機械学習技術の進展とその数理基盤

鈴木大慈
東京大学大学院情報理工学系研究科数理情報学専攻
JSTさきがけ
理研AIP

数理システムユーザーコンファレンス
2017年11月2日

人工知能と機械学習

人工知能

本日の様相
「人工知能」≒「機械学習」

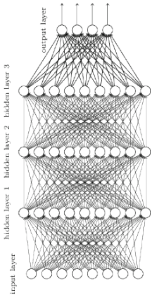
自分で問題設定ができ，
その解決もできる．

“汎用人工知能”

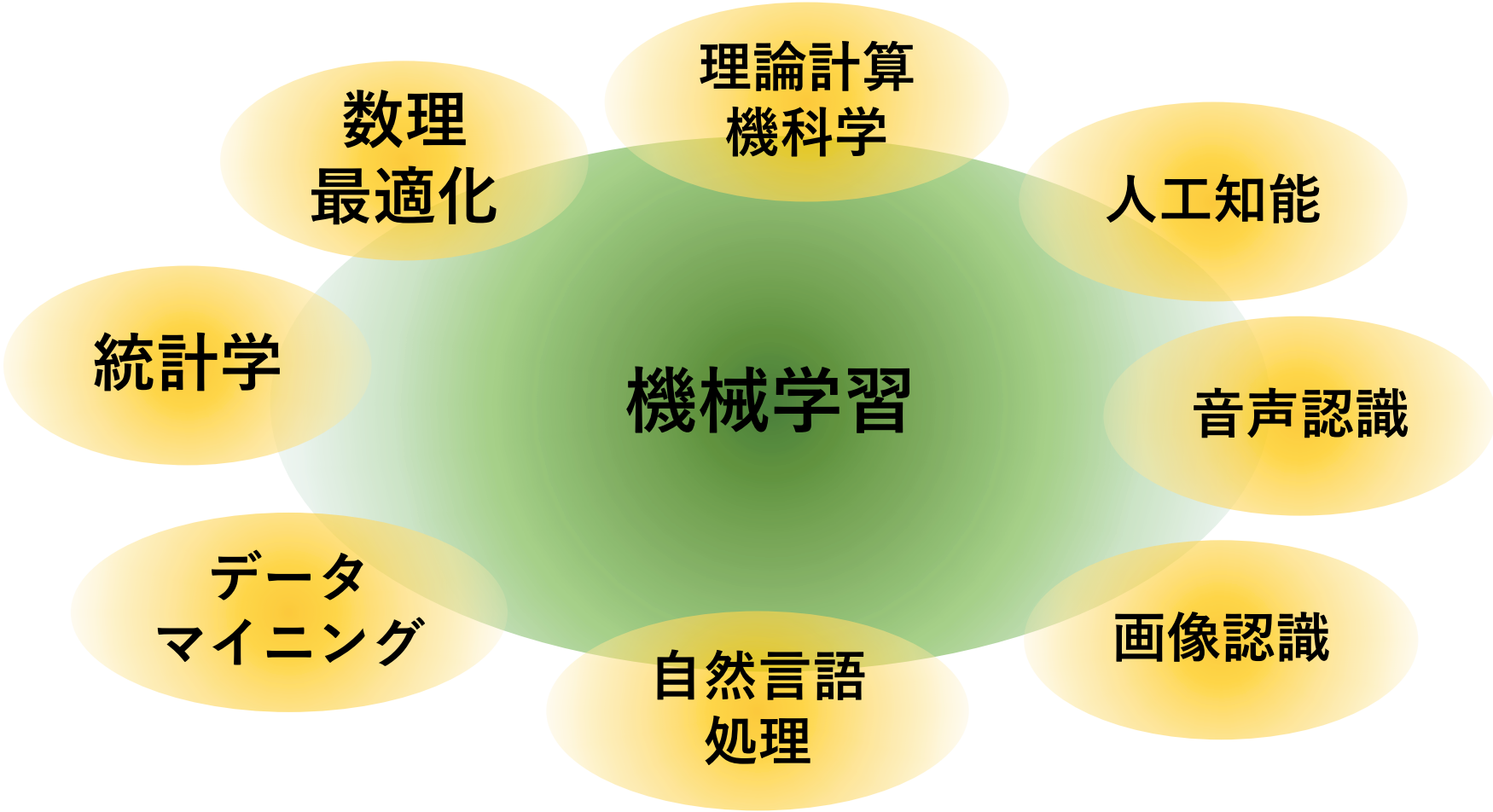
機械学習
統計的アプローチ

SVM
トピックモデル
スパース学習
テンソル学習
…

深層学習



機械学習の立ち位置



さまざまな分野の複合領域

機械学習コミュニティの実体

機械学習の主戦場は「**国際会議**」

NIPS (Neural Information Processing Systems)
ICML (International Conference of Machine Learning)
COLT (Conference of Learning Theory)
AISTATS, UAI, ECML, …

- 全て査読あり：ダブルブラインド，採択率20～25%
  - NIPS2016 (568/2500, 22.7%), ICML2016 (322/1327, 24.3%)
  - ストーリー＋理論＋実験＋読みやすさ
- これらの会議に論文が通っていることが重要
- 各国際会議はワークショップが併設

- 速報性
- テーマの変遷が速い
- 顔が見える

ICML2016@NYC

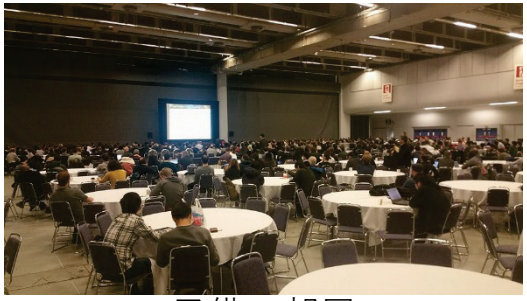

NIPS2015@Montreal

http://www.tml.cs.uni-tuebingen.de/team/luxburg/misc/nips2016/index.php

NIPS2015の様子

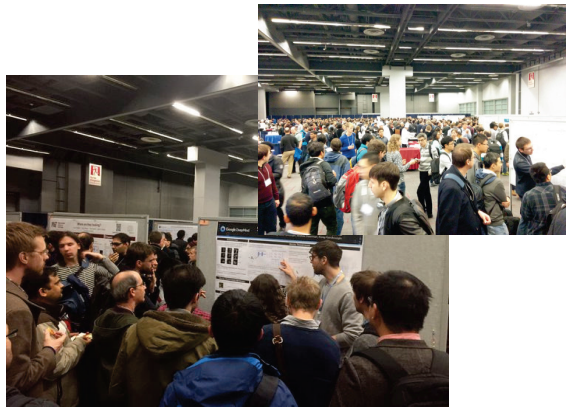
・初日はチュートリアル
・３日間の本会議
・２日間のワークショップ

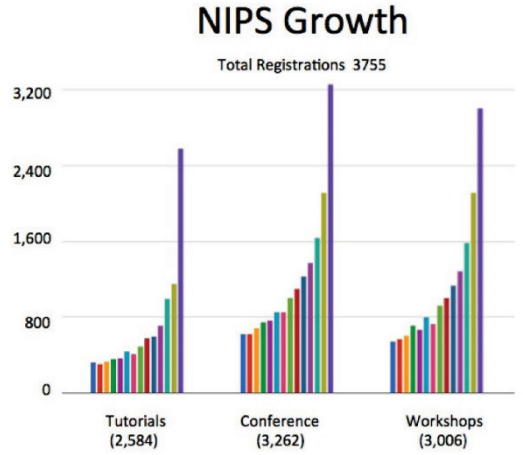

Deep learning tutorial


予備の部屋
入れなかった人たちはこちらへ

・本会議：
  - 朝から夕方までシングルトラック：招待講演＋オーラル発表(全体の3%!)
  - 夜はポスター（ほとんどの論文）：午後７時から**<u>午後１２時</u>**まで×**四日間.**
・ワークショップ：
  - 40種類のワークショップ＝20種類×２日間
  - Deep learning，最適化，ビッグデータ，機械学習ソフトウェア，…
・企業ブース多数，毎夜各企業のパーティーが開催（リクルーティング）


メイン会場
(2000人＋α)


ポスター発表



参加者は約4000人，2年で約2倍

周辺分野の国際会議

- データマイニング
  KDD，ICDM，WWW，WISDM，SIGIR，SDM
- コンピュータビジョン
  CVPR，ICCV，ECCV
- 自然言語処理
  ACL，NAACL，EMNLP，COLING
- 人工知能
  IJCAI，AAAI
- 理論計算機科学
  STOC，FOCS，SODA
- データベース
  VLDB，ICDE

※機械学習に関係の深い統計学と数理(連続)最適化はジャーナル文化

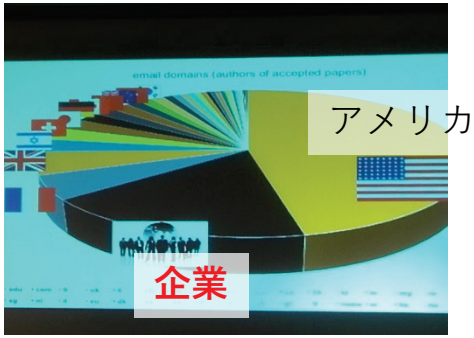
人の出入りと重複が激しい
会議文化としての共通点がある

企業との関係

- 多くの企業が論文を投稿/採録.
  Google, Facebook, Microsoft, Yahoo, Amazon, Baidu, …

例： NIPS2017に採択された論文中約30本に
Deep mindの著者

ICML2016採択比率



- 国際会議のスポンサー
  - 会議はリクルーティングの場でもある.
- 大学との共同研究，寄付
- インターン多数
  - 学生：実データを用いた研究＋経歴＋給料(月50万円～)
  - 企業：大学の優秀な学生と研究ができる．有名研究室とのコネ.
- 公開ライブラリ，公開データ
- 機械学習プラットフォーム

→ 良い研究をすることは企業のブランディングにとっても重要.
できるだけ若く優秀な研究者・開発者を雇用したい.

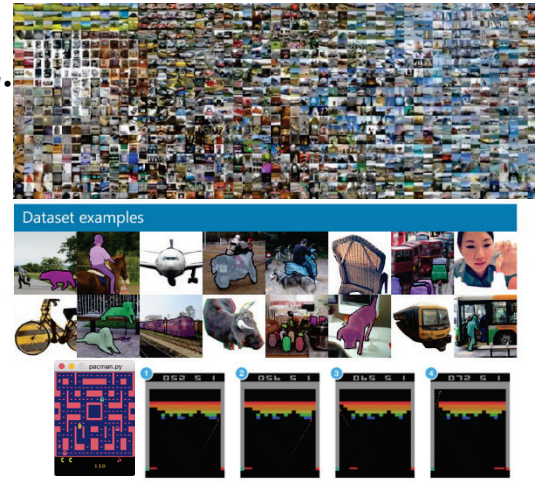
オープンデータ/オープンソース

オープンデータ

公開されているデータが多く，比較がしやすい．

古
- UCIリポジトリ：３５１データセット．回帰，判別，クラスタリングなど．数年前までの定番データセット．
- MNISTデータセット：文字認識データセット．
- Caltech101データセット：101ラベル．一般画像認識．
- 20 Newsgroups：自然言語処理データセット

新
- ImageNet：約1400万枚の画像．毎年コンペ．
- COCOデータセット：セグメンテーションなど．
- Yahoo Flickr Creative Commons 100M：画像+タグ
- OpenAI Gym:強化学習開発プラットフォーム



オープンソース：無料で使える機械学習ライブラリ

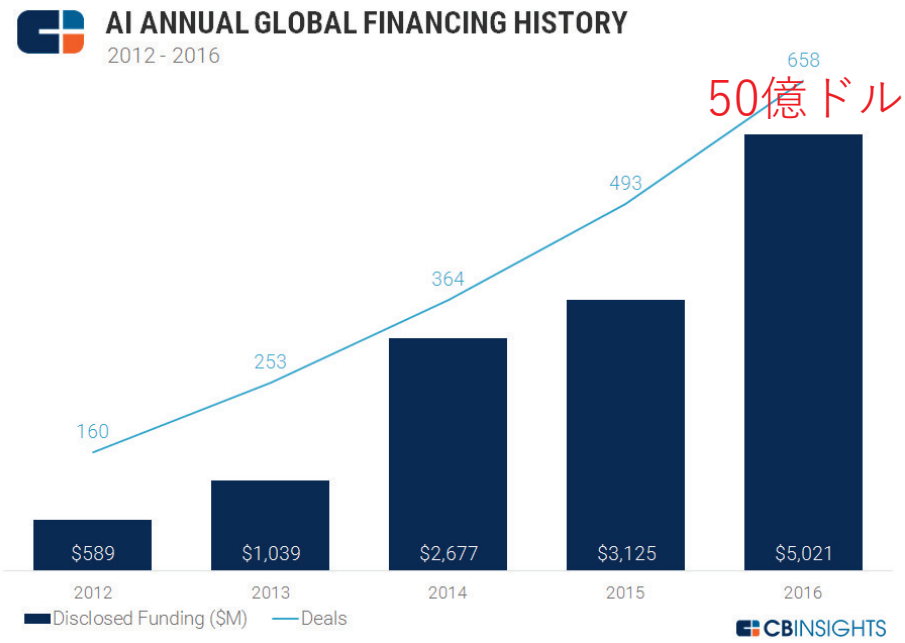
- LibSVM
- scikit-learn for python
- Spark MLlib

産学双方からの貢献

深層学習用ライブラリ

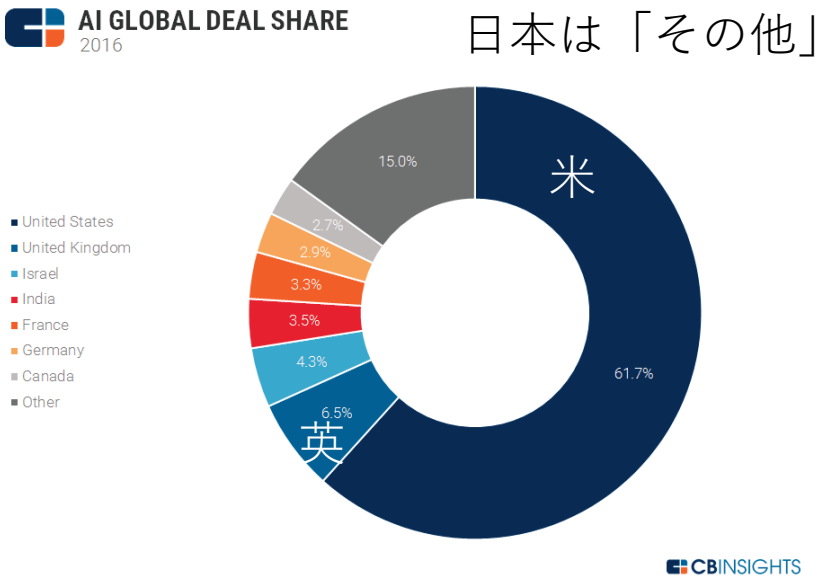
- TensorFlow (Google)
- MXNet
- Caffe (UC Berkley)
- Theano (U. of Montreal)
- Torch (R. Collobert@Facebookら)
- Chainer (Preferred Networks)

人工知能投資額



50億ドル

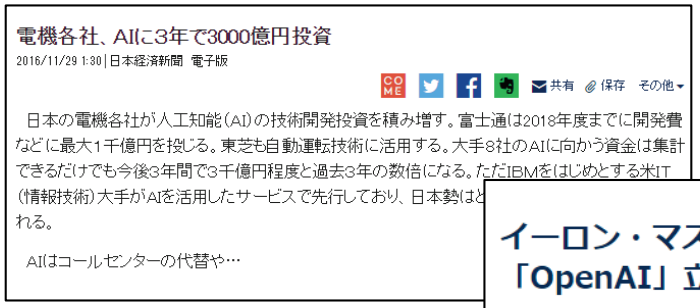
世界のAI投資額



日本は「その他」

国別割合 (2016年)

[CB Insights, “The 2016 AI Recap: Startups See Record High In Deals And Funding ”]



電機各社、AIに3年で3000億円投資

2016/11/29 1:30 | 日本経済新聞 電子版

共有 保存 その他

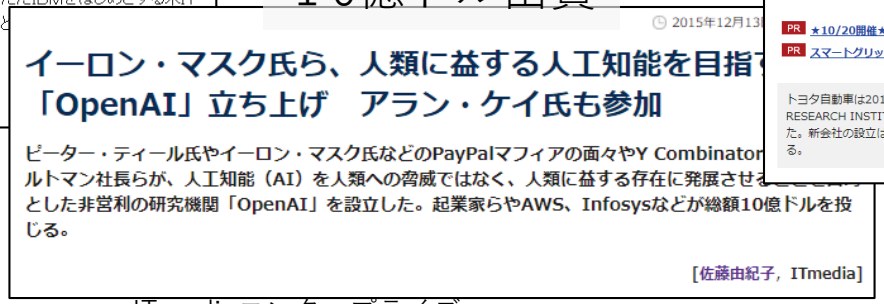
日本の電機各社が人工知能（AI）の技術開発投資を積み増す。富士通は2018年度までに開発費などに最大1千億円を投じる。東芝も自動運転技術に活用する。大手8社のAIに向かう資金は集計できるだけでも今後3年間で3千億円程度と過去3年の数倍になる。ただIBMをはじめとする米IT（情報技術）大手がAIを活用したサービスで先行しており、日本勢はど
れる。

AIはコールセンターの代替や…

日本経済新聞電子版

3年で3000億円

10億ドル出資



2015年12月13

イーロン・マスク氏ら、人類に益する人工知能を目指
「OpenAI」立ち上げ　アラン・ケイ氏も参加

ピーター・ティール氏やイーロン・マスク氏などのPayPalマフィアの面々やY Combinator
ルトマン社長らが、人工知能（AI）を人類への脅威ではなく、人類に益する存在に発展させ
とした非営利の研究機関「OpenAI」を設立した。起業家らやAWS、Infosysなどが総額10億ドルを投
じる。

[佐藤由紀子，ITmedia]

ITmediaエンタープライズ

ビッグデータ　ビッグデータ記事一覧へ

[市場動向]

トヨタ、シリコンバレーにAI技術の研究開発会社を設立へ、5年で1200億円投入

2015年11月6日(金)　河原 潤（IT Leaders編集委員／クラウド&データセンター完全ガイド編集長）

いいね！ シェア ツイート B! Bookmark G+ Pocket

PR ★10/20開催★レッドハットOSS最新事例が多数集結！NTTドコモ・NRI他

PR スマートグリッドからM2M/IoTまで　最新情報をメルマガでお届け！

トヨタ自動車は2015年11月6日、人工知能（AI）技術の研究開発拠点として、新会社TOYOTA RESEARCH INSTITUTE, INC（TRI）を、米カリフォルニア州シリコンバレー地区に設立すると発表した。新会社の設立は2016年1月で、今後5年間で約10億ドル（約1200億円）を投入してこの領域に注力する。

