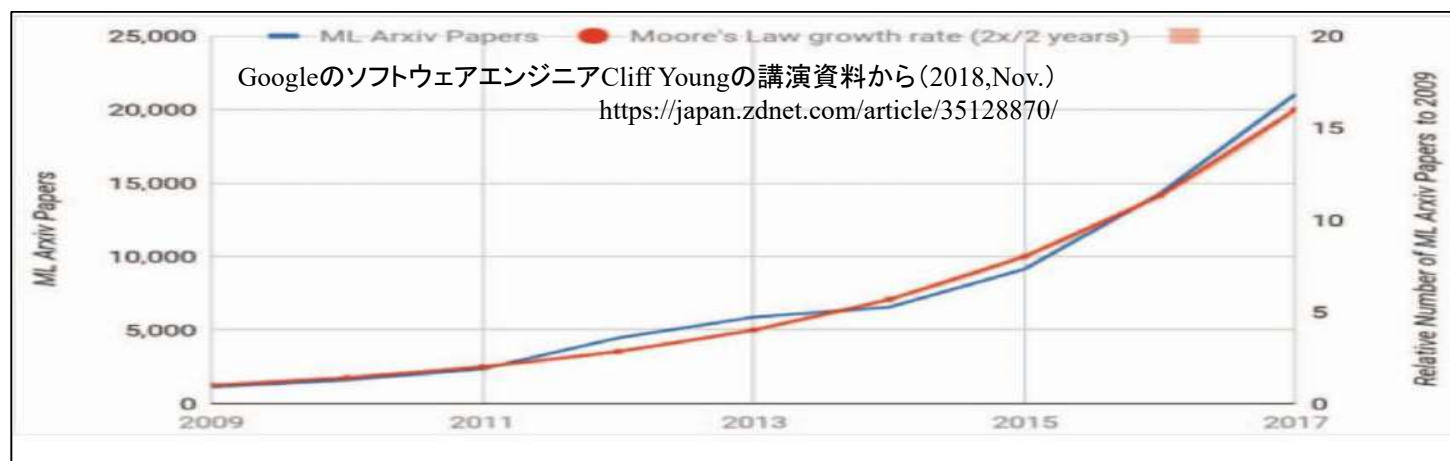
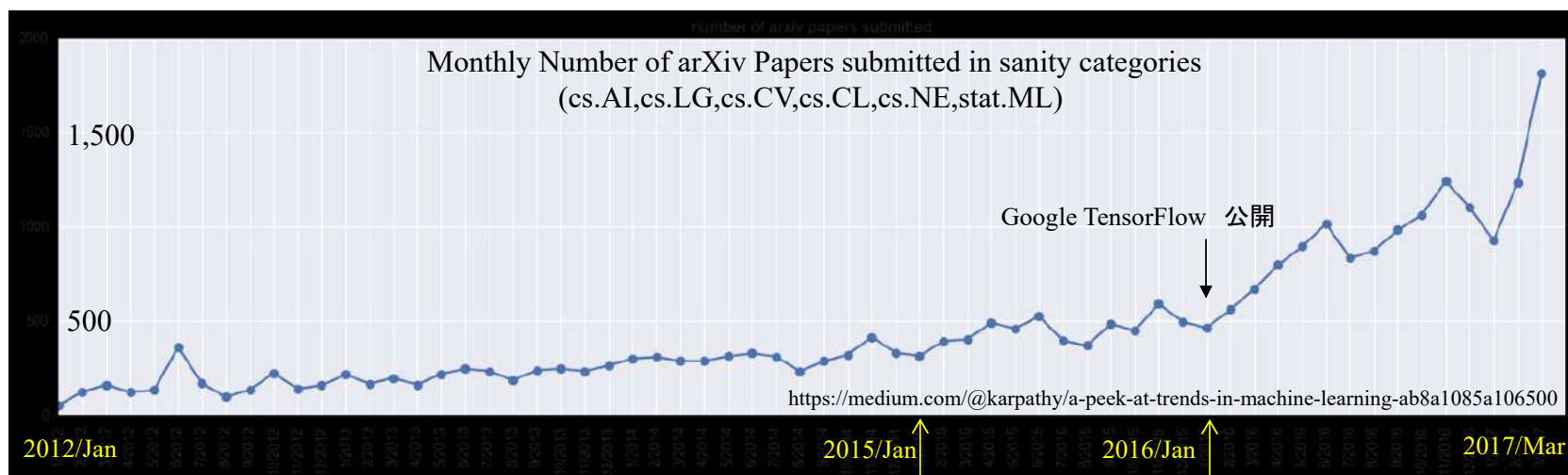
ビッグデータ

