SASによる新しい大規模統計学入門Ⅱ
～きみはSASのDeepLearningを体験したか？～

小野　潔　　松澤　一徳
㈱　インテック　金融ソリューション開発センター

Large scale statistical forecasting using by SAS

Kiyoshi Ono,　Kazunori Matsuzawa
Financial Solutions Development Center,
Financial Solutions Service Division, INTEC Inc

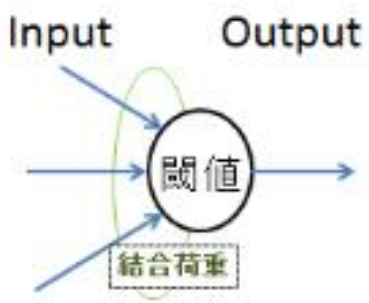
株式会社インテック（INTEC Inc.）

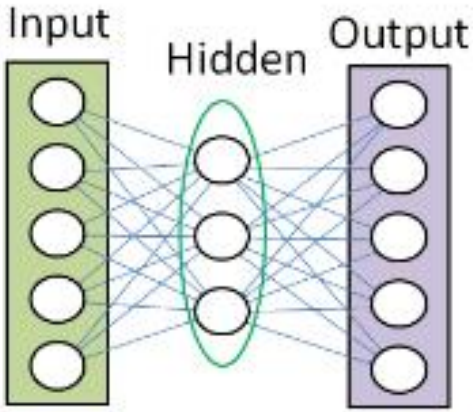
設立　　1964年1月11日
資本金　208億30百万円
代表者　代表取締役社長　日下　茂樹
東証一部上場　（証券コード：3626　ITホールディングス株式会社　TIS株会社と当社の共同持株会社）
従業員数 3,666名　グループ会社60社　　グループ従業員　19,472名（2014/9/30現在）
事業内容 技術研究、ICTコンサルティング、ソフトウェア開発
　　システム・インテグレーション、ネットワークサービス、アウトソーシングサービス