IT Leaders

5年で1200億円

日本の人工知能拠点

三省一体となって事業を推進

総務省
脳情報通信融合研究センター
情報通信研究機構

- 脳情報通信
- 音声認識
- 多言語音声翻訳
- 社会知解析
- 革新的ネットワーク

情報通信技術の統合的なプラットフォームの構築

文部科学省
革新知能統合研究センター
理化学研究所

- 基礎研究
- 革新的な科学技術成果の創出
- 次世代の萌芽的な基盤技術の創出
- 大型計算機資源
- 人材育成

卓越した科学技術研究を活用するためのプラットフォームの構築

経済産業省
人工知能研究センター
産業技術総合研究所

- 応用研究、実用化・社会への適用
- 標準的評価手法等の共通基盤技術の整備
- 標準化
- 大規模目的研究

基礎研究を社会実装につなげるセンター

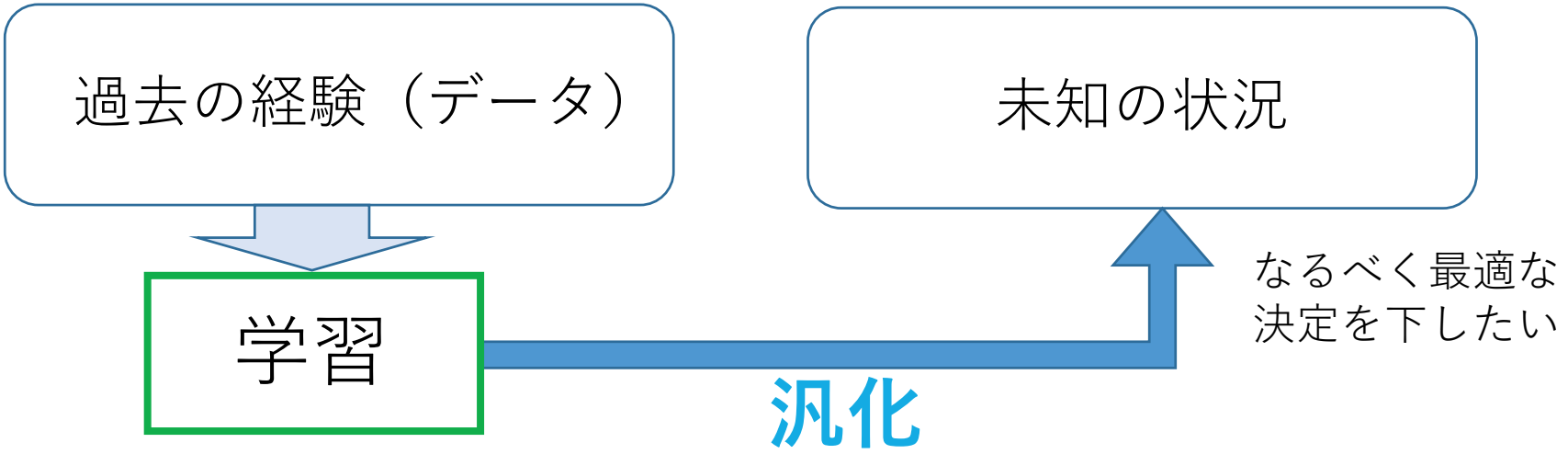
- 機械学習は一つのコア要素
- 人材の育成や基礎研究も重要な課題

第13回科学技術・学術審議会 総合政策特別委員会（H28.6.14）資料１－５より

機械学習の目的

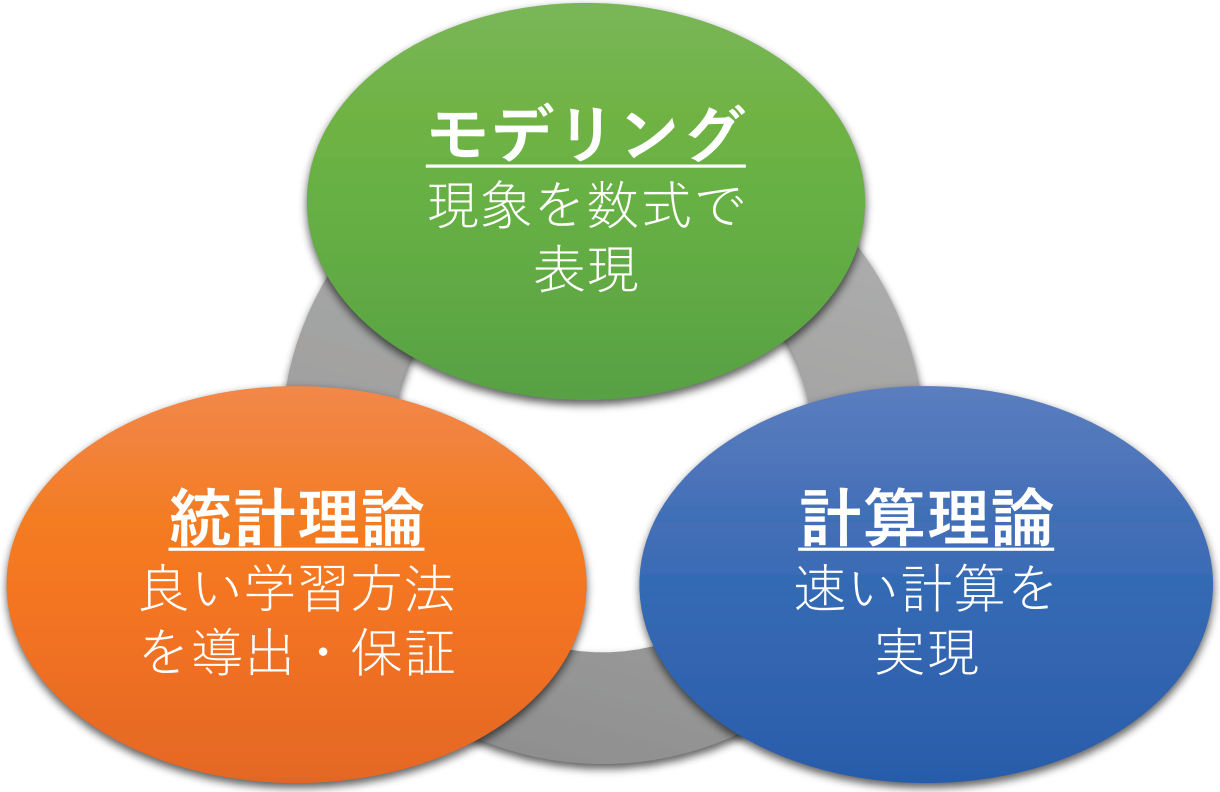
- 人間と同様の知的情報処理を計算機で実現するための技術・手法

Arthur Samuel 「Field of study that gives computers the ability to learn without being explicitly programmed 」（1959）



機械学習の「研究」

ビジネスシーンが変わっても**不変な理論的基盤**



記述言語：数学

- 確率-統計，線形代数，関数解析，最適化理論

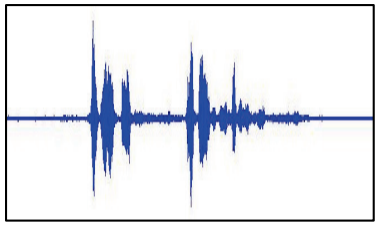
予測と推測

機械学習の活用法

予測

（より機械学習的）

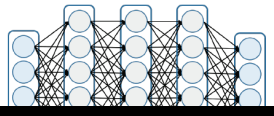
➡ 船



➡ “Hello”

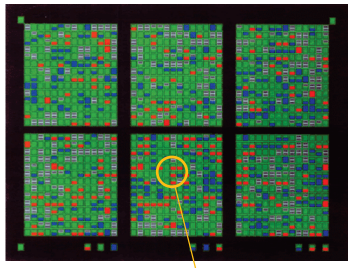
- Outcomeを正しく当てる.
- 解釈よりも予測精度を重視.

例：深層学習



推測

（より数理統計的）

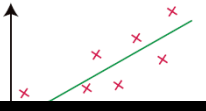


⟷ 肺癌

第〇〇遺伝子が肺癌に寄与
有意水準5%

- 原因の究明.
- 仮説検定は典型例.

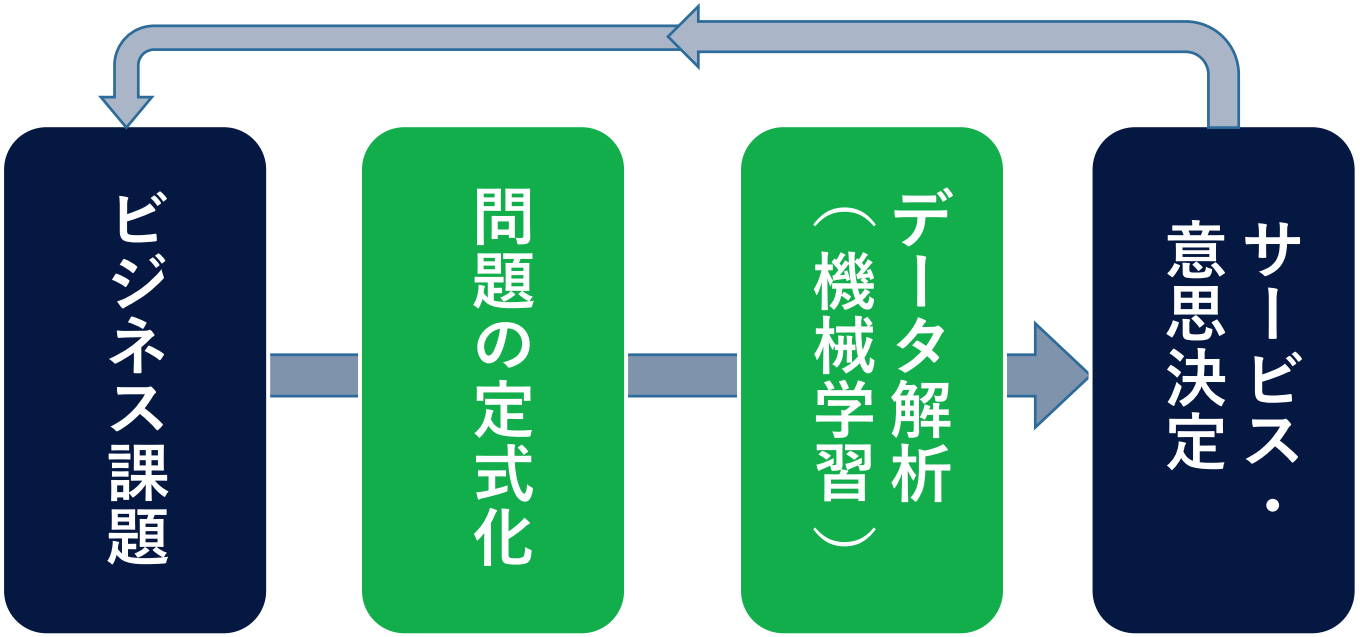
例：線形回帰分析



理論的にはこの二つはある種のトレードオフの関係にある.

機械学習のビジネス利活用

- 目的に応じて手法を選択する必要



Q:収益を上げたい.（予測？推測？）

- 収益を正確に予測？→予測精度が良くてもそれ自体は意味がない.
- どうすれば収益を上げられるか，そのファクターを見つけたい.

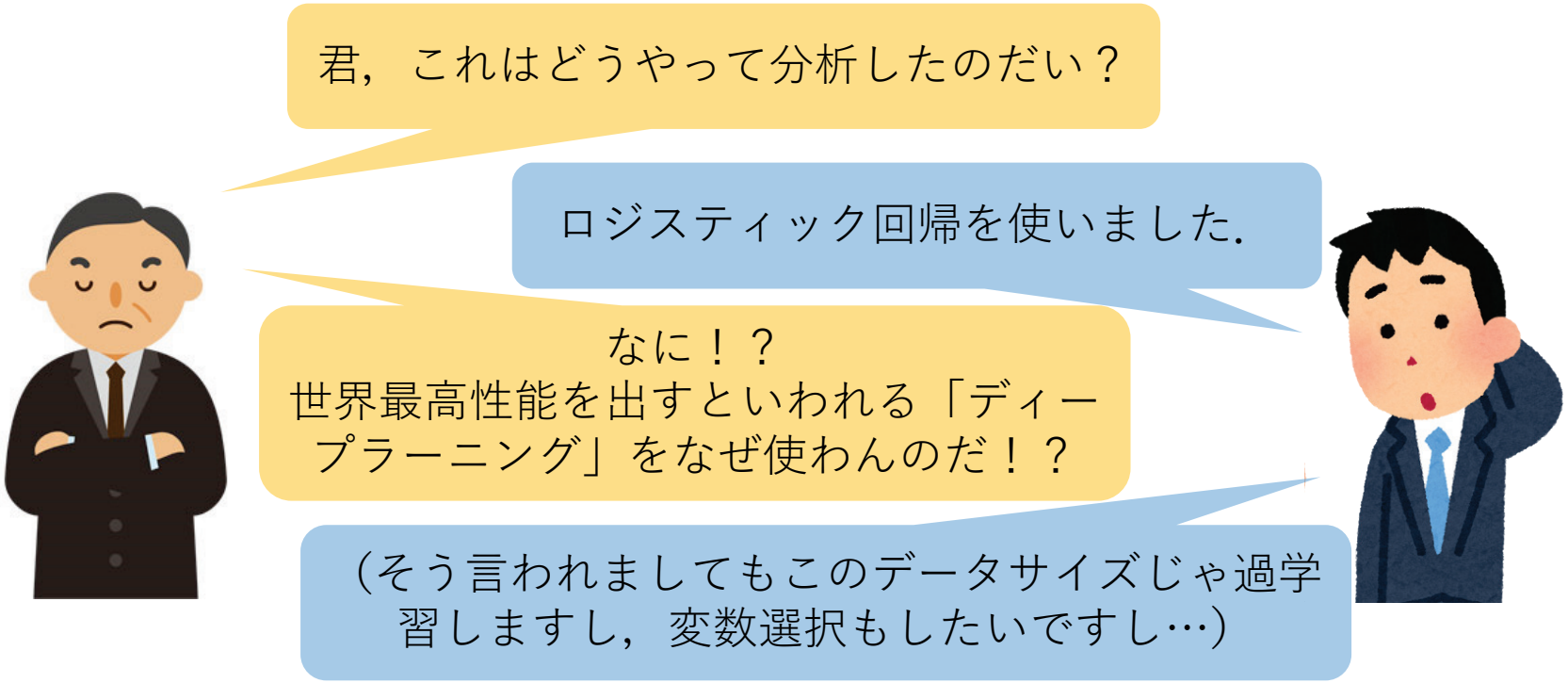
各種機械学習手法で何ができるのか？
→仕組み/理論を把握する重要性

機械学習の「難しさ」

大が小を兼ねるわけではない.

- 目的に応じた手法の選択.
  (予測 vs 推測)
- **過学習の問題.** (後述)

データの種類（画像，テキスト，音声,...），データサイズ，教師あり/なし，判別/回帰，予測/推測，高次元/低次元，ベクトル/行列，...



機械学習の基本事項

機械学習と人工知能の歴史

1946： ENIAC，高い計算能力
フォン・ノイマン「俺の次に頭の良い奴ができた」
1952： A. Samuelによるチェッカーズプログラム

統計的学習

ルールベース

1960年代前半：
ELIZA（イライザ），
擬似心理療法士

1980年代：
エキスパートシステム

人手による学習ルールの作りこみの限界
「膨大な数の例外」

Siriなどにつながる

1957：Perceptron，ニューラルネットワークの先駆け
第一次ニューラルネットワークブーム
1963：線形サポートベクトルマシン

線形モデルの限界

1980年代：多層パーセプトロン，誤差逆伝搬，
畳み込みネット
第二次ニューラルネットワークブーム
1992： 非線形サポートベクトルマシン
（カーネル法）

非凸性の問題

1996： スパース学習（Lasso）
2003： トピックモデル（LDA）
2012： Supervision（Alex-net）
第三次ニューラルネットワークブーム

データの増加
＋計算機の強化

人間

四則演算

単純なルール

機械

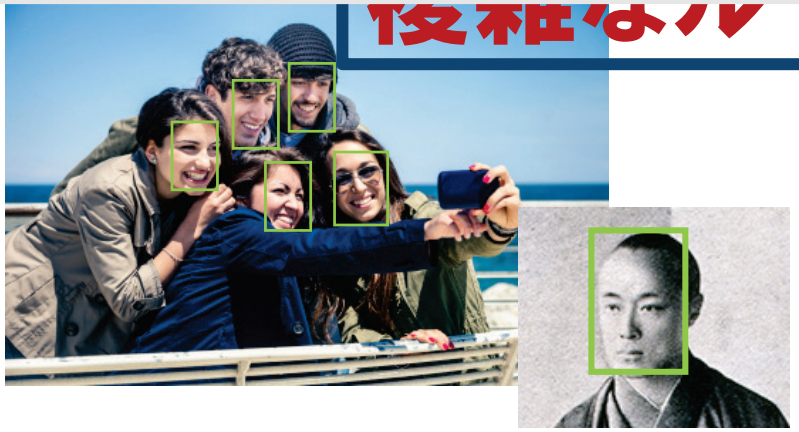
難

$193707721 \times 761838257287 - 2^{67} = -1$

易

人の手でプログラムするのは無理

learn without being explicitly programmed



易

難

人によって顔が違う，照明の当たり方で見え方・色が変わる，
表情の違い，髪型の違い，顔の向きの違い，．．．．

統計的学習の考え方

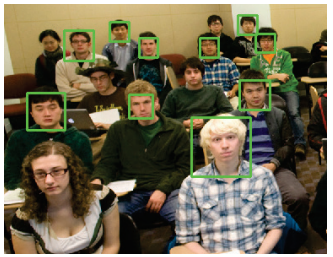
- 人がプログラムするのは認識の仕方ではなく学習の仕方

→数学が必要

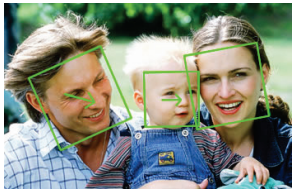
大量画像データ

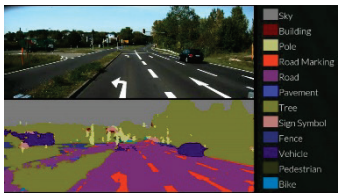
学習

初見の画像も
正しく認識
(汎化)



- 強い将棋ソフトを作りたい → 大量の棋譜データで学習
- 顔認識ソフトを作りたい → 大量の画像データで学習
- 車道を認識したい → 大量の車載カメラ画像で学習



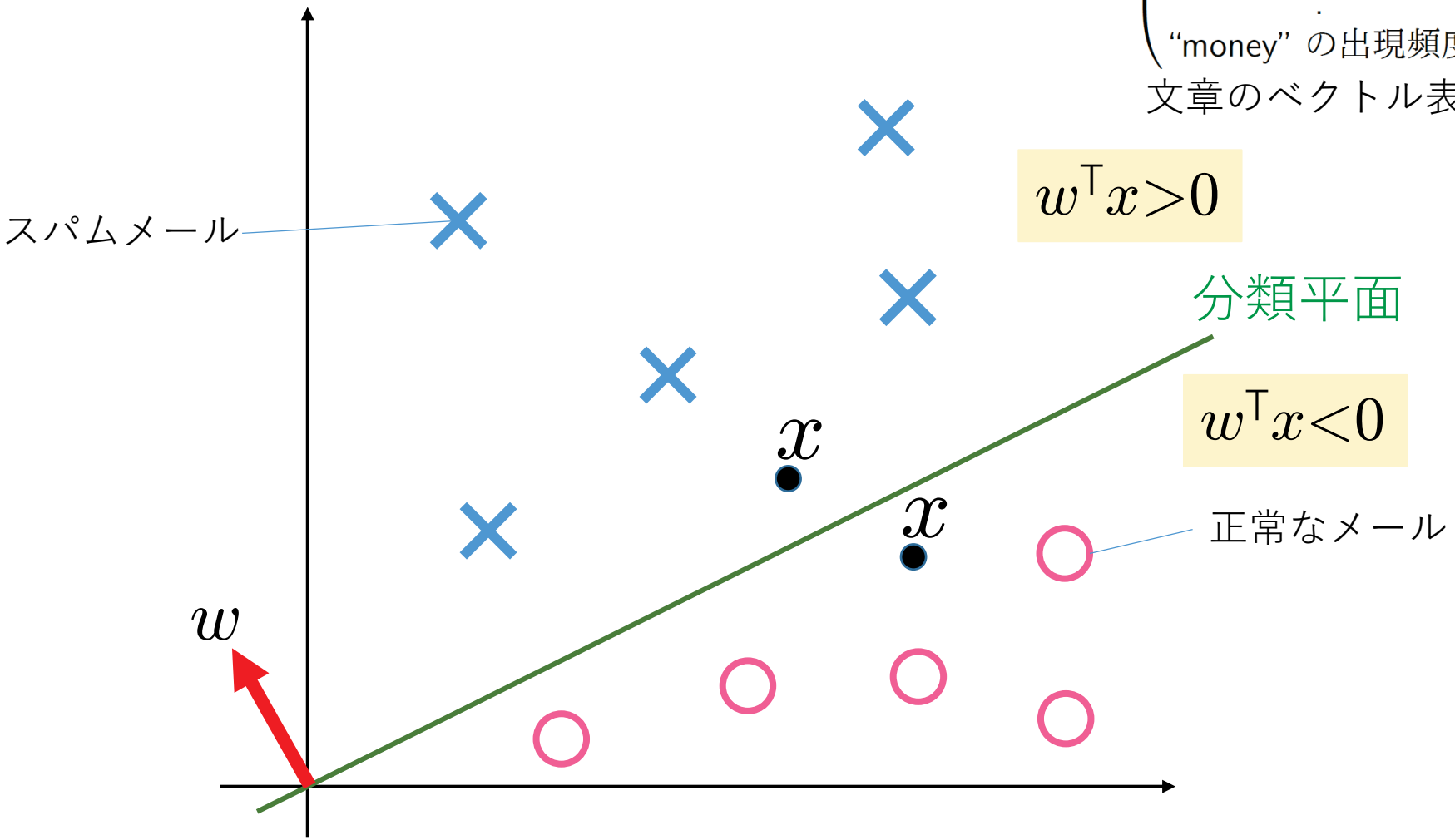


線形分類機

Bag-of-words

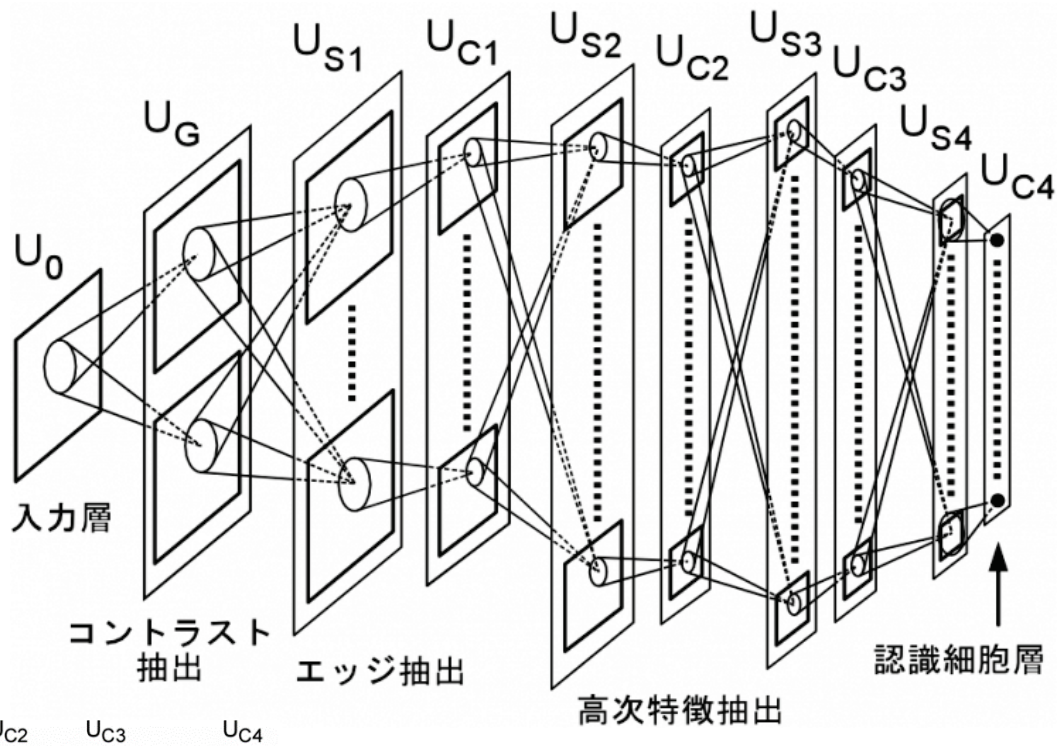
$$x = \begin{pmatrix} \text{“please” の出現頻度} \\ \text{“credit” の出現頻度} \\ \vdots \\ \text{“money” の出現頻度} \end{pmatrix}$$