無目的・副産物的にデータが蓄積され、経験則のみでOK(目的-予測・判別-が達成)

ビッグデータ時代以前: 良質な空間を見つけること

1) 説明変数の選択 2) 線形性 3) 少ないパラメータ

近年の動き：ムーアの法則 $2x/1.5\text{years}$ を超える勢い



樋口知之, ムーアの法則を超える論文数
化学と工業 72(9) 747 - 747, 2019

機械学習の有用性がさらに浸透する ダウンスケーリング：超解像

[教師あり学習]

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



初期値、境界条件

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

低解像度初期条件・
境界条件
→ 高解像度の解

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

[シミュレーション]

計測
データ

サロゲートモデル

統計的ダウンスケーリング

高解像度
シミュレー
タ

コンポーネントとし
てプラグイン

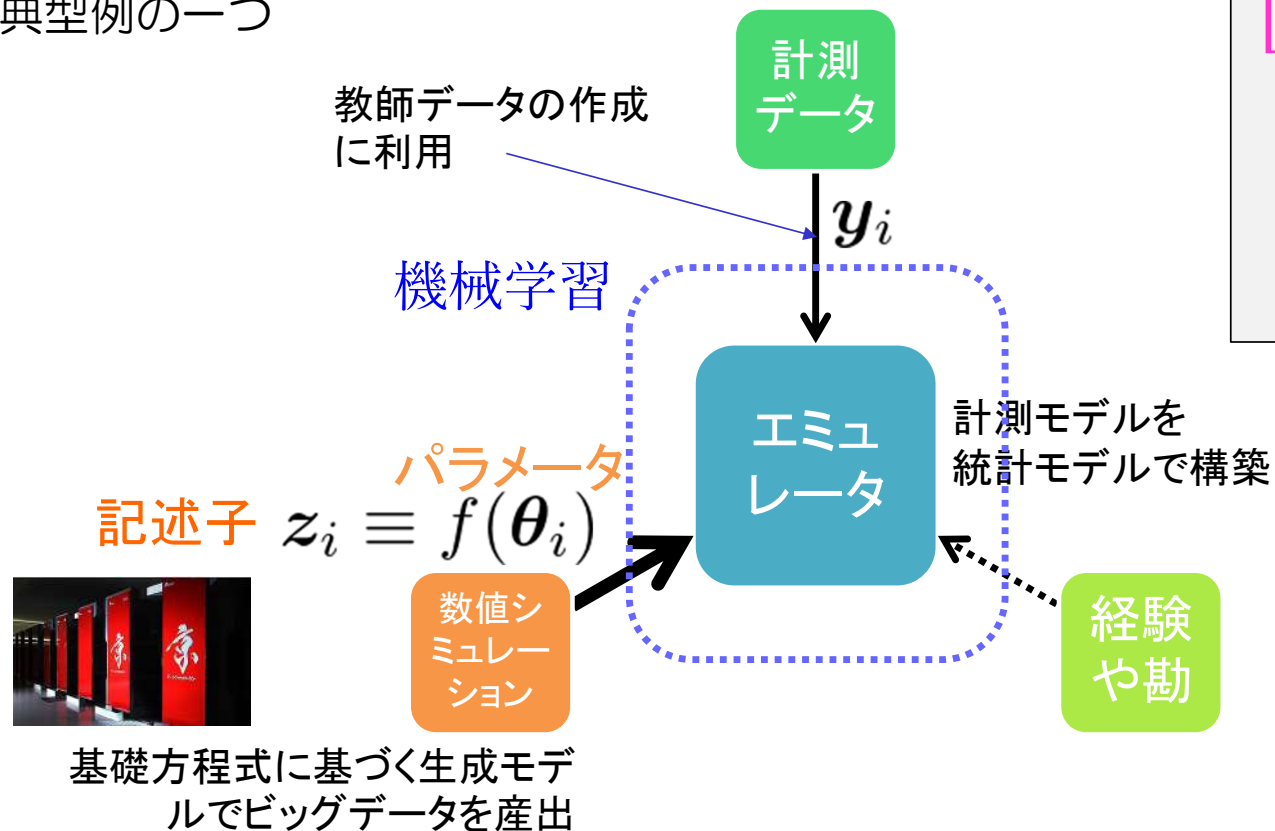
経験
や勘

数値シ
ミュレー
ション

数学的ダウンスケーリング

融合が進む：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]