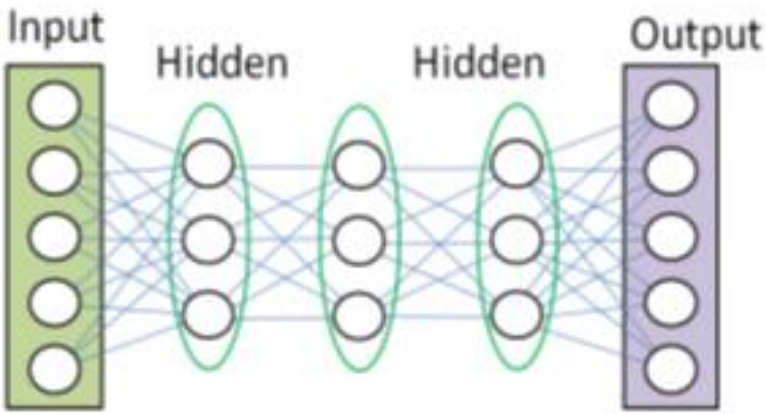
AIとDeep Learning の歴史

西　暦	イベント/人物/企業	内　　容
1946年	IBM	世界初のコンピュータENIAC
（1956～1960年代）	第一次AIブーム	探索・推論問題の解法（定理証明（1957）、遺伝子アルゴリズム（1958）） ダートマスワークショップ
1956年	ダートマスワークショップ	初めて人工知能（Artificial Intelligence）という言葉が出現
1958年	ローゼンブラット 第一次ニューロブーム	パーセプトロンを発表（ニューロの最初のニューロンモデル）。
1962年	ミンスキー（人工知能の父）	パーセプトロンは線形分離しか適用できないことを指摘。第一次ニューロブーム終焉
1964～1966年	MITのジョセフ・ハイゼンバウム	対話システムELIZA（人工無能）登場、入力文章に含まれるパターンをオウム返しし、会話を理解しているように見せかける　⇒　チューニング・テスト合格？
1970年代	AIの冬の時代到来	機械翻訳絶望、現実問題がとけず
1973～1976年	スタンフォード大学	MYCIN（マイシン）、エキスパートシステムの開発
1980年代	第二次AIブーム	知識工学の時代　　エキスパートシステム、自然言語・画像・音声理解システム
1982～1994年	通産省	第5世代コンピュータプロジェクト（1981年）に570億円。第二次人工知能ブーム。
1986年	ラメルハート、マクレランド、ヒントン 第二次ニューロブーム	ニューロの中間層以降を学習させるバックプロパゲーションを発表。ニューロを利用して非線形分離問題も解くことが可能に。
	ホップフィールド	ニューロによる最適化問題と連想記憶モデルを発表
	コホネン	ニューロによる自己組織化マップを発表
1990年～	再びAIの冬の時代到来	知識（ルール）獲得の失敗
1990年代		オントロジーの発達（概念集合の体系化）
1992年	ヴァプニック	サポートベクターマシンを発表
2003年～	IBM	IBM　Practical intelligent Question Answering Technology プロジェクト 論理形式分析と機械翻訳を結合するも、成果あげられず
2006年	ヒントン 第三次ニューロブーム	オートエンコーダー（自己符号化器）を発表。DeepLearni ngの発端。
2010年～	第三次AIブーム	自己学習、表現の時代 ビックデータ出現、Web広がり、DeepLearningの発見
2011年	IBM	質問応答システムWatsonが米国クイズ番組「Jeopardyジェパディ」でクイズ王に勝利
2012年	大規模画像認識コンテスト(ILSVRC)	DeepLearningがコンペティションで圧勝(以後3年連続優勝)
	Google	トロント大Hinton教授と学生の会社を買収
	ニューヨークタイムズ誌	トップ記事で、グーグル猫を掲載
2013年	Facebook	ニューヨーク大学のYann LeCun教授を所長に招き人工知能研究所を設立
2014年	Google	Deep Mind Technologies（英国）を4億ドルで買収
	Baidu（中国）	スタンフォード大学のAndrew Ng教授を所長に迎えてシリコンバレーにDeepLearningの研究所を開設（3億ドルの研究予算）
	Facebook	人工知能のVicarious社に4000万ドルの投資
	ドワンゴ、リクルート（日本）	各社AI研究所設立
	ロシアのAI	Eugene(ユージーン）君13歳がチューリングテストに合格
・・・・・・・・・・	・・・・・・・・・・	・・・・・・・・・・・・・・・・・・・・・・・・・・・・・
2045年頃	AIが人類を越す年	シンギュラリティ（技術的特異点）、2015/2/6 総務省の研究会開催