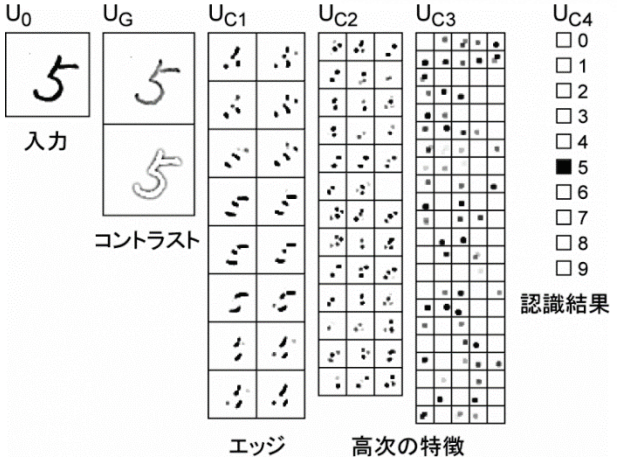
文章のベクトル表現例

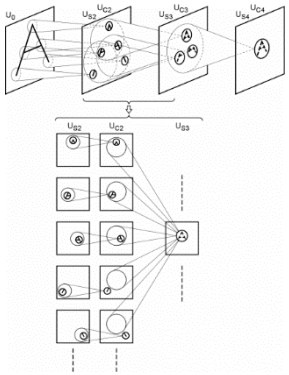


ネオコグニトロン

[福島,79]



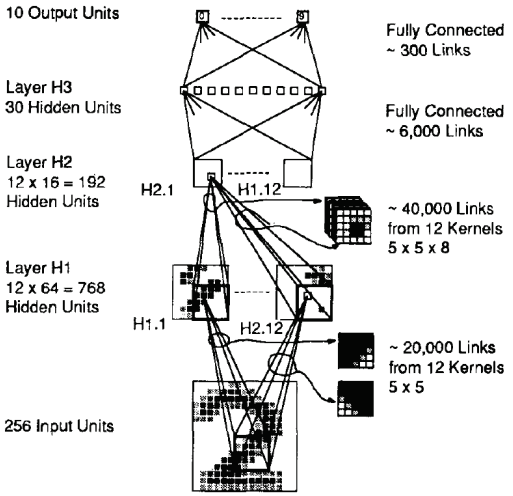




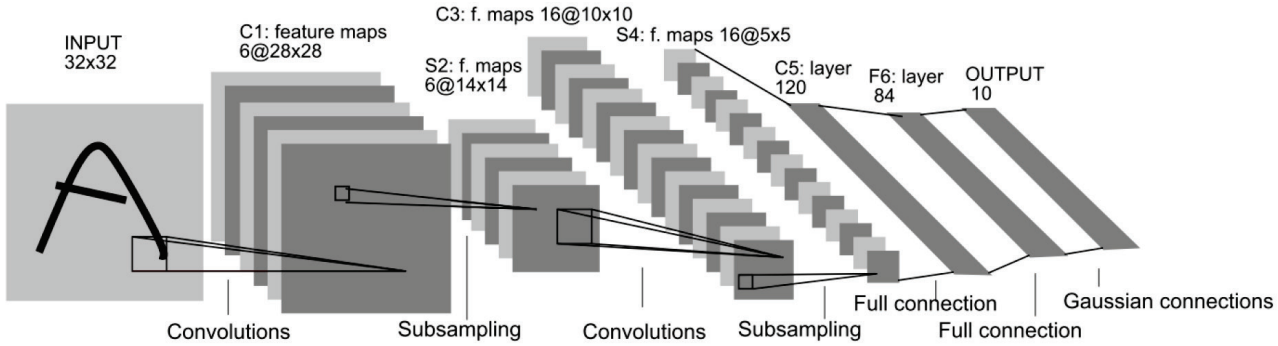
- 人間の脳を模倣
- 畳み込みネットの初期型
- 自己組織型学習
  →素子を足してゆく

LeNet

[LeCun+etal,89]



LeNet-5
[LeCun etal,98]



- 畳み込み＋プーリング：現在も使われている構造
- 誤差逆伝搬法でパラメータを更新
- 手書き文字認識データセット（MNIST）で９９％の精度を達成

機械学習と人工知能の歴史

1946： ENIAC，高い計算能力
フォン・ノイマン「俺の次に頭の良い奴ができた」
1952： A. Samuelによるチェッカーズプログラム

統計的学習

ルールベース

1960年代前半：
ELIZA（イライザ），
擬似心理療法士

1980年代：
エキスパートシステム

人手による学習ルールの作りこみの限界
「膨大な数の例外」

Siriなどにつながる

1957：Perceptron，ニューラルネットワークの先駆け
第一次ニューラルネットワークブーム
1963：線形サポートベクトルマシン

線形モデルの限界

1980年代：多層パーセプトロン，誤差逆伝搬，
畳み込みネット
第二次ニューラルネットワークブーム
1992： 非線形サポートベクトルマシン
（カーネル法）

非凸性の問題

1996： スパース学習（Lasso）
2003： トピックモデル（LDA）
2012： Supervision（Alex-net）
第三次ニューラルネットワークブーム

データの増加
＋計算機の強化

問題点

- 誰でも実装できるわけではなかった.

  e.g. 「LeNetはYan LeCunしか実装できない」といった噂
- 様々な職人芸的なノウハウが存在.
  - パラメータのチューニング：学習率，層の数，層の幅
- 大域的最適解の計算が難しい.

  局所最適解しか得られない.

→これらは現代でも未解決

(実装に関してはライブラリの充実でかなり解決)

誰でも実装できて，最適解が一つの手法が欲しい.

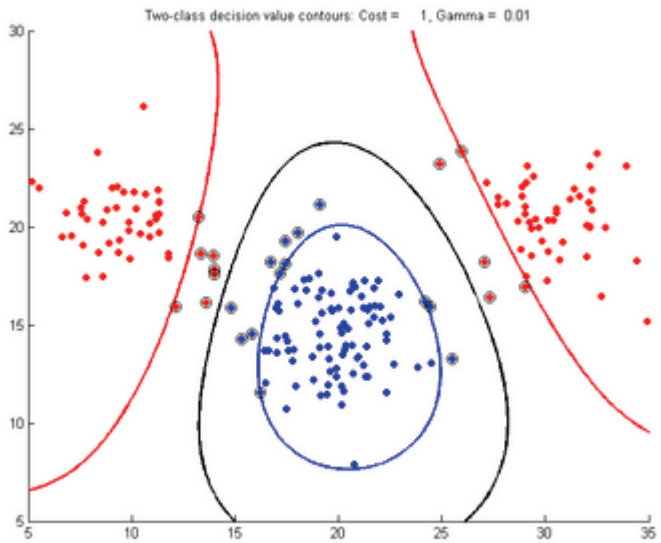
→　カーネルを用いたサポートベクトルマシン

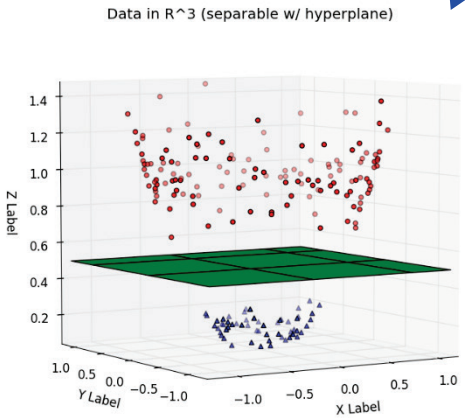
カーネルを用いたSVM

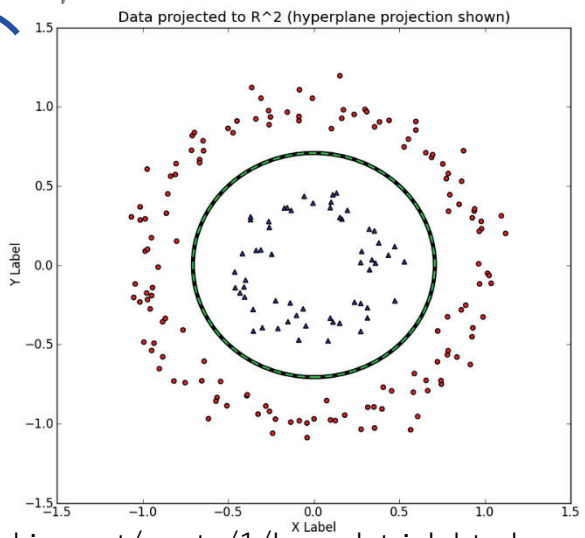
カーネルトリック

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

非線形写像 ϕ





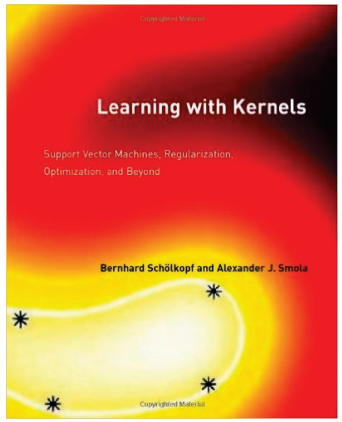


http://wiki.eigenvector.com/index.php?title=Svmda

http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html

$$\min_{\alpha_i, b} \sum_{i=1}^{n} \max\left\{1 - y_i\left(\sum_{j=1}^{n} k(x_j, x_i)\alpha_j + b\right), 0\right\} + C \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$

- **凸最適化問題**で解ける.
  - ✓ 効率的な最適化手法が存在.
  - ✓ 解は一つ. 誰が解いても同じ答えが返ってくる.
- **VC理論**による汎化誤差の保証.



９０年代以降

データ解析としての機械学習

統計学やデータマイニングとの融合

- 高次元スパース学習
- ベイズモデリング
- オンライン学習，確率的最適化

　→　ビッグデータ解析への活用

メディアに出る「人工知能」はここに属することが多い

- データの増加＋計算機の強化

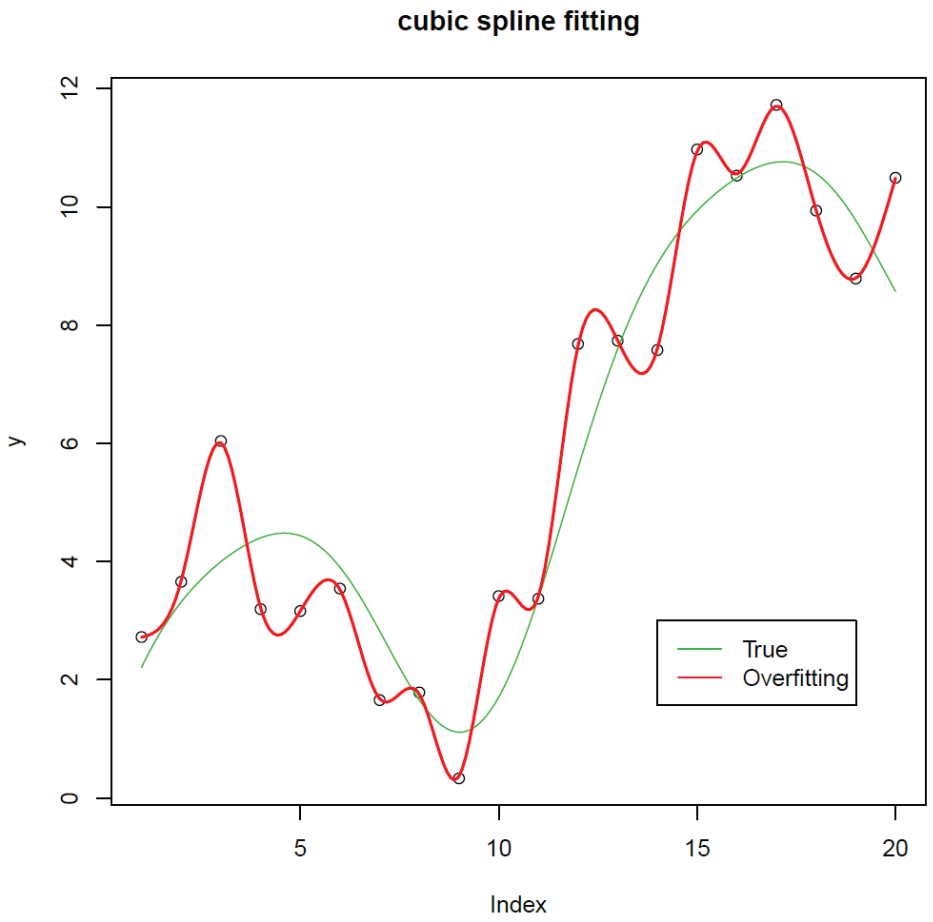
　→深層学習再興，**第三次ニューラルネットワークブーム**へ

機械学習の数理
-過学習とモデルの複雑さ-

過学習

複雑なモデル（例えば深層ニューラルネット）を用いるのが常に良い選択か？

→ そうとは限らない．「**過学習**」に注意する必要あり．



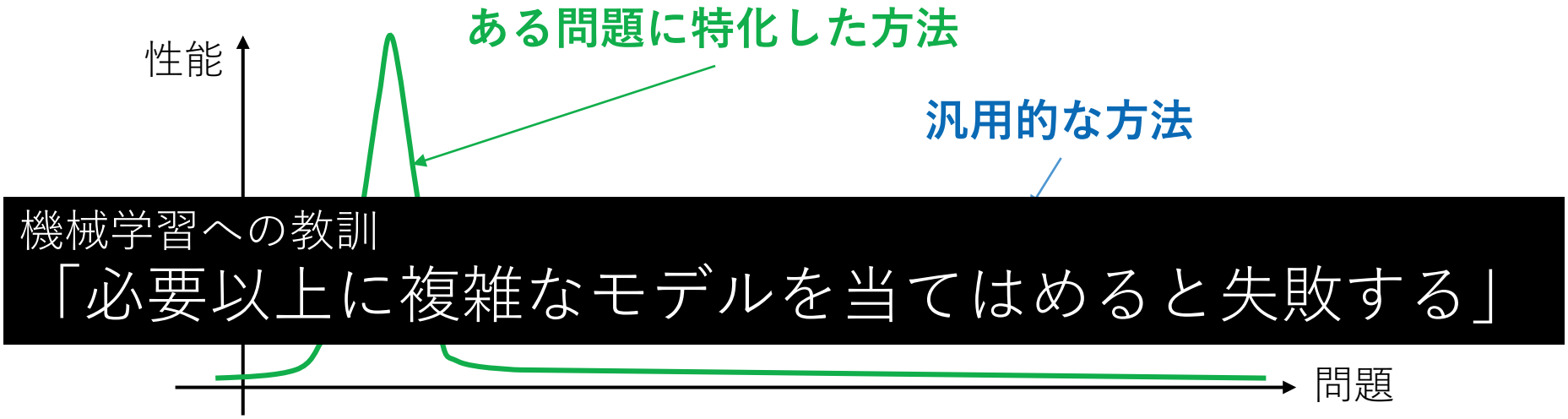
学習機の複雑さと学習能力

- オッカムの剃刀

「ある事柄を説明するためには、必要以上に多くを仮定するべきでない」とする指針

- No free lunch theorem

「あらゆる問題で性能の良い汎用的学習機は実現不可能であり，ある問題に特殊化された手法に勝てない」



William of Ockham：1285-1347．スコラ学の神学者，哲学者．

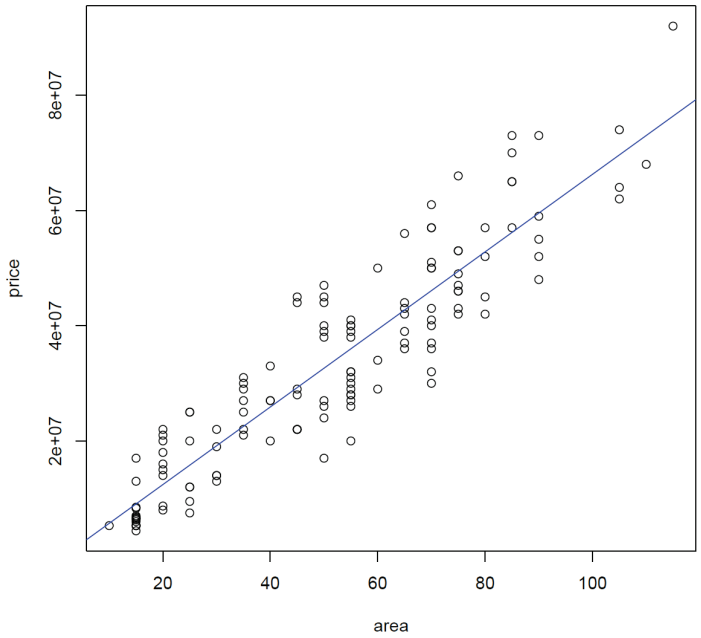
No free lunch theorem: [D.H.Wolpert: 1996]
[D.H.Wolpert and W.G. Macready: 1995,1997][Y.C. Ho and D.L. Pepyne: 2002]

線形モデル

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \beta_0 + \epsilon$$

y:従属変数, x:特徴ベクトル

マンション価格=$\beta_1 \times$ 床面積 $+ \beta_2 \times$ 築年数 $+ \beta_3$+(揺らぎ)