$$\{y_i, z_i \equiv f(\theta_i)\}_{i=1}^N$$

記述子 $\downarrow$

$$\phi(y|\theta^*)$$

さらに帰納法の中でも：機械学習vs.統計的モデリング

深層学習 vs. 旧世代 機械学習

JSM2019 David van Dyk (Imperial College)
<https://wwwf.imperial.ac.uk/~dvandyk/>



Astrostatisticsの専門家
LIGO（重力波）、ブラックホール撮像

- Data-driven vs. Science-driven
- Predictive model vs. Descriptive model
- Correlation vs. Causality

「認識科学」から「設計科学」へのシフト

「対象理解」から「機能の最適化」
へ興味がシフト

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）

$$p(\mathbf{y} \mid \boldsymbol{\theta}, \mathbf{z})$$

データ パラメータ 説明変数(アспект)

クラス分類問題

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

例: $C=1$ or 0

$$p(C = C_i \mid \mathbf{y}, \boldsymbol{\theta}) = \frac{p(\mathbf{y}, C = C_i \mid \boldsymbol{\theta})}{\sum_i p(\mathbf{y}, C = C_i \mid \boldsymbol{\theta})}$$

生成モデル（狭義）

$$p(\mathbf{y}, C \mid \boldsymbol{\theta}) \quad \text{同時分布}$$

データ クラス

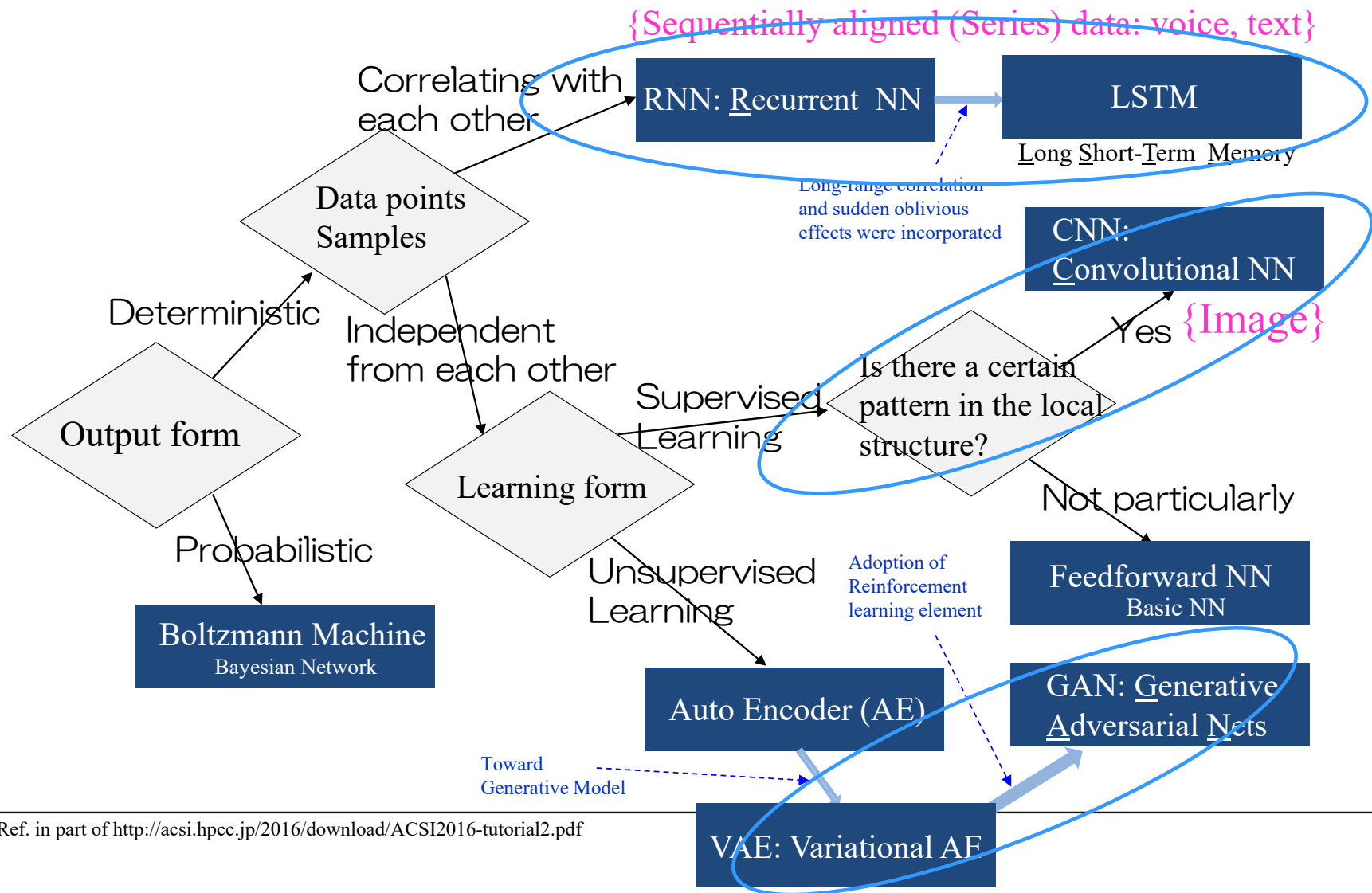
$$p(\mathbf{y} \mid C, \boldsymbol{\theta}_2) \cdot p(C \mid \boldsymbol{\theta}_1)$$

識別モデル

$$p(C \mid \mathbf{y}, \boldsymbol{\theta}) \quad \text{条件付き分布}$$

識別関数: $\mathbf{y}$ を C に写像する関数

Classification of D N N

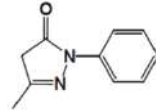


Ref. in part of <http://acsi.hpcc.jp/2016/download/ACSI2016-tutorial2.pdf>

非効率（現実には機能しない）探索法

構造

x



$f(x)$

物理あるいは化学的指数や係数、特性の計算値

機能発現

y

その(実験値あるいは経験値)データ

膨大な数のモンテカルロ計算

x の候補を恣意的に多数発生

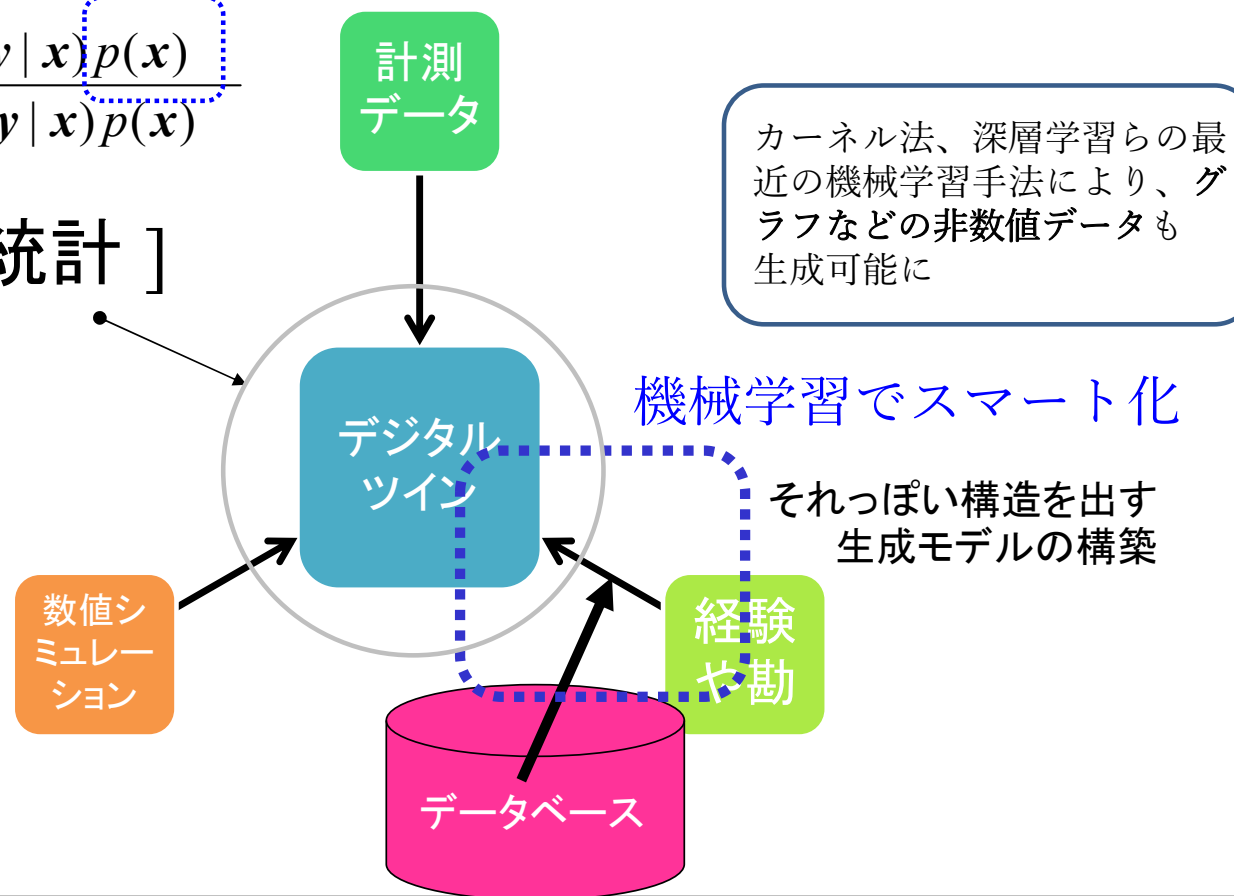
一つ一つの x に対して順解析(MDなどの第一原理計算)を実施。データとの整合性 $p(\underline{y} | f(\underline{x}))$ を計算