正則化学習法

正則化：データに合った単純なモデルを当てはめる

→ 過学習を回避

正則化訓練誤差最小化

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 + \lambda \psi(\beta)$$

手元にあるデータへの当てはまり　　正則化項

複雑さへの罰則

代表的な例：リッジ正則化（L2ノルム）

$$\psi(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{d} \theta_j^2$$

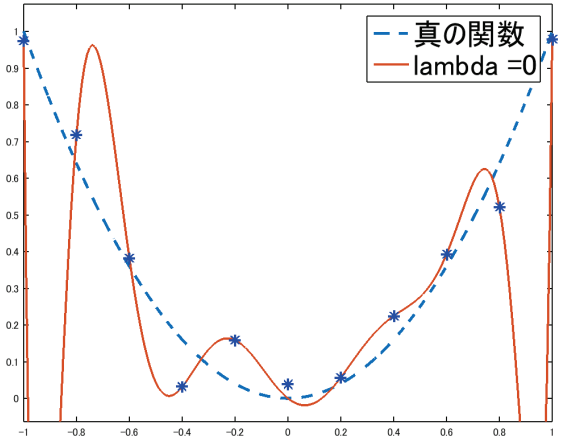
正則化の例

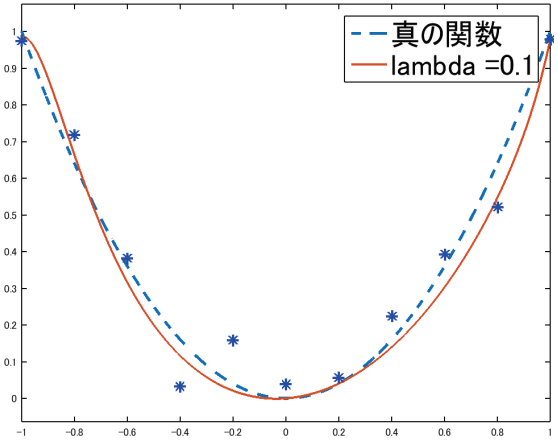
多項式回帰（１５次多項式，リッジ回帰）

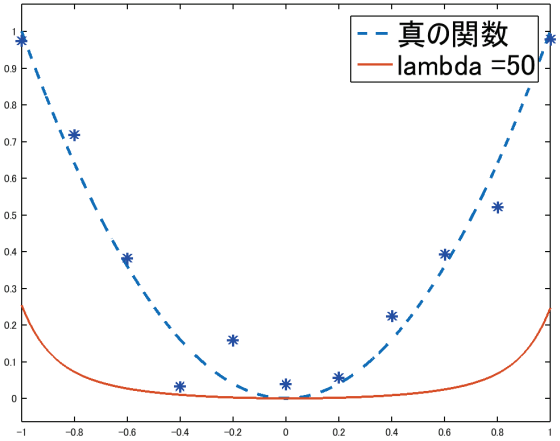
$$\min_{\beta \in \mathbb{R}^{15}} \sum_{i=1}^{n} \{y_i - (\beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{15} x_i^{15})\}^2 + \lambda \|\beta\|_2^2$$

リッジ正則化と言う

手元のデータには良くあてはまるが真の関数からは遠い







$\lambda = 0$　　$\lambda = 0.1$　　$\lambda = 50$

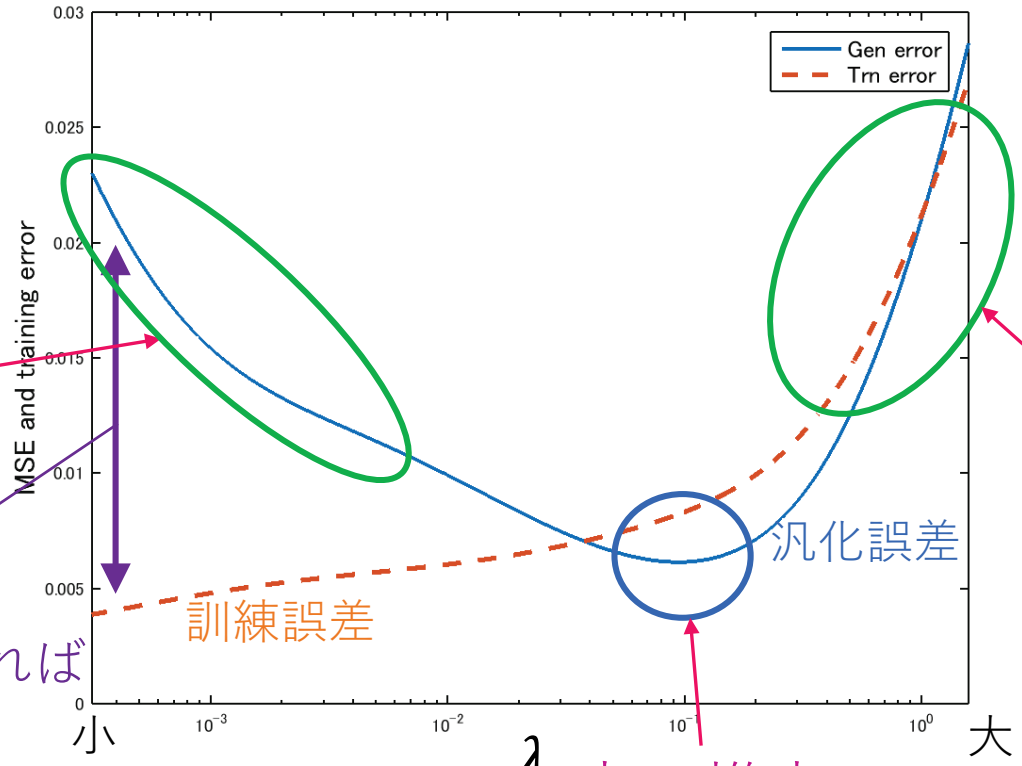
過学習　　良い推定　　過小学習

正則化によってあまり複雑にならないよう制御がかかる

正則化の強さと汎化誤差の関係

多項式回帰（１５次多項式，リッジ回帰）

$$\min_{\beta \in \mathbb{R}^{15}} \sum_{i=1}^{n} \{y_i - (\beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_{15} x_i^{15})\}^2 + \lambda \|\beta\|_2^2$$



過学習

訓練誤差が小さければ
よいわけではない

訓練誤差

汎化誤差

過小学習

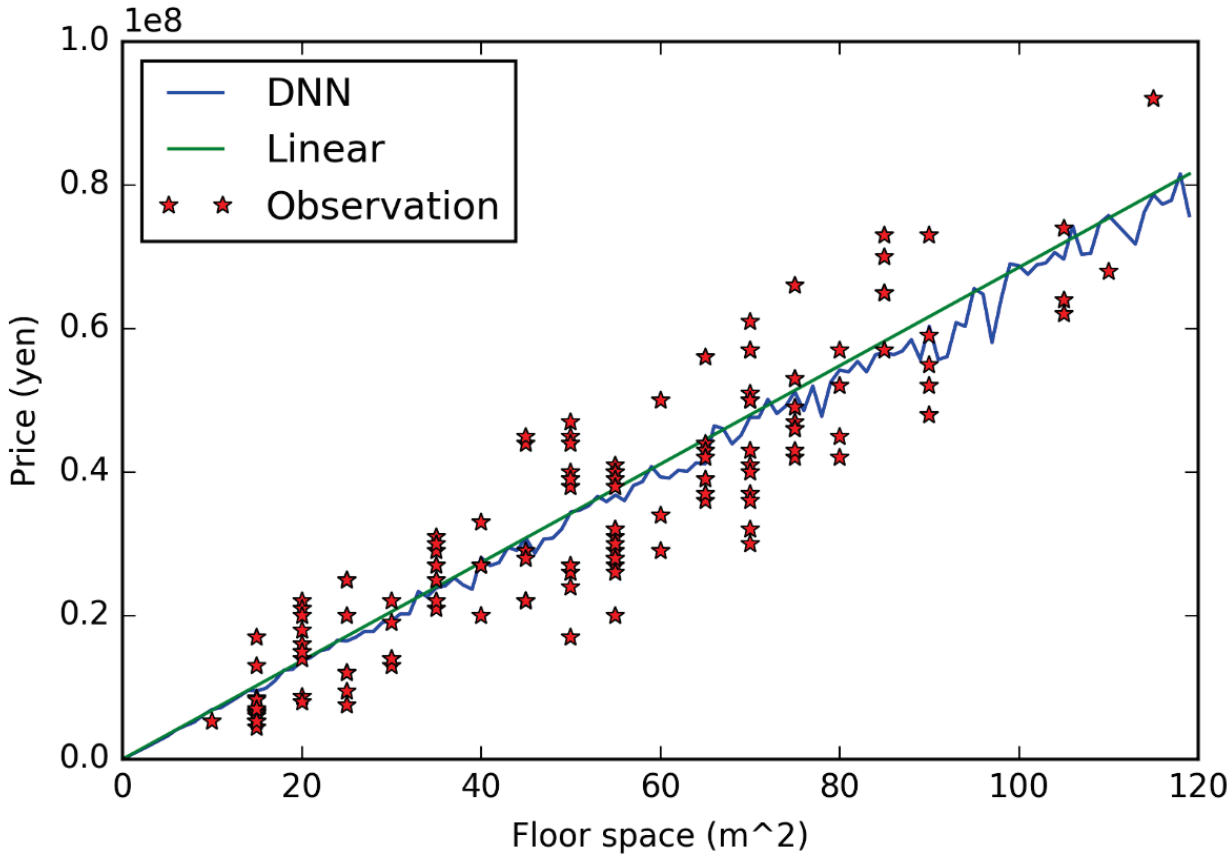
良い推定

横軸：正則化パラメータ(log-スケール)．　縦軸：汎化誤差（青），訓練誤差（赤）．

適切な λ を選ぶ方法→交差検証法，Mallows' Cp

線形モデル vs 深層学習

過学習の例



深層学習を使うには簡単&データが少なすぎる

マンションの価格推定

DNN: 中間層２層横幅100の深層NN, **Linear**: 単回帰モデル

汎化誤差（平均二乗誤差）： DNN: 1.30×10^{15}, Linear: 6.26×10^{13}

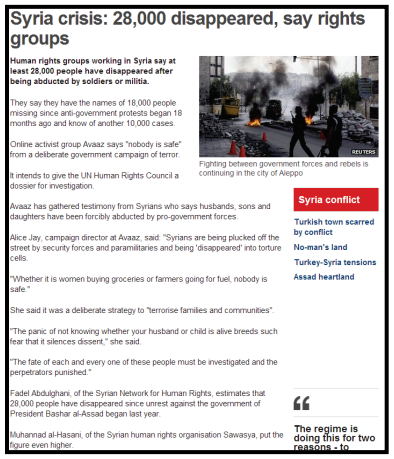
一概に何でもかんでも深層学習が良いとは言えない

高次元データ

インターネットや計測機器の発達により多様なデータが取得可能

多くの場合で高次元

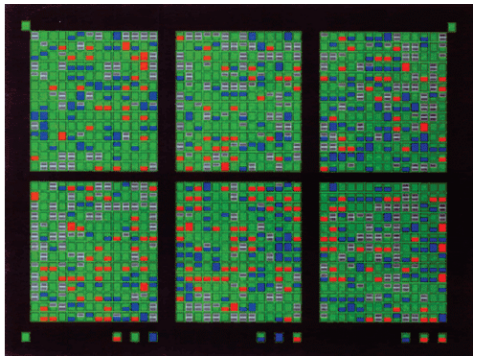
- 遺伝子データ
- テキストデータ
- マーケティングデータ
- 金融データ



Bag of words

数百万次元

Syria	13
people	5
bomb	7
economy	1
immigrants	2
soccer	0
walk	1

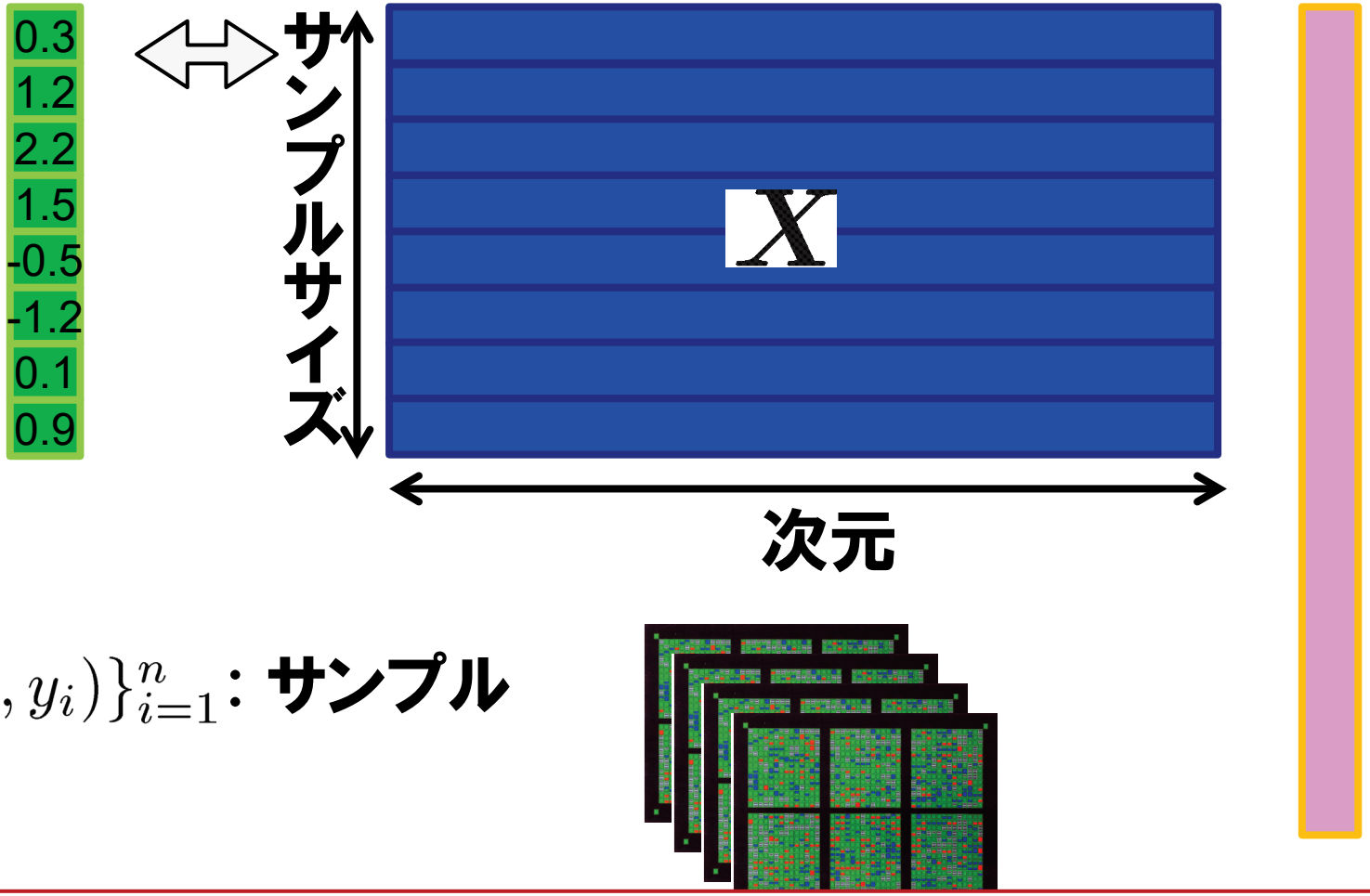


遺伝子発現量

数万次元

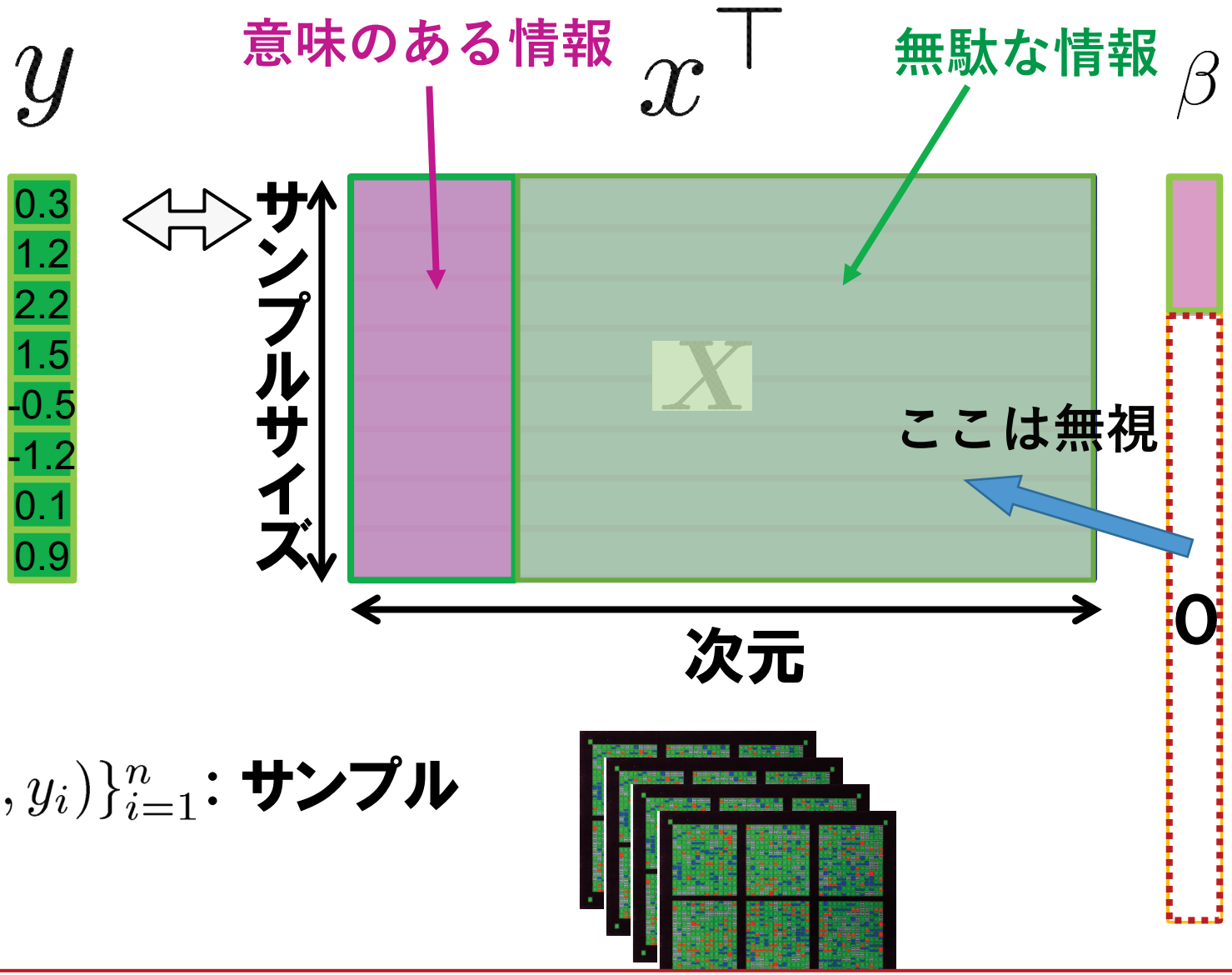
0.5
2.4
4.2
0.2
1.3
0.1
5.3

y $x^\top$ β



$\{(x_i, y_i)\}_{i=1}^n$: サンプル

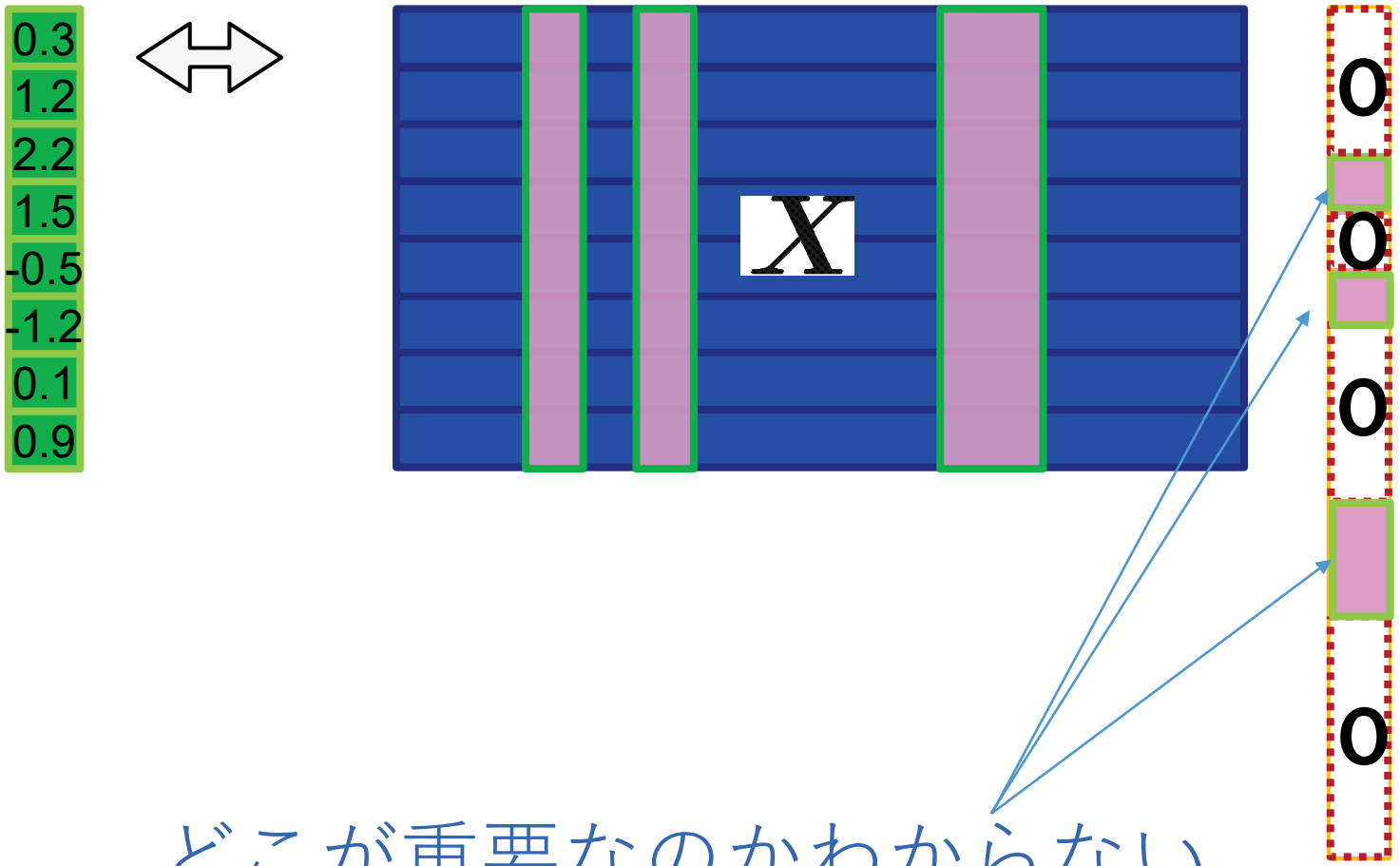
次元 > サンプルサイズ → 余分な情報を落としたい



$\{(x_i, y_i)\}_{i=1}^n$: サンプル

次元 > サンプルサイズ → 余分な情報を落としたい

スパースモデリング



どこが重要なのかわからない

→ 特徴選択：データから学習

予測に寄与する特徴量を特定できれば解釈性も上がる

AICによる特徴選択（組み合わせ的方法）

AIC: 赤池情報量規準 → 最尤推定量の予測誤差の不偏推定量

AIC最小化

$$\hat{\beta}_{\mathrm{AIC}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \|Y - X\beta\|^2 + 2\sigma^2 \|\beta\|_0$$

データへの当てはまり

次元に対する罰則（正則化）

ただし$\|\beta\|_0 = \beta$の非ゼロ要素の個数 ：L_0ノルムと言う.