エキスパートが目と勘で判断

暗黙知（経験と勘）を具現化する生成モデル

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

[ベイズ統計]



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

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

ベイズの反転公式

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

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



■ Virtual screening, QSAR modeling: $P(Y|G)$ Forward Problem

Quantitative Structure-Activity(Affinity) Relationship

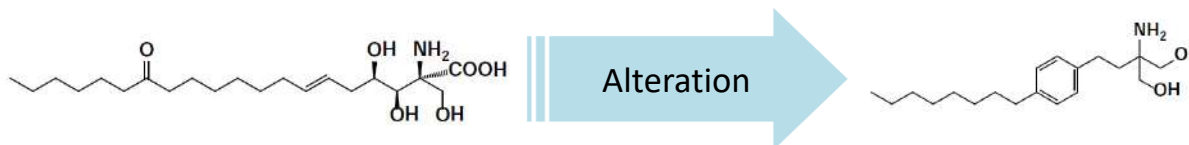
Develop statistical models to predict biochemical or physiochemical activities Y of an input chemical structure G



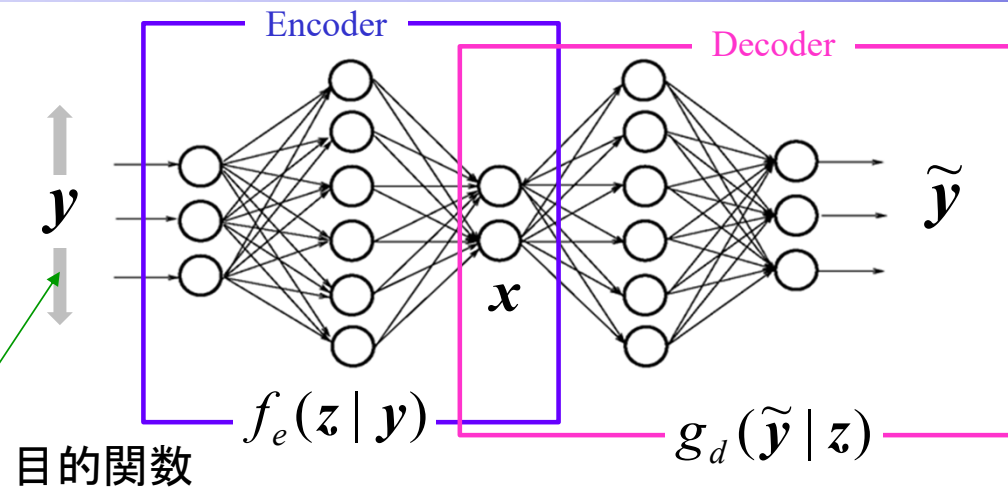
■ Chemical design, Inverse-QSAR: $P(G|Y=y)$ Inverse Problem

Generate novel chemical structures G achieving desired activities $Y=y$

Preimage reconstruction of the graph kernel using a MCMC algorithm



Denoising AE and Generalized DAE



$$L = \max E_y [\log p(\tilde{\mathbf{y}} | \mathbf{y})] \Rightarrow \min \sum_{y_i} (\mathbf{y}_i - g_d(f_e(\mathbf{y}_i)))^2$$

$C(\mathbf{y}' | \mathbf{y})$: known corruption process (Conditional distribution $p(\mathbf{y}' | \mathbf{y})$)

DAE

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, Pierre-Antoine Manzagol, Extracting and composing robust features with denoising autoencoders, Proceedings of the 25th international conference on Machine learning, 2008

Surrogate Analysis, Kernel Bootstrap,...

2008年

$$L = E_{\mathbf{y}, \mathbf{y}'} [\log p(\tilde{\mathbf{y}} | \mathbf{y}) - \lambda R(\mathbf{y}, \mathbf{y}')] \quad \text{Regularization function}$$

GDAE

Yoshua Bengio, Li Yao, Guillaume Alain, Pascal Vincent, Generalized denoising auto-encoders as generative models, Advances in Neural Information Processing Systems, 2013

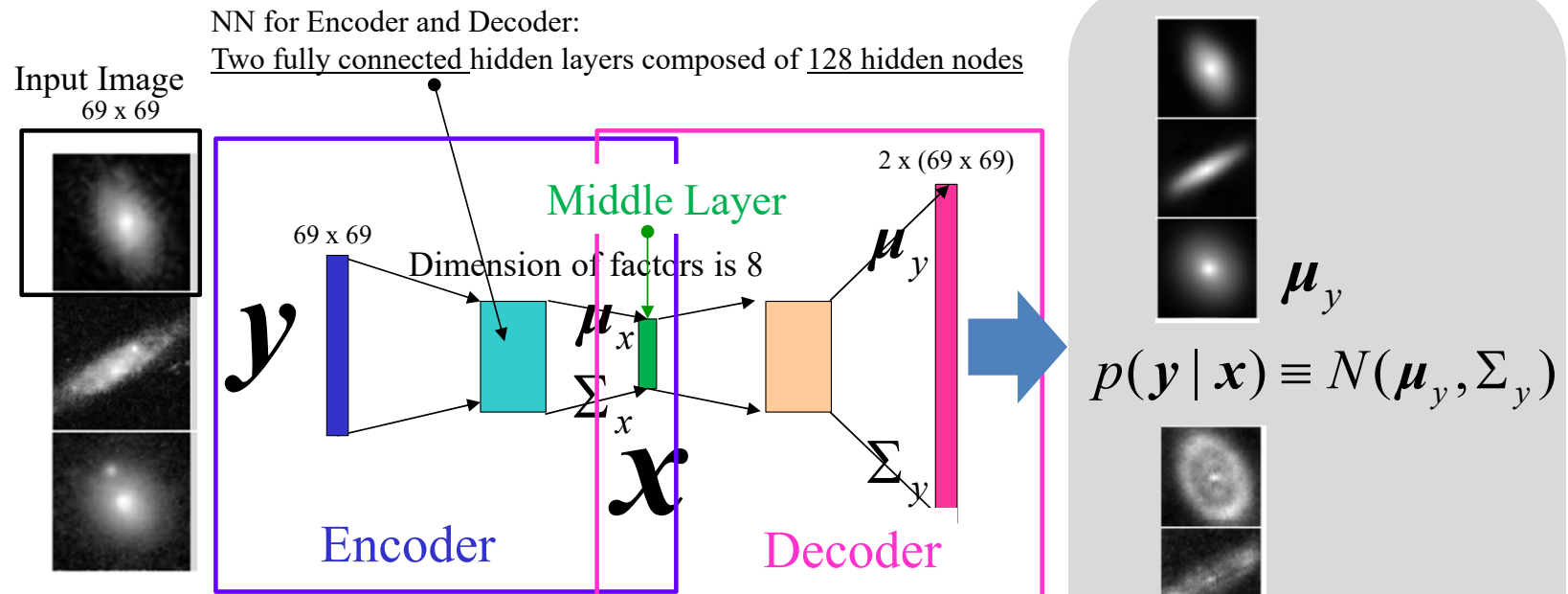
2013年

Variational Auto-encoder

Generative Model

Galaxy shape modeling with probabilistic auto-encoders (2014)

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In ICLR,



2017年4月

A look at deep learning for science

Scientific use cases show promise, but challenges remain for complex data analytics.