- 予測誤差を近似的に最小化
- 変数の組み合わせの数： 2^p個の候補（膨大）
- NP困難

線形モデルを仮定

$$Y = X\beta^* + \xi$$

サンプルサイズn, 次元p

観測ノイズ：分散σ^2の正規分布

LASSOによる特徴選択（凸最適化）

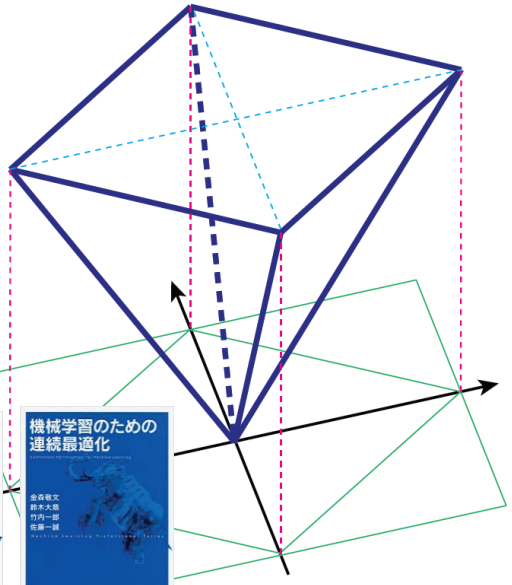
Lasso [L_1正則化] (R. Tsibshirani (1996))

$$\hat{\beta}_{\text{Lasso}} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \underbrace{\|Y - X\beta\|^2}_{\text{データへの当てはまり}} + \underbrace{\lambda \|\beta\|_1}_{\text{次元に対する罰則（正則化）}}$$

ただし $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ ：L_1ノルムと言う.

Lassoは**凸最適化**と呼ばれる問題のクラス

- **高速に解ける（近接勾配法等）**
- L_1ノルムはL_0ノルムを最も良く近似する凸関数
- パラメータλはクロスバリデーションで選べば良い.
- 理論が豊富.



書籍：確率的最適化，機械学習のための連続最適化

Lassoのスパース性

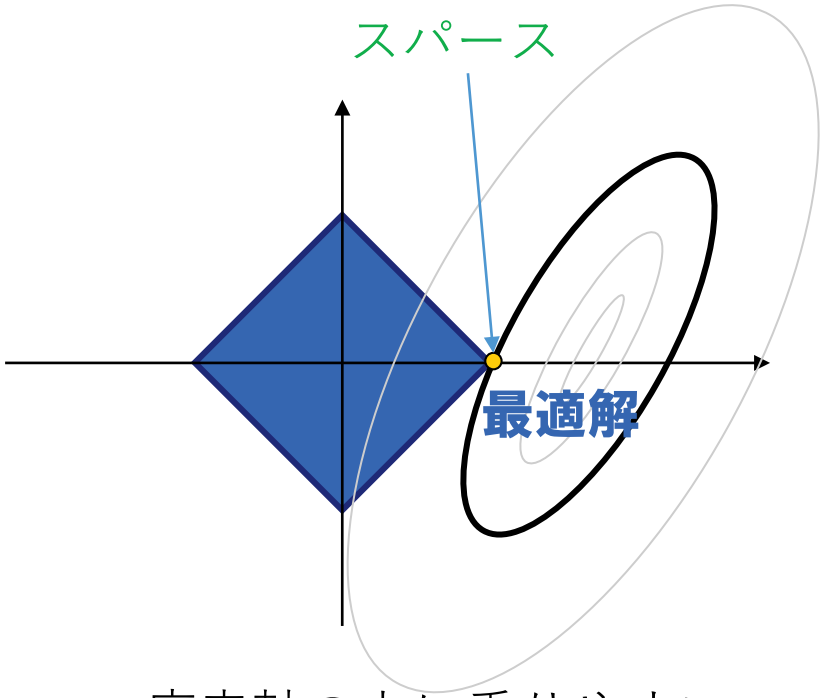
$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \quad \Leftrightarrow \quad \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|^2 \text{ s.t. } \sum_{j=1}^{p} |\beta_j| \leq C$$

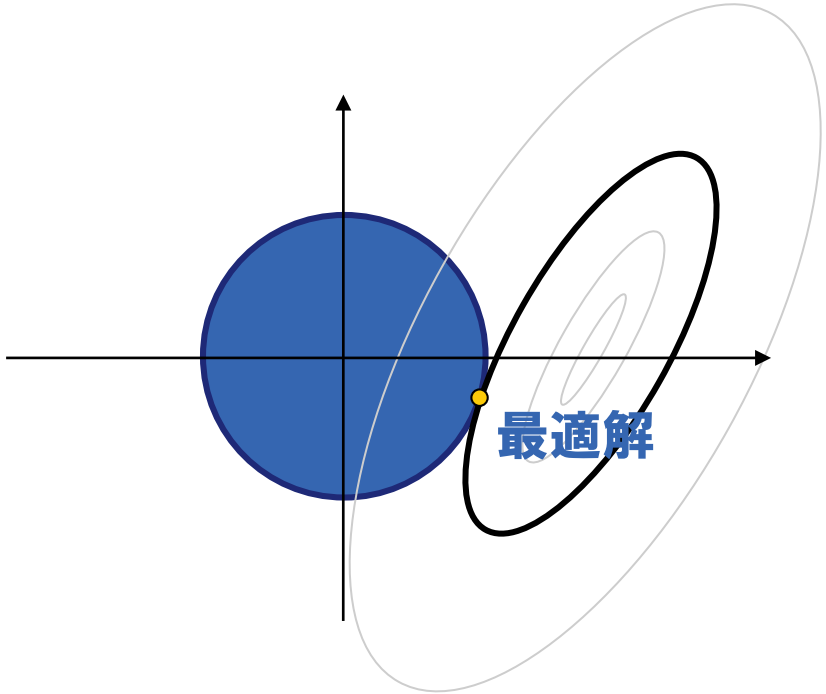
L1正則化

$$\|\beta\|_1 = |\beta_1| + \cdots + |\beta_p|$$

L２正則化（リッジ正則化）

$$\|\beta\|_2^2 = \beta_1^2 + \cdots + \beta_p^2$$





座表軸の上に乗りやすい

スパース推定によって予測に必要な変数が自動的に選ばれる

スパース性の恩恵

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \|X\beta - Y\|_2^2 + \lambda_n \sum_{j=1}^{p} |\beta_j|$$

d = 真のベクトルβ^*の非ゼロ要素の数 (予測に寄与する変数の数)

定理 (Lassoの収束レート (Bickel et al., 2009; Zhang, 2009))

ある条件のもと（制限等長性など），ある定数Cが存在して,

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \frac{d \log(p)}{n}$$

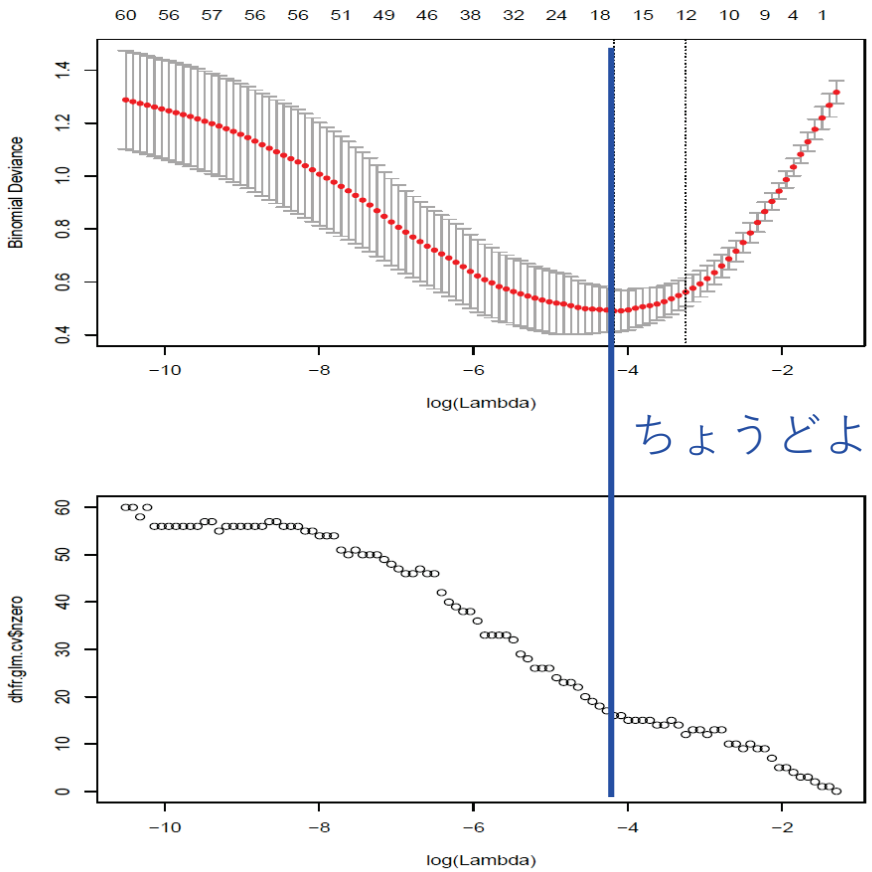
- 全体の次元p はたかだか$O(\log(p))$でしか影響しない!
- 実質的次元d が支配的.
- 高次元スパースな問題を精度よく解くことができる.

} **過学習を防止**

推定誤差 $\frac{d \log(p)}{n} \ll \frac{p}{n}$ (最小二乗法)
過学習してしまう

低次元性（スパース性）をうまく利用できている.

ジヒドロ葉酸レダクターゼデータにおける実験



CVスコア
(予測精度)

ちょうどよい正則化

非ゼロ要素の個数

スパース性と汎化誤差

横軸：正則化パラメータ．縦軸：(上段) CV スコア,(下段) 非ゼロ要素個数

低ランク行列補完

ベクトルから**行列**の学習へ

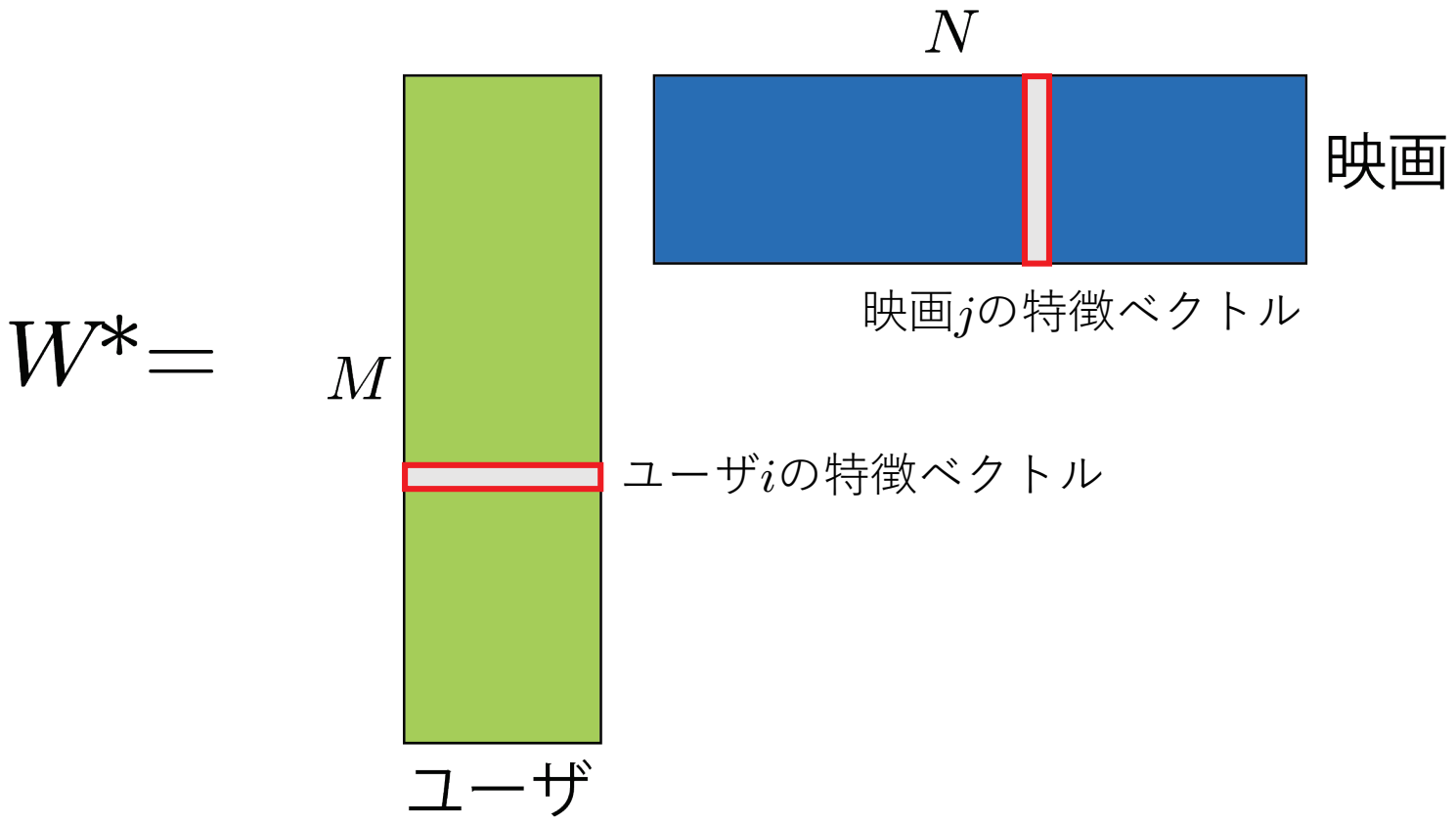
- 推薦システム

	映画 A	映画 B	映画 C	・・・	映画 X
ユーザ 1	4	8	4	・・・	2
ユーザ 2	2	4	2	・・・	1
ユーザ 3	2	4	2	・・・	1
⋮					

ランク 1 と仮定

各ユーザーが各映画をどれだけ好むかという部分的情報がある.
→　残りの部分（*の部分）を埋めたい.
　　低ランク行列補完で可能.

e.g., Netflix prize (100万ドルの賞金,　48万ユーザ×１万8千映画)

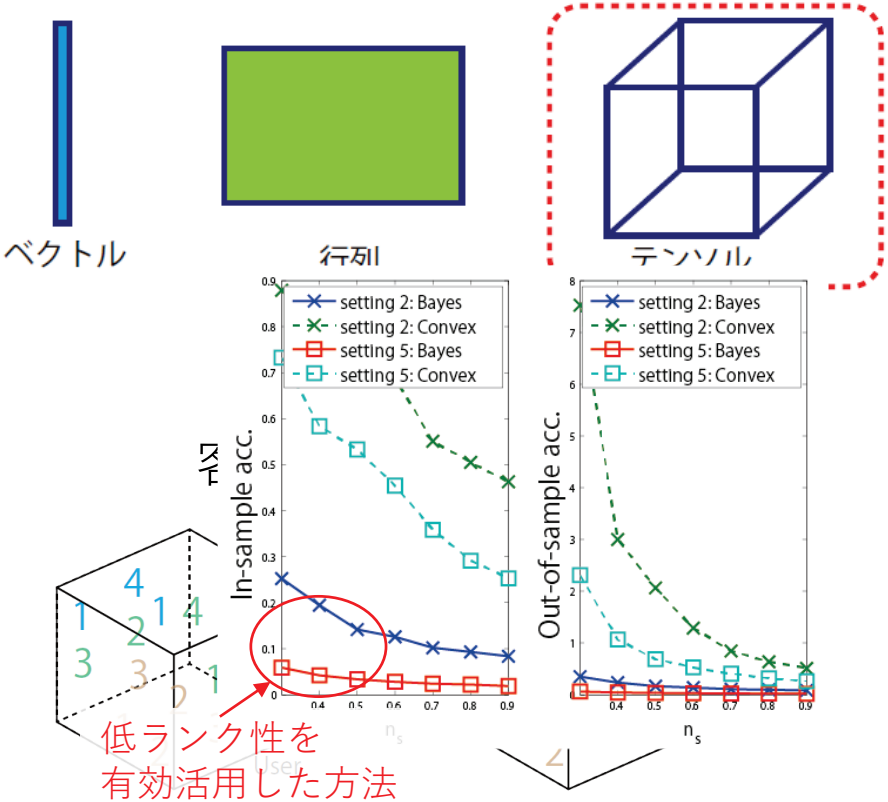


低ランク行列の学習は「ユーザ」と「映画」の低次元表現を学習することに他ならない.
→交互最適化法やトレースノルム正則化法で学習可能

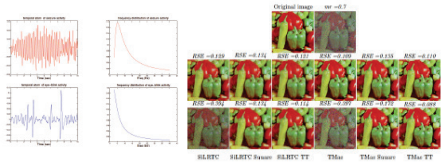
推定誤差 $O\left(\frac{r(M+N)}{n}\right) \ll O\left(\frac{MN}{n}\right)$ (低ランク性を利用しない最小二乗法)

r:ランク

テンソルの学習

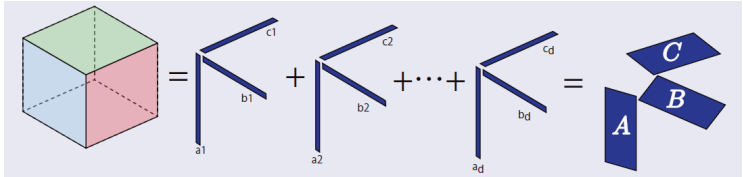


他の応用例：
- 時空間データ解析
- 画像処理
- 自然言語処理
- 深層学習



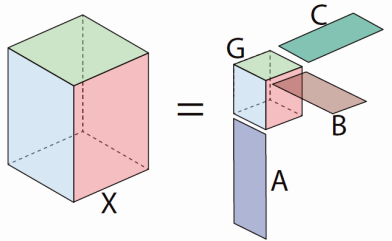
テンソルの“ランク”