By Prabhat, April 3, 2017

<https://www.oreilly.com/ideas/a-look-at-deep-learning-for-science>

$$\mathbf{x} = \mu_x + \Sigma_x \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N(0, I_{8 \times 8})$$

潜在変数ベクトル

From <http://people.eecs.berkeley.edu/~jregier/publications/regier2015deep.pdf>

Deep Convolutional Generative Adversarial Network (DCGAN)

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

2016年

敵対的生成ネットワーク

Click on the person who is real.

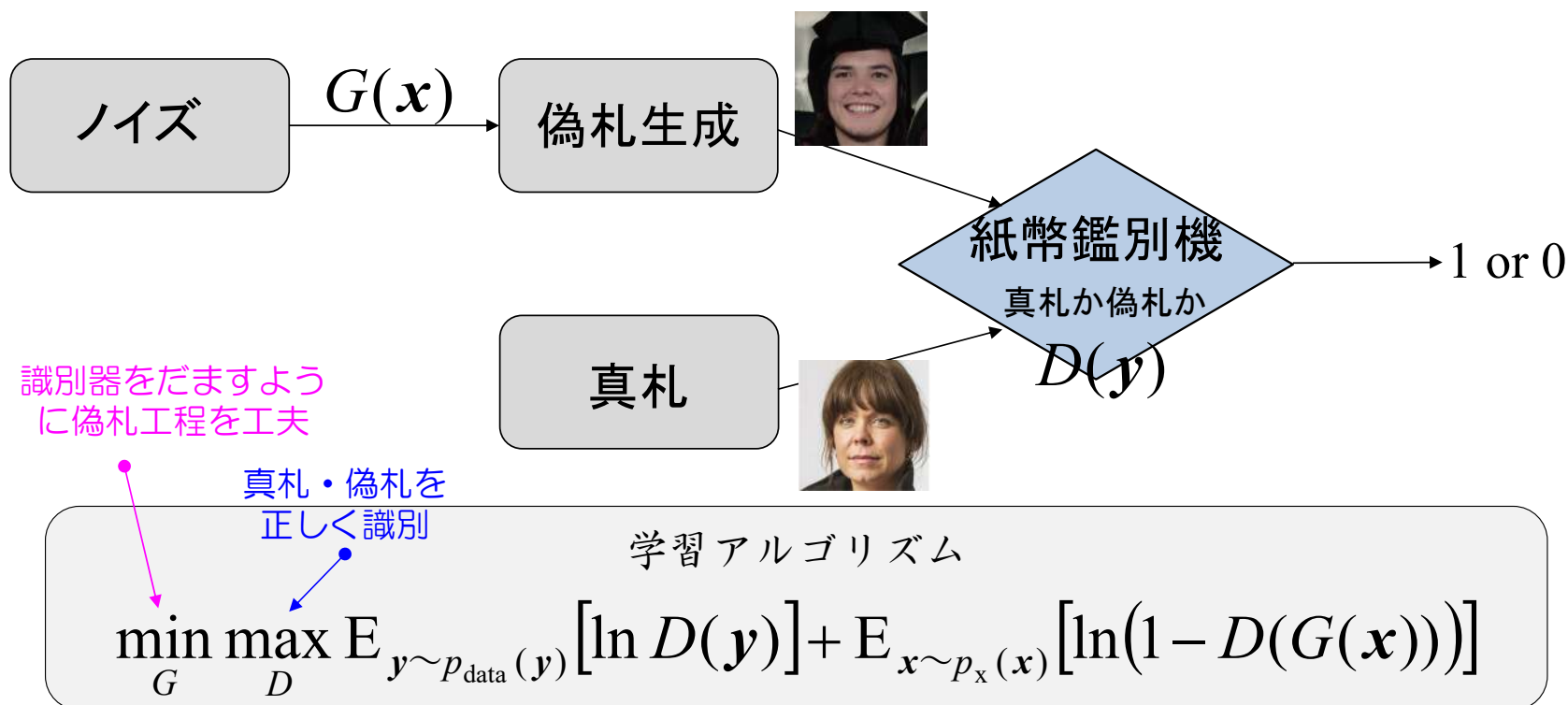
<http://www.whichfaceisreal.com/index.php>



いたちごっこ：GANの基本アルゴリズムの比喩的解説

識別モデル $D(\mathbf{y})$ ：真札である確率

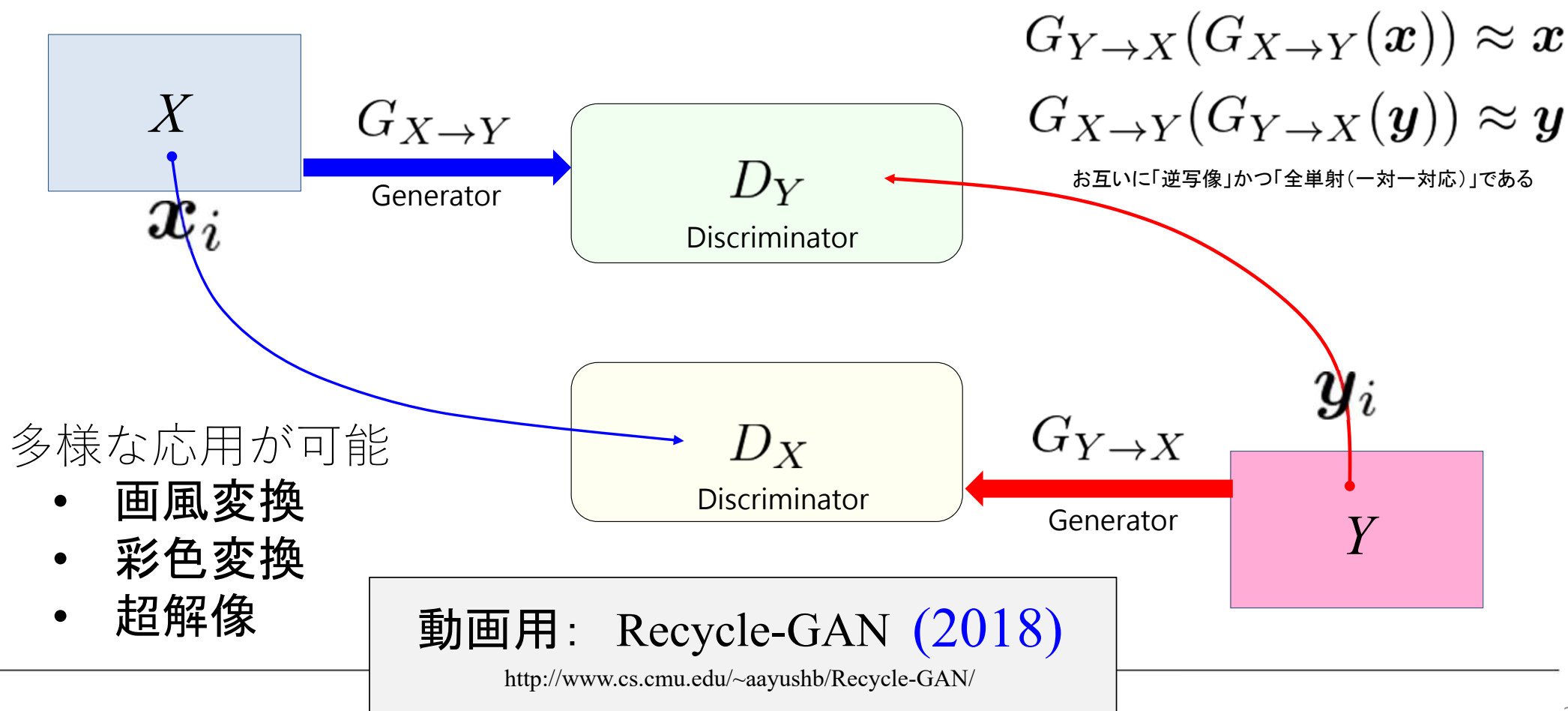
生成モデル $G(\mathbf{x})$ ：偽札を生成するモデル(偽札工程)



GANを使った報告は、うまくいった例しか出さないケースが多々あり！

CycleGAN: 画像の変換と生成モデルの自動生成 Jun-Yan Zhu, ICCV (2017)

教師無し学習: 非ペアの教師データから画像変換を実現



AIの影 研究不正: 特定不正行為

研究不正ガイドライン文科省 (p10)

1. ねつ造
2. 改ざん
3. 剽窃

超解像(心眼)

- 捏造とは、存在しないデータ、研究成果等を作成すること。
- 改ざんとは、研究資料、機器・過程を変更する操作を行い、データ、研究活動によって得られた結果等を真正でないものに加工すること。