CP-分解/ランク



Canonical Polyadic 分解(Hitchcock, 1927; Hitchcock, 1927)
CANDECOMP/PARAFAC (Carroll & Chang, 1970; Harshman, 1970)

Tucker-分解/ランク

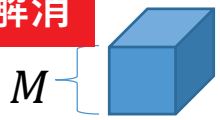


(Tucker, 1966)

通常(最小二乗法)　　低ランク性を利用した方法
(ベイズ推定法等)

$$M^K/n \Rightarrow dKM/n$$

次元の呪いを解消

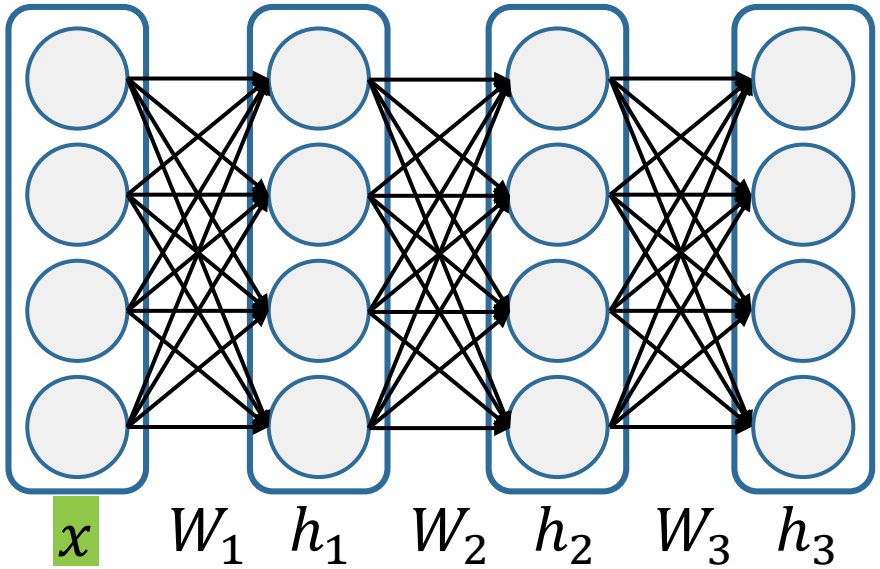


K:次元
d:ランク($\ll M$)

[Suzuki,ICML2015;Kanagawa+et al.,ICML2016;Suzuki+et al.,NIPS2016]

Deep Learning
深層学習

深層学習の構造



基本的に「線形変換」と「非線形活性化関数」の繰り返し．

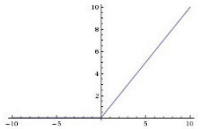
x → W_1x → $h_1(W_1x)$ → $W_2h_1(W_1x)$ → $h_2(W_2h_1(W_1x))$

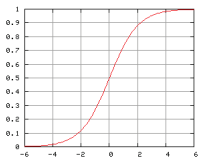
入力　線形変換　非線形変換　線形変換　非線形変換

$$h_1(u) = [h_{11}(u_1), h_{12}(u_2), \ldots, h_{1d}(u_d)]^T$$

活性化関数は通常要素ごとにかかる．Poolingのように要素ごとでない非線形変換もある．

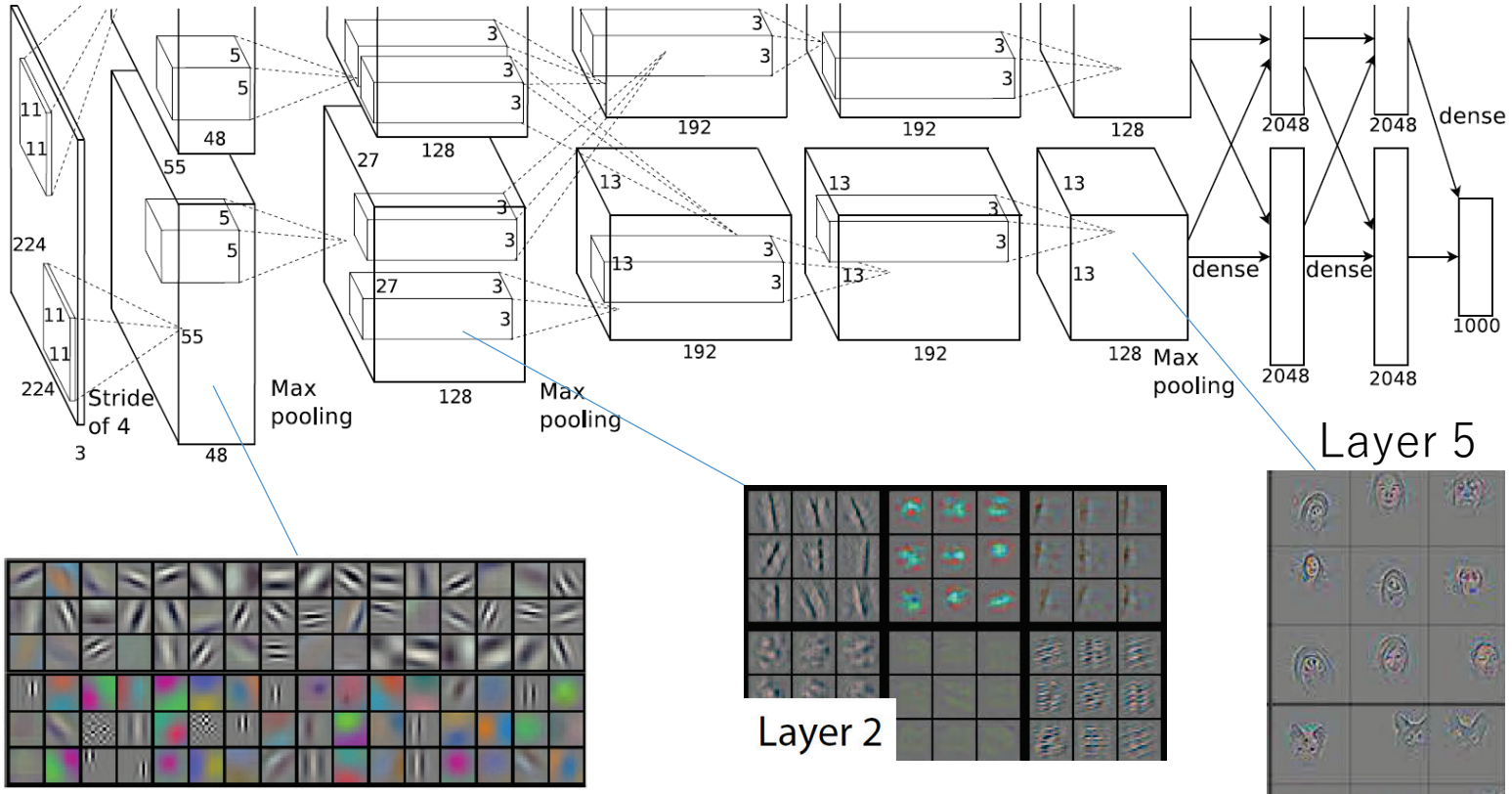
- ☆ReLU (Rectified Linear Unit)：　$h(u) = \max\{u, 0\}$
- シグモイド関数：　$h(u) = \dfrac{1}{1+e^{-u}}$





Alex-net [Krizhevsky, Sutskever + Hinton, 2012]

畳み込みニューラルネットを 5 層積み重ね（+pooling+3層の全結合層）

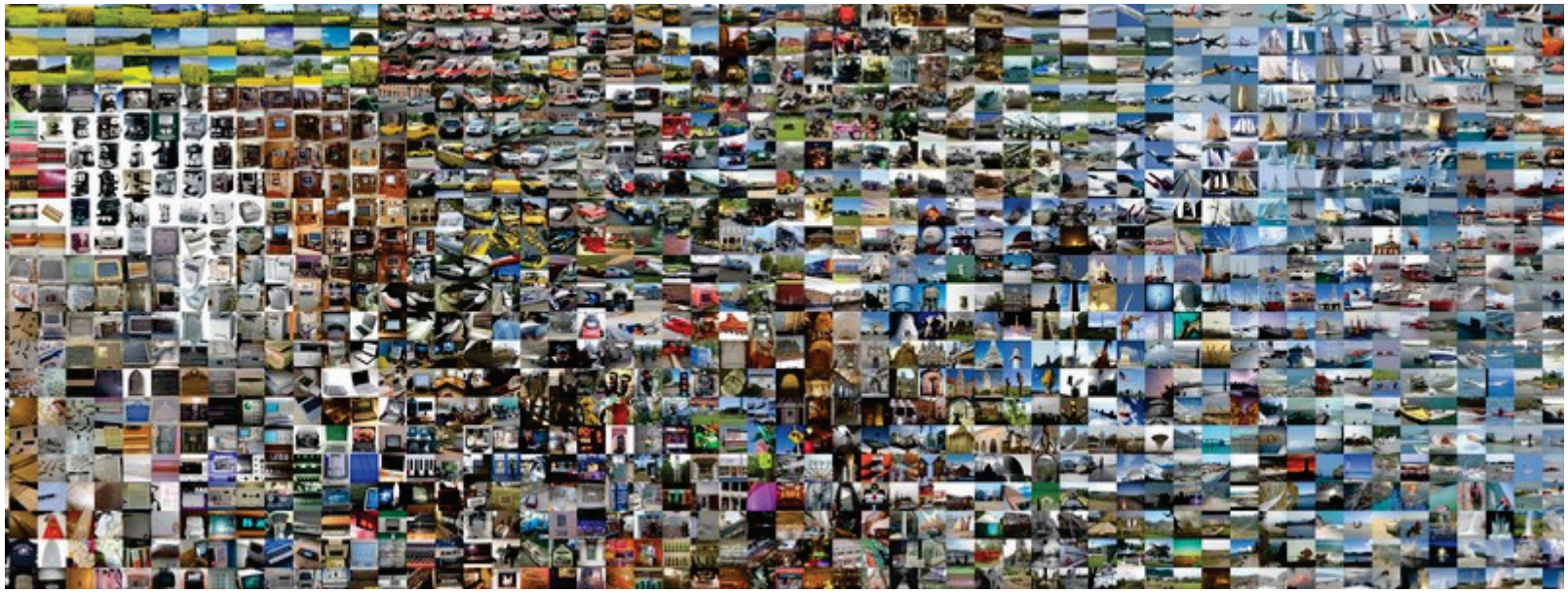


イメージパッチのようなものが学習されている

⇒　特徴量の自動学習

中間層ではより抽象的な情報がコードされる

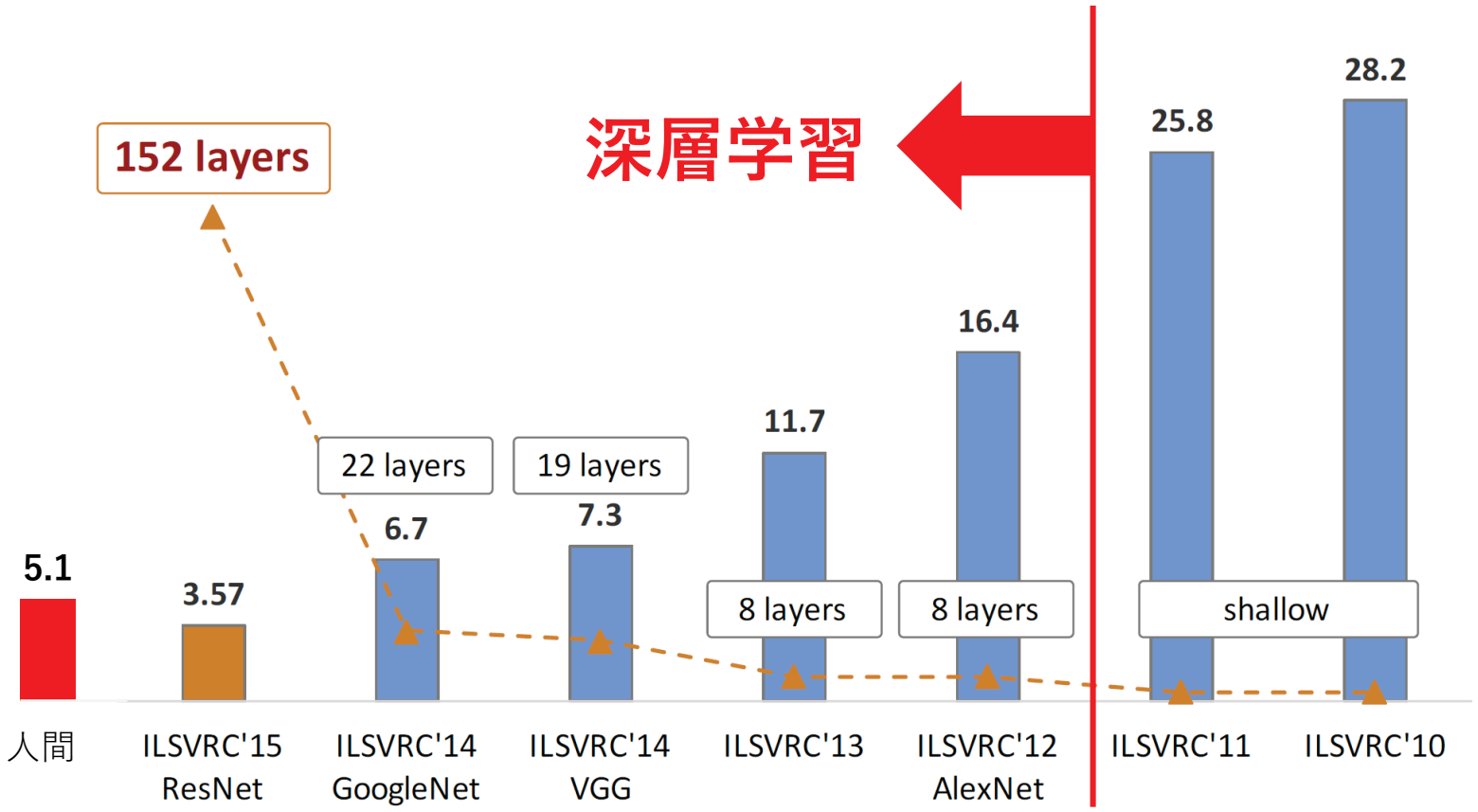
ImageNet



ImageNet: 21,841カテゴリ，14,197,122枚の訓練画像データ

[J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei.
ImageNet: A Large-Scale Hierarchical Image Database. In CVPR09, 2009.]

ResNet (Deep Residual Net)



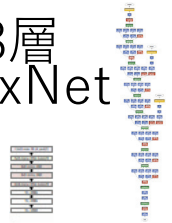
He, Zhang, Ren, & Sun. “Deep Residual Learning for Image Recognition”. CVPR 2016. (CVPR2016 best paper award)

He. “Deep Residual Network”. ICML2016 tutorial.

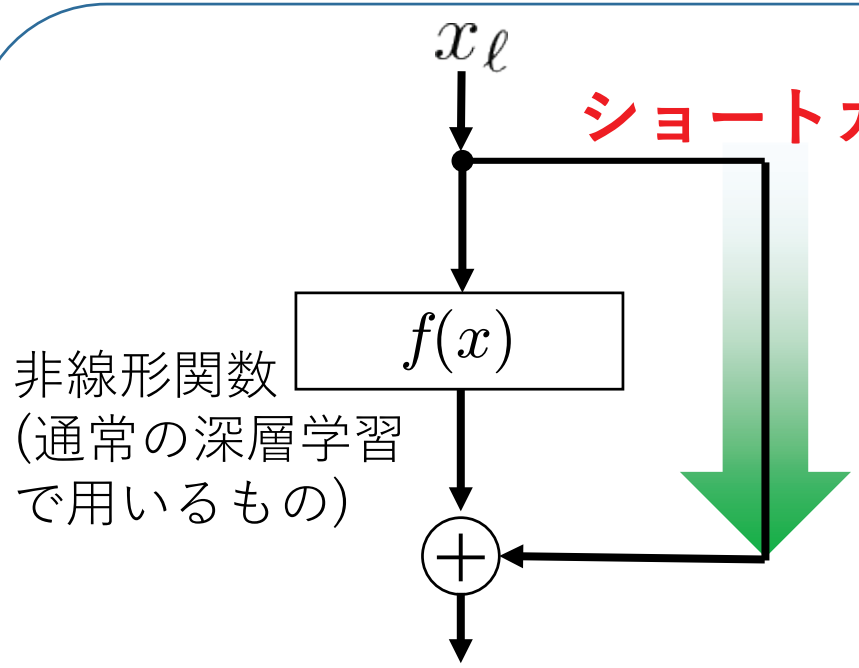
ResNet
152層

22層
GoogleNet

8層
AlexNet



ResNetの構造



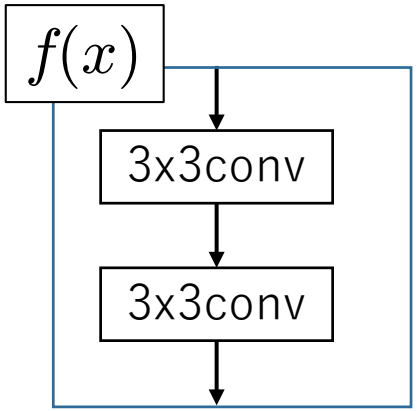
$$x_{\ell+1} = x_\ell + f(x_\ell)$$

$$x_{\ell+2} = x_\ell + f(x_\ell) + f(x_{\ell+1})$$

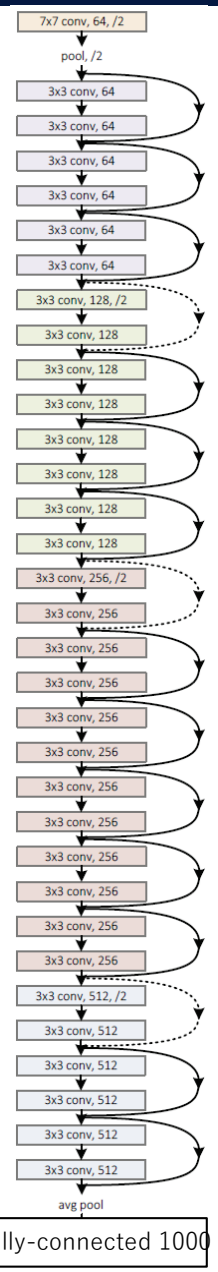
$$\vdots$$

$$x_L = x_\ell + \sum_{k=\ell}^{L-1} f(x_k)$$

情報が減衰せずに伝わる



CIFAR-10などの画像認識タスクでは$f(x)$として2層の畳み込み層を用いたものが良かった.



1000層を超えるものもある

ResNetの変種

- **Stochastic Depth**

  [Huang,Sun,Liu,Sedra,Weinberger: Deep Networks with Stochastic Depth, 2016]

  学習中に接続を確率的に切る.

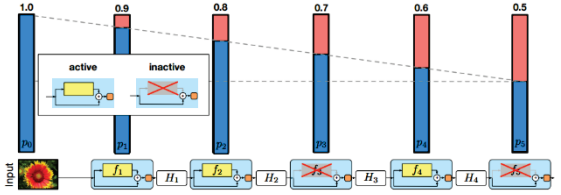
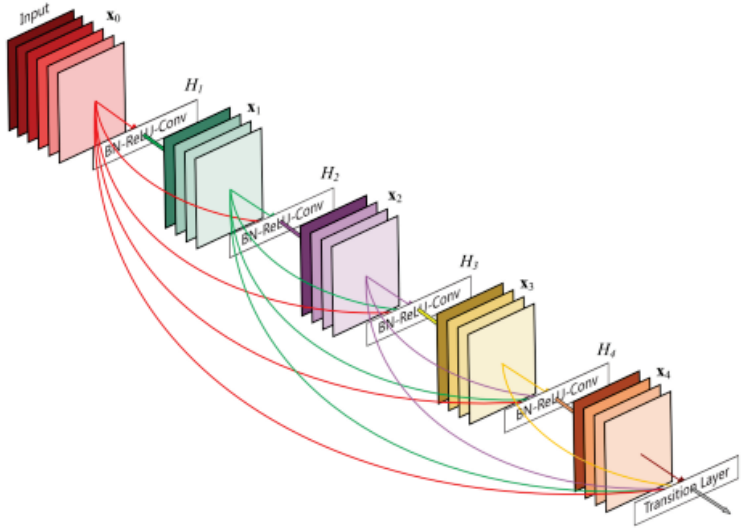
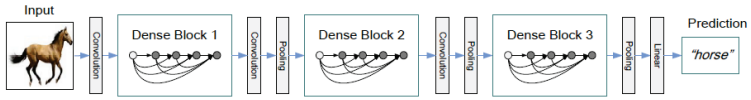


Fig. 2. The linear decay of p_ℓ illustrated on a ResNet with stochastic depth for $p_0 = 1$ and $p_L = 0.5$. Conceptually, we treat the input to the first ResBlock as H_0, which is always active.

- **DenseNet** [Huang, Liu, Weinberger, van der Maaten: Densely Connected Convolutional Networks, 2016]
  (CVPR2017 best paper award)

長いスキップを用いて密な結合を用いる



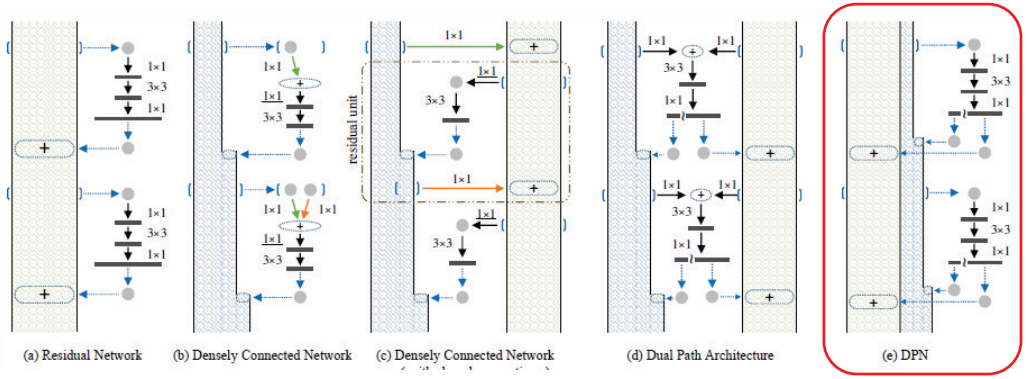


DenseNetの様子

- **Dual Path Networks**

  [Chen, Li, Xiao, Jin, Yan, Feng: Dual Path Networks, 2017]

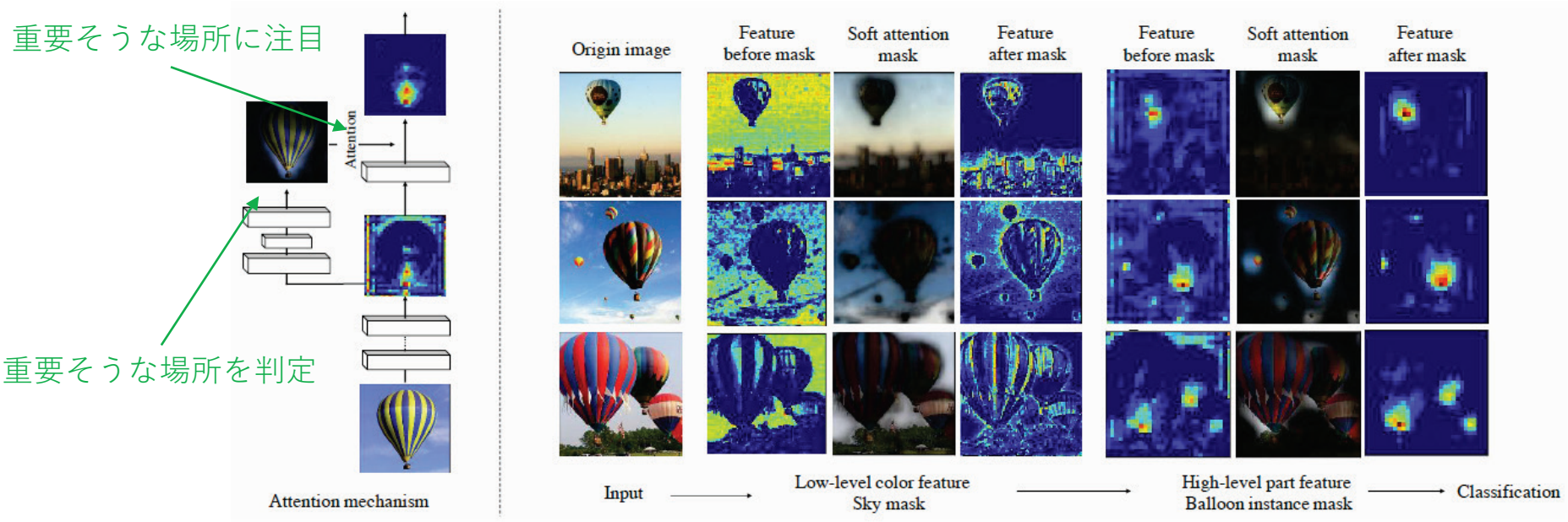
Dual Path



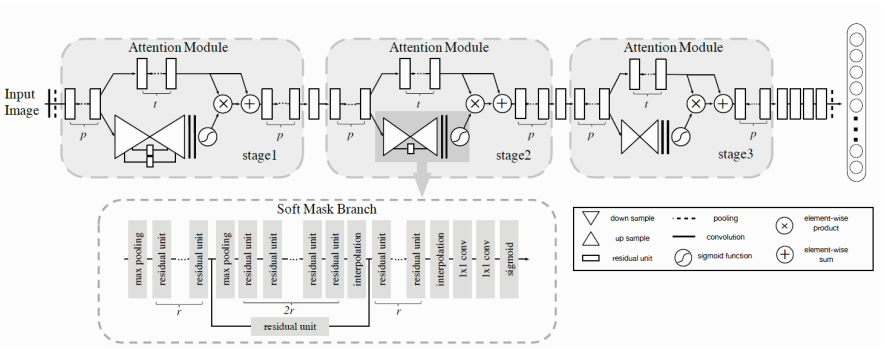
ResNetとDenseNetの良い部分を組み合わせ.
ILSVRC2017のObject localization部門で１位.

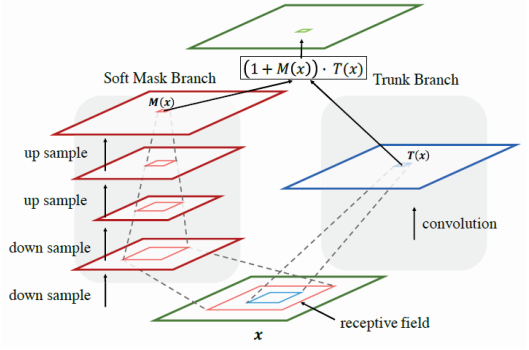
Residual Attention Network

ILSVRC2017のObject detection部門１位



ResNetに選択的注意の機構を付与





[Wang F, Jiang M, Qian C, et al. Residual Attention Network for Image Classification. arXiv:1704.06904, 2017]

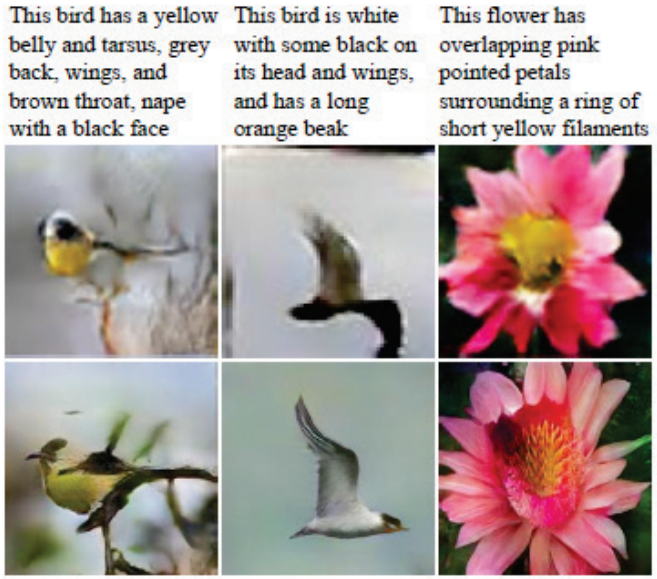
生成モデル

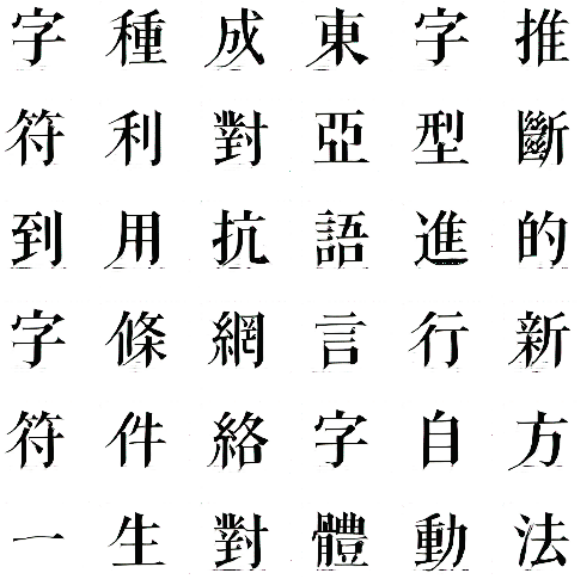
本物らしいデータを生成したい

生成データ　　訓練データ

(a) Stage-I images

(b) Stage-II images





[Tian: zi2zi, Master Chinese Calligraphy with Conditional Adversarial Networks,2017]

深層学習が生成した画像

生成モデル

目標：本物らしい画像を生成したい.

- GAN (Generative Adversarial Network) [Goodfellow+et al., 2014]

2つの構成要素

Generator: $x = G(z)$

Discriminator: $D(x) = P(x\text{が本物})$

G: 画像の素z (乱数) から偽画像xを生成. Dを騙そうとする.
D: 画像xが本物か偽物か判別. Gに騙されないようにする.

最適化問題

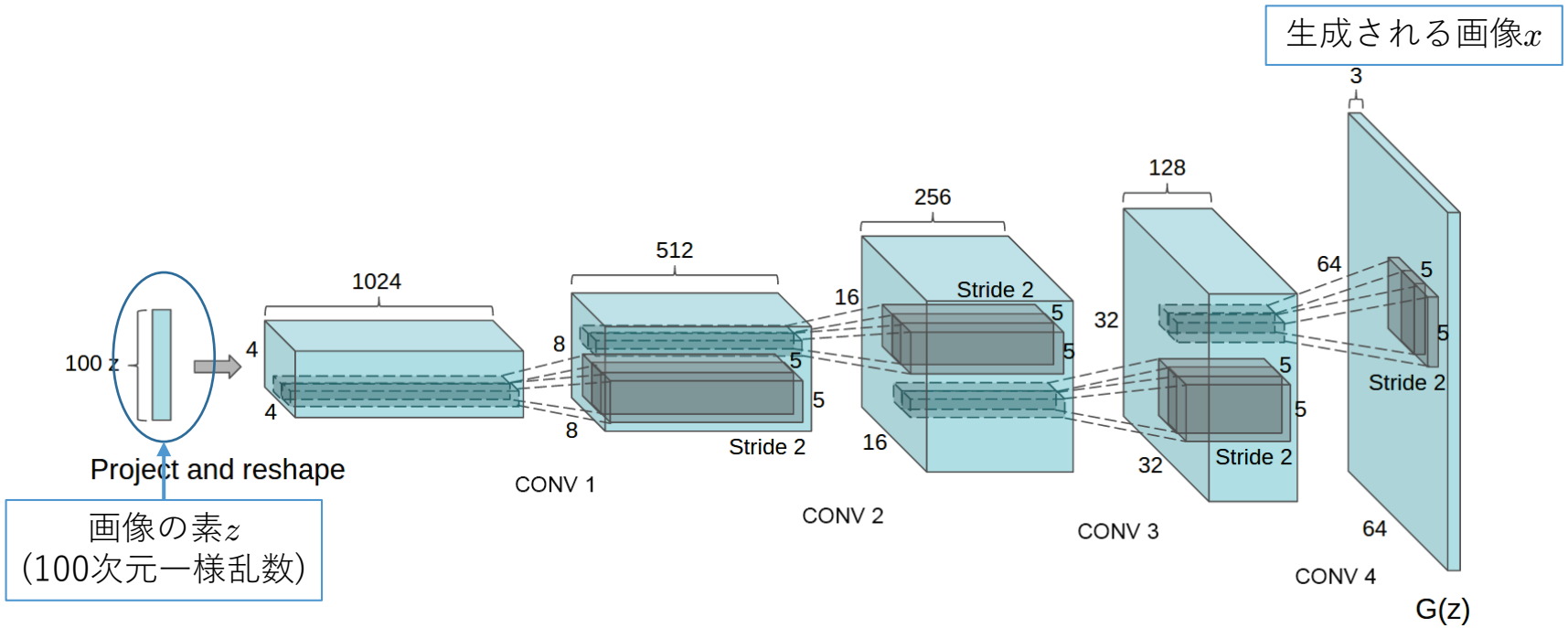
$$\min_G \max_D \mathrm{E}_{x\sim p_{\mathrm{data}}}[\log D(x)] + \mathrm{E}_{z\sim p_x}[\log(1 - D(G(z)))]$$

本当の画像を
本物と判別する確率

偽物の画像を
偽物と判別する確率

2016-2017にかなり流行
GANの変種まとめ：https://github.com/hindupuravinash/the-gan-zoo
※GANの他にもVAE (Variational Auto-Encoder)と呼ばれる方法もよく用いられている.

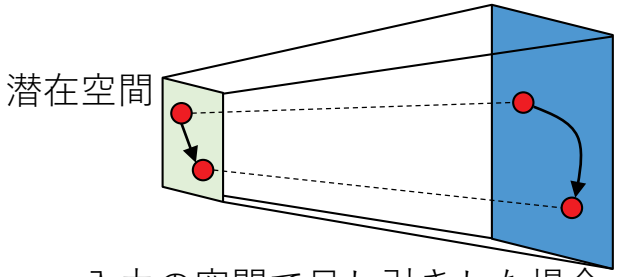
DCGAN (Deep Convolutional GAN)
畳み込みネットを用いたGAN

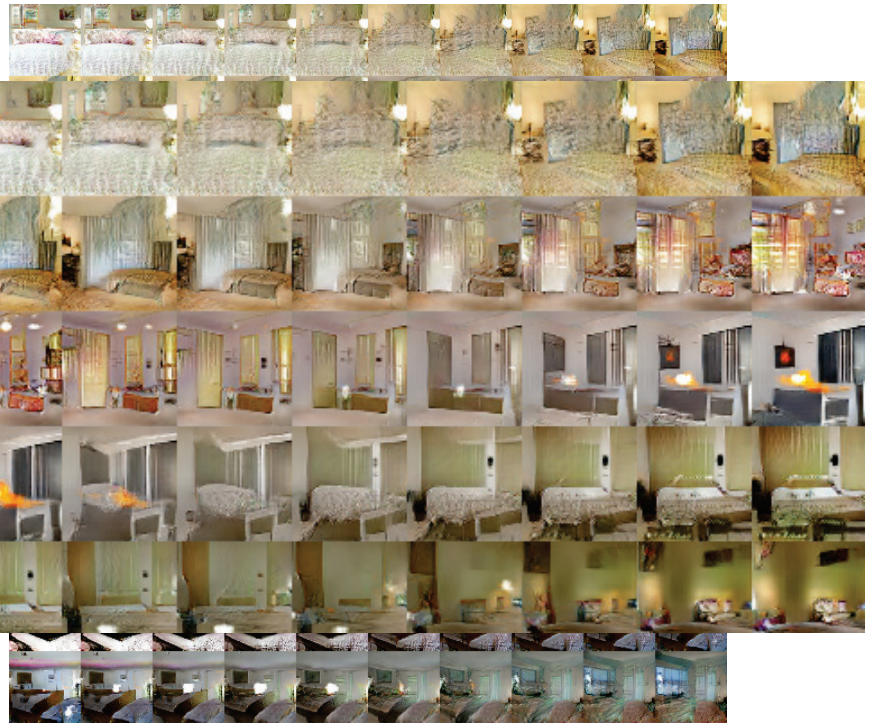


DCGANのGenerator

- 入力zは画像の低次元ベクトル表現にもなっている.
- Discriminatorも畳み込みネットを用いる.

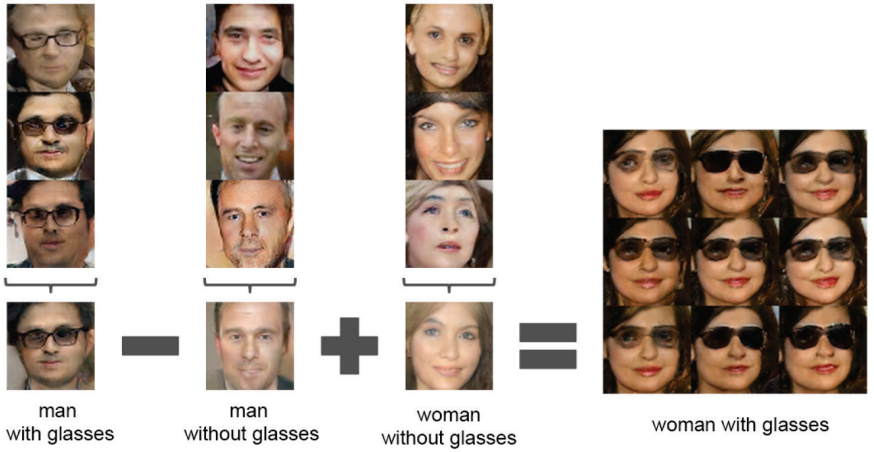
Radford, Metz & Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” ICLR2016.





生成されたベッドルーム画像

入力の空間で足し引きした場合



ピクセルごとに足し引きした場合

生成されたベッドルーム画像
入力zの凸結合で中間的画像が
得られる.

入力zを足し引きすることで意味
の足し引きが実現されている.
cf. word2vec.

Radford, Metz & Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks.” ICLR2016.

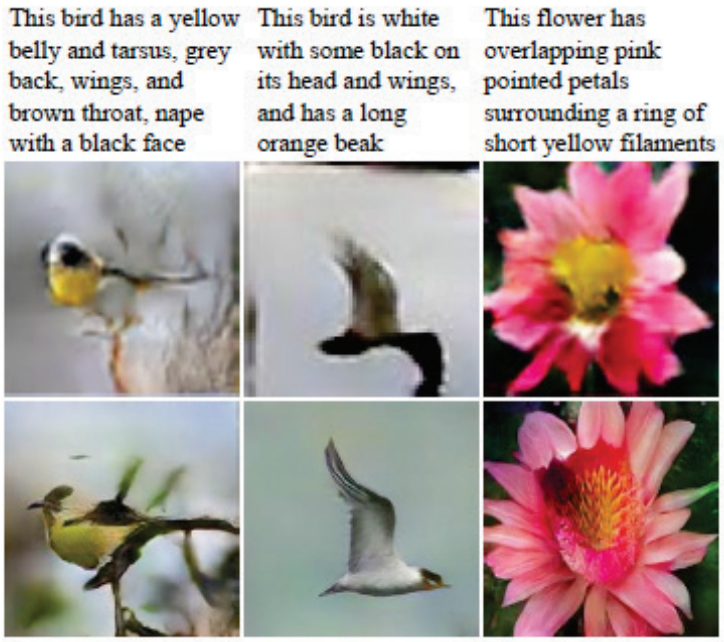
StackGAN

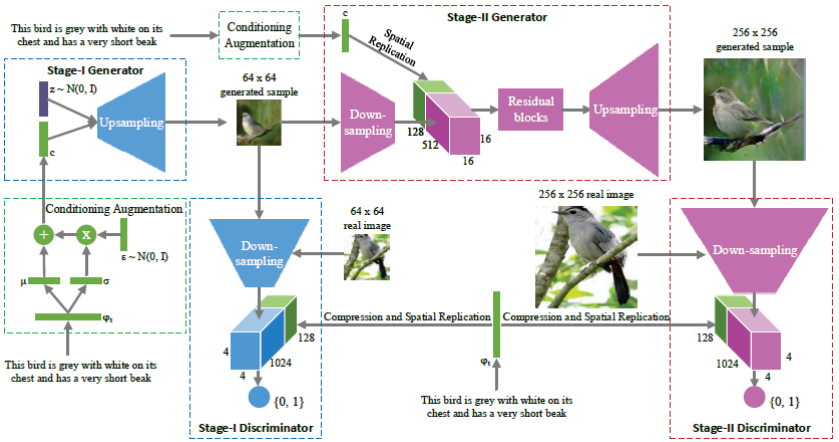
StackGAN [Zhang+etal.2016]　荒い画像を生成してからそれを高精細に修正（超解像）

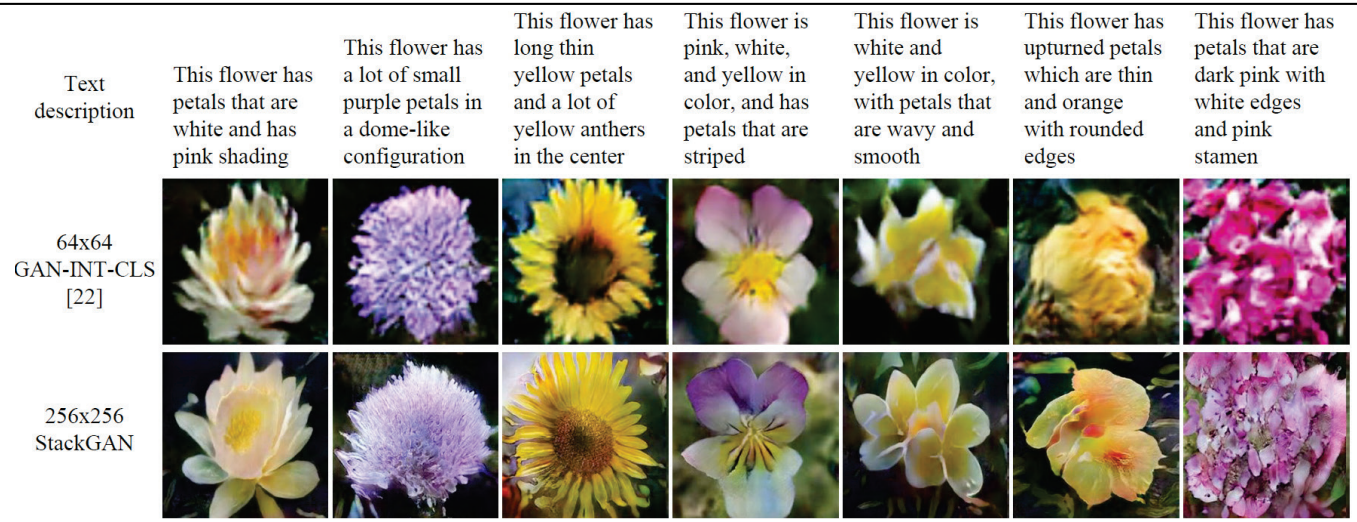
入力文章

荒い画像を生成

さらにこうなる





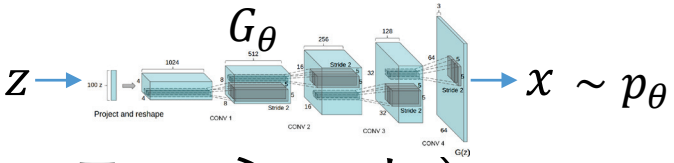


既存手法

StackGAN

GANの仕組み

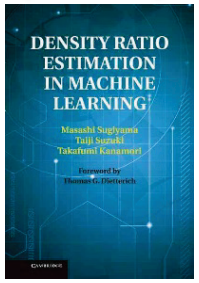
- z: 乱数（一様分布など）
- $x = G_\theta(z)$　（変数変換 by ニューラルネット）



適当な乱数を変数変換して目的の乱数（画像など）を生成

xの分布p_θ ⟷ 真の分布q　（近づけたい）

双対の関係



密度比推定の方法論

[Sugiyama, Suzuki, Kanamori: Density ratio matching. 2012.]

f-divergenceの最小化

f-divergence:
分布の距離(のようなもの)

$$\int q(x) f\left(\frac{p_\theta(x)}{q(x)}\right) \mathrm{d}x$$

GANはJensen-Shannon divergenceに対応

f-GAN [Nowozin, Cseke, Tomioka, 2016]

密度比$p_\theta(x)/q(x)$を1に近づける

Bregman-divergence:

$$\mathrm{BR}_f(\hat{r}) = \int q(x)\{f'(\hat{r}(x)) - f(\hat{r}(x)) + f(r_\theta(x))\}\mathrm{d}x - \int p_\theta(x) f'(\hat{r}(x))\mathrm{d}x$$

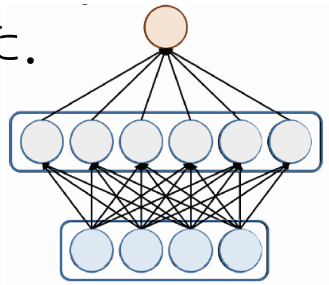
真の密度比とのBregman-divergenceを最小化して密度比を推定

B-GAN [Uehara+et al., 2016]

深層学習の理論

万能近似能力

ニューラルネットの関数近似能力は80年代に盛んに研究された.



$$f(x) = \sum_{j=1}^{m} v_j h(w_j^\top x + b_j)$$

なる関数が$m \to \infty$で任意の関数を任意の精度で近似できるか？

（「任意の関数」や「任意の精度」の意味はどのような関数空間を考えるかに依存）

hがシグモイド関数やReLUなら万能性を有する.

年		基底関数	空間
1987	Hecht-Nielsen	対象毎に構成	$C(R^d)$
1988	Gallant & White	Cos	$L_2(K)$
	Irie & Miyake	integrable	$L_2(R^d)$
1989	Carroll & Dickinson	Continuous sigmoidal	$L_2(K)$
	Cybenko	Continuous sigmoidal	$C(K)$
	Funahashi	Monotone & bounded	$C(K)$
1993	Mhaskar + Micchelli	Polynomial growth	$C(K)$
2015	Sonoda + Murata	Unbounded, admissible	$L_1(R^d)$

Kは任意のコンパクト集合

参考：園田, “ニューラルネットの 積分表現理論”, 2015.

理論より三層パーセプトロンでも中間層のユニット数を
無限に増やせば任意の関数を任意の精度で近似できる.

歴史的には後にSVMの理論に繋がってゆく.

(例：Gaussian kernelの万能性)

Q：ではなぜ深い方が良いのか？

A：深さに対して**指数的に表現力が増大**するから.

表現力と層の数

NNの“表現力”：領域を何個の多面体に分けられるか？

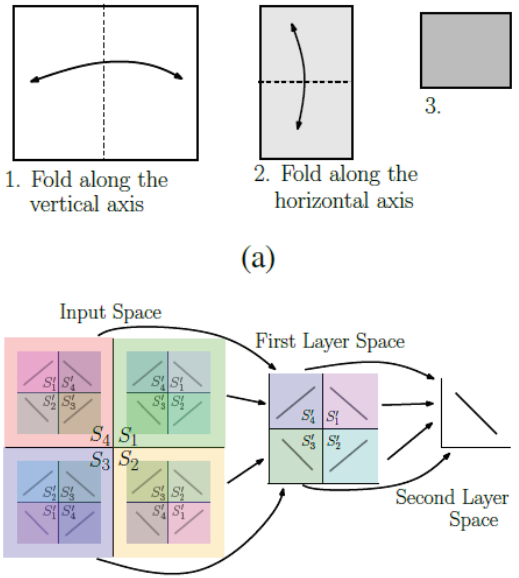
- 層の数に対して表現力は**指数的**に上がる.

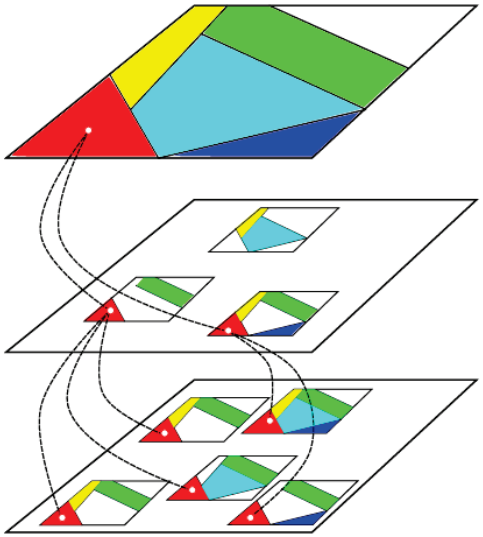
$$\left(\frac{n}{n_0}\right)^{L-1}\sum_{j=0}^{n_0}\binom{n}{j}$$

- 中間層のユニット数 (横幅) に対しては**多項式的**.

$$\sum_{j=0}^{n_0}\binom{n}{j}$$

L：層の数
n：中間層の横幅
n_0：入力の次元





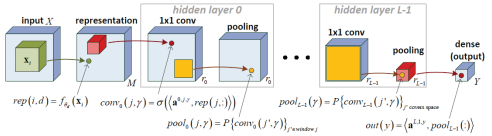
折り紙のイメージ

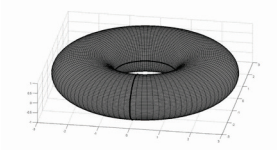
Montufar, Guido F., et al. "On the number of linear regions of deep neural networks." 2014.

多層で得する理由

他にも同様の結論を出している論文多数

- **多項式展開，テンソル解析** [Cohen et al., 2016; Cohen & Shashua, 2016]
  単項式の次数
- **代数トポロジー** [Bianchini & Scarselli, 2014]
  ベッチ数(Pfaffian)
- **リーマン幾何 + 平均場理論** [Poole et al., 2016]
  埋め込み曲率

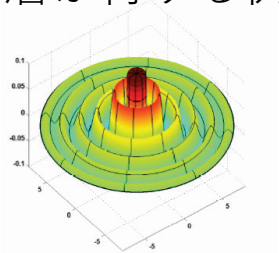




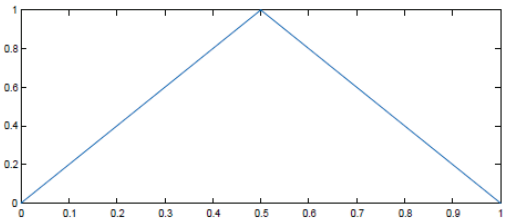
対称性の高い関数は，特に層を深くすることで得をする．

$$h(x) = \begin{cases} 2x & (0 \le x \le 1/2) \\ 2(1-x) & (1/2 \le x \le 1) \\ 0 & (\text{otherwise}). \end{cases}$$

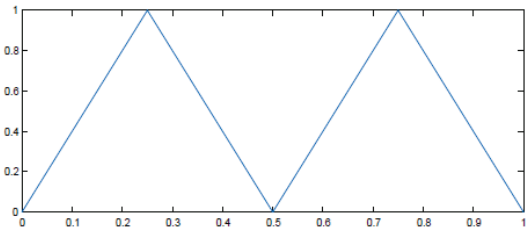
多層が得する例



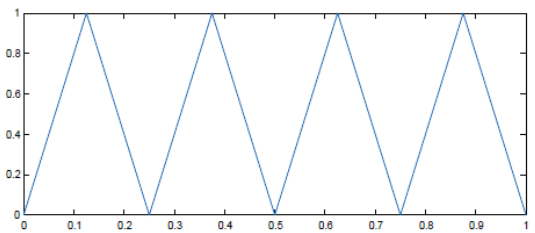
[Eldan,Shamir,ALT2016]



$h(x)$



$h \circ h(x)$



$h \circ h \circ h(x)$

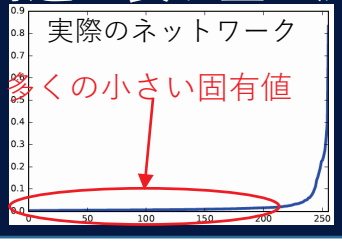
深層学習の汎化誤差理論

[T. Suzuki. Fast learning rate of deep learning via a kernel perspective. arXiv:1705.10182, 2017.]

深層学習のネットワーク構造を決定する指針はないか？

- 「自由度」という深層ネットワークの構造を表す量が汎化誤差に影響

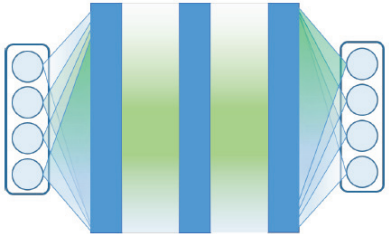
「自由度」は中間層の出力の分散共分散行列の固有値に対応して決まる



小さい固有値が多ければ横幅は狭くて良い

深層学習の汎化誤差を決める要素は何か？
→**自由度**が重要

深層NNの積分表現（真の関数）

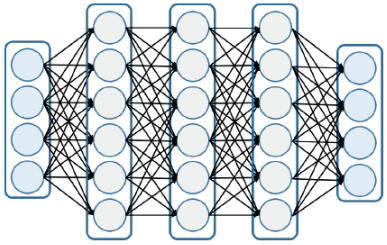


$$F_\ell(\tau, x) = \int_{\mathcal{T}_\ell} \underbrace{h_\ell^o(\tau,\tau')}_{\text{Weight}} \eta(F_{\ell-1}(\tau', x))\mathrm{d}Q_\ell(\tau') + \underbrace{b_\ell^o(\tau)}_{\text{Bias}}$$

有限近似

求積法の理論

有限次元モデル



汎化誤差=真の関数とモデルのずれ（バイアス）＋モデルの複雑さ（バリアンス）

- 真の関数を表すために積分表現を導入
- 積分表現から各層に再生核ヒルベルト空間を定義
- 空間に付随した「自由度」を定義

自由度 $N_\ell(\lambda) = \sum_{j=1}^{\infty} \frac{\mu_j^{(\ell)}}{\mu_j^{(\ell)} + \lambda}$

ただし$\mu_j^{(\ell)}$は各層に対応するカーネルの固有値

（各層の実質的次元）

定理 第ℓ層の横幅m_ℓが$m_\ell \geq N_\ell(\lambda_\ell)$ならば

$$\|\hat{f} - f^*\|_{L^2}^2 \leq 2(\hat{\delta}^2 + \epsilon_n^2)$$

バイアス $\hat{\delta} = \sum_{\ell=2}^{L} 2\sqrt{\hat{c}_\delta^{L-\ell-1} R^{L-\ell} \sqrt{\lambda_\ell}}$

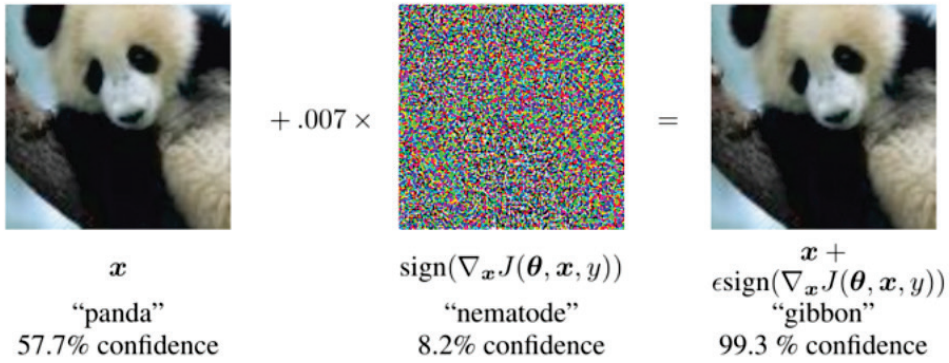
バリアンス $\epsilon_n = C\sigma\sqrt{\frac{\sum_{\ell=1}^{L} m_{\ell+1} m_\ell}{n}} \log(n).$

一般の深層NN $\sum_{\ell=1}^{L} n^{-\frac{1}{1+2s_\ell}} \log(n)$

3深NN（カーネル法） $n^{-\frac{1}{1+s_1}} \log(n)$

深層学習の脆さ

- Adversarial example



少し作為的ノイズを入れただけでパンダをテナガザルと間違える.
しかもかなり強い自信をもって間違える.

[Szegedy et al.: Intriguing properties of neural networks. ICLR2014.]

標識をハックすることで誤認識を誘発.

「STOP」を「スピード制限時速45mile」と誤認識

[Evtimov et al.: Robust Physical-World Attacks on Machine Learning Models. 2017]

敵対的入力 (adversarial example) に関する研究は現在盛り上がってる.
様々な対処法 (dropoutやVirtual Adversarial Trainingなど) も提案されている.
しかし，<u>深層学習の信頼性評価はまだ難しい</u>

これからの人工知能

- 足りないものは？

機械学習主体の方法 データからの帰納	人間の「知能」 論理による演繹

大きなギャップ

- 機械の得意とするところ：

  情報の検索，高次元データの扱い，高速な計算

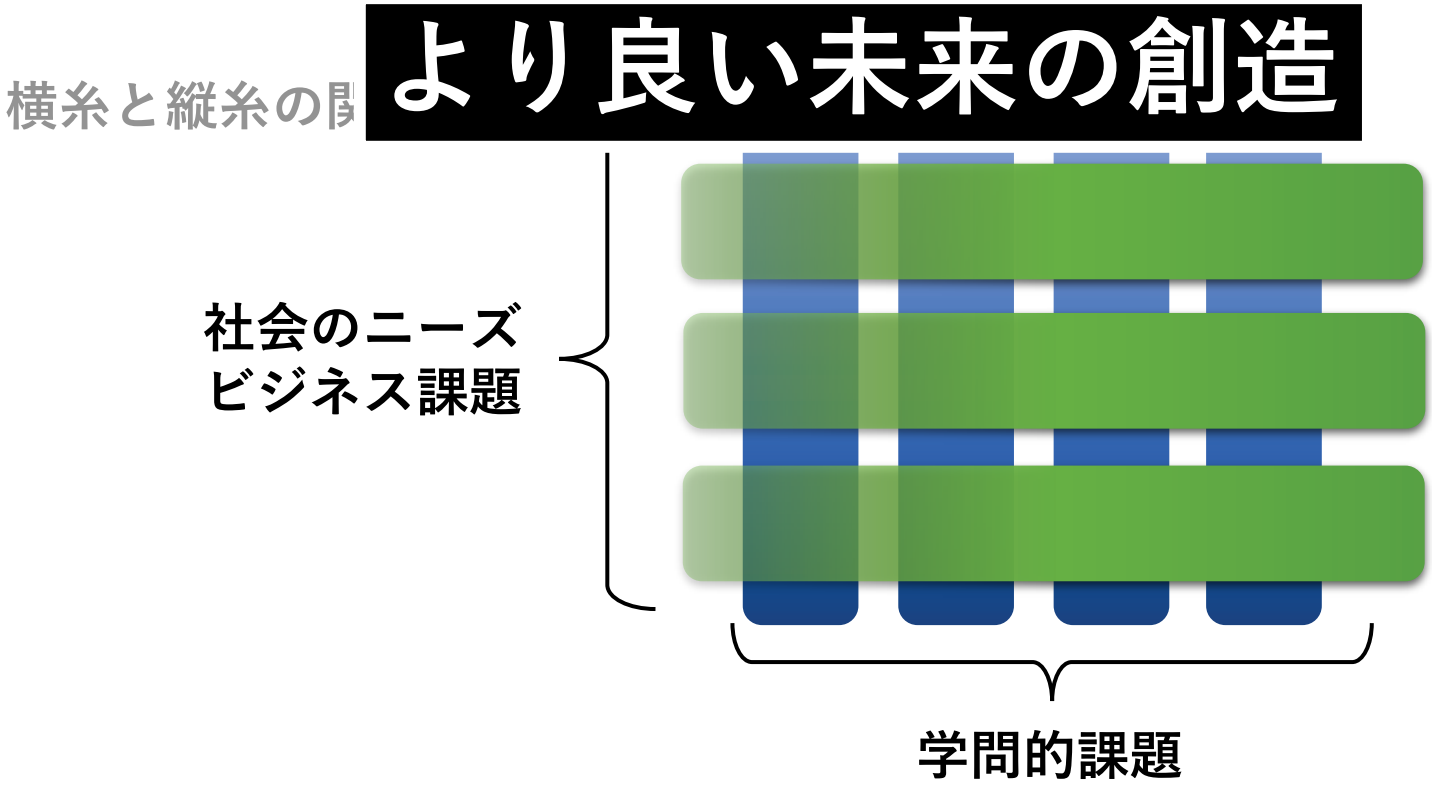
- 大量データを有する計算機の世界とのインターフェイスとしての「人工知能」

- 「知能」の本質的理解はまだまだ先

最後に

- 機械学習基礎研究による各種方法論の正しい理解
- 社会のニーズに応える産業界と時代によらない普遍的価値を求めるアカデミアの交流

「機械学習は両業界の距離が近い」



ご清聴ありがとうございました.