ニューラルネットワークの発達







ニューロン
モデル

線形判別

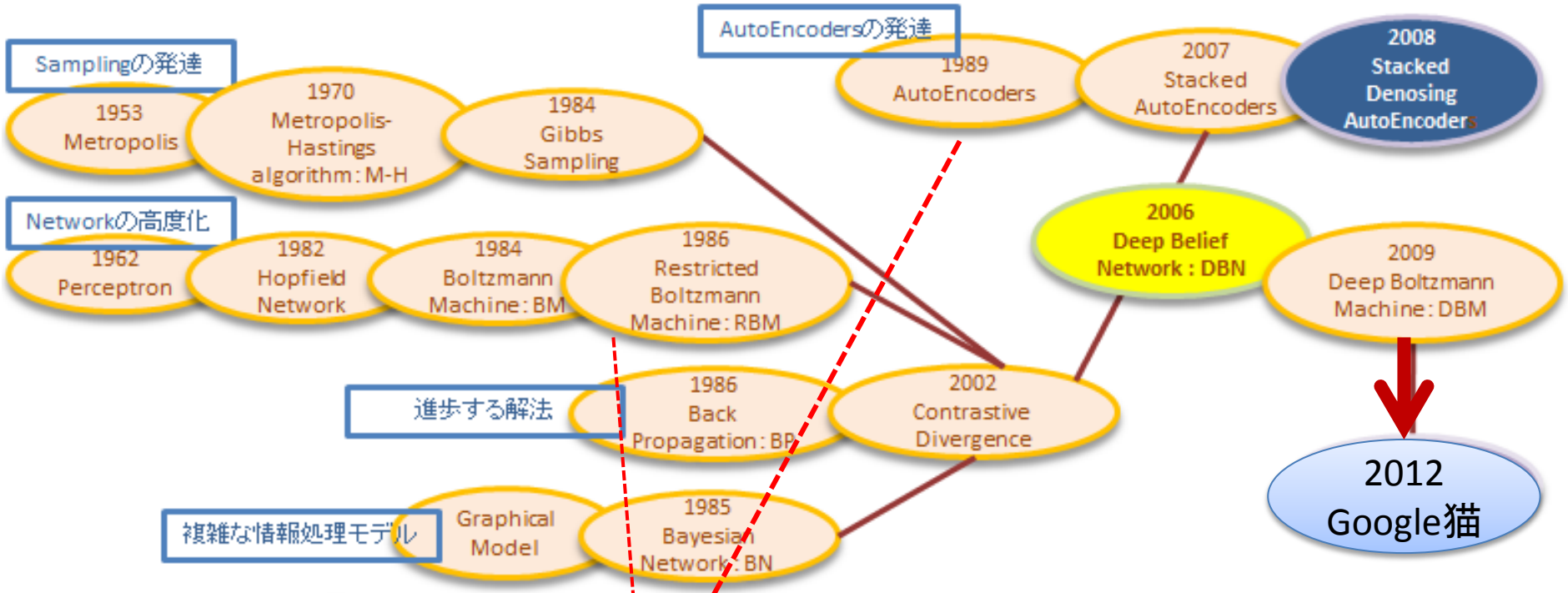
3層ニューラル
ネットワーク

単純な非線形

多層ニューラル
ネットワーク

複雑な非線形？

DeepLearning系譜



WHY NOW?

DeepLearningの基礎技術RBM & AutoEncoders は30年前の技術

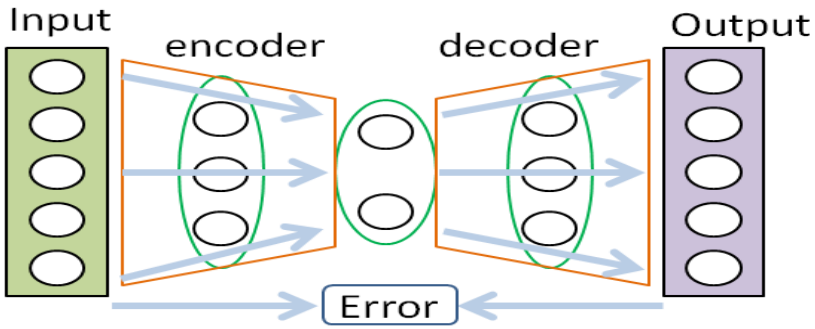
コンピュータの発達
Web等のデータ収集
分析技法／最適化法の発達

WHY MISTAKES?

データセットやモデルの規模を削減した結果、
高次元の特徴抽出も削り取った

Google猫　32×32（失敗）　200×200（成功）　人700万画素？

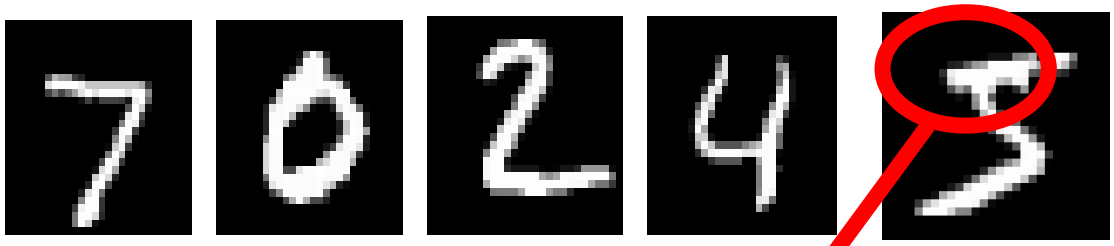
AutoEncoderの事前学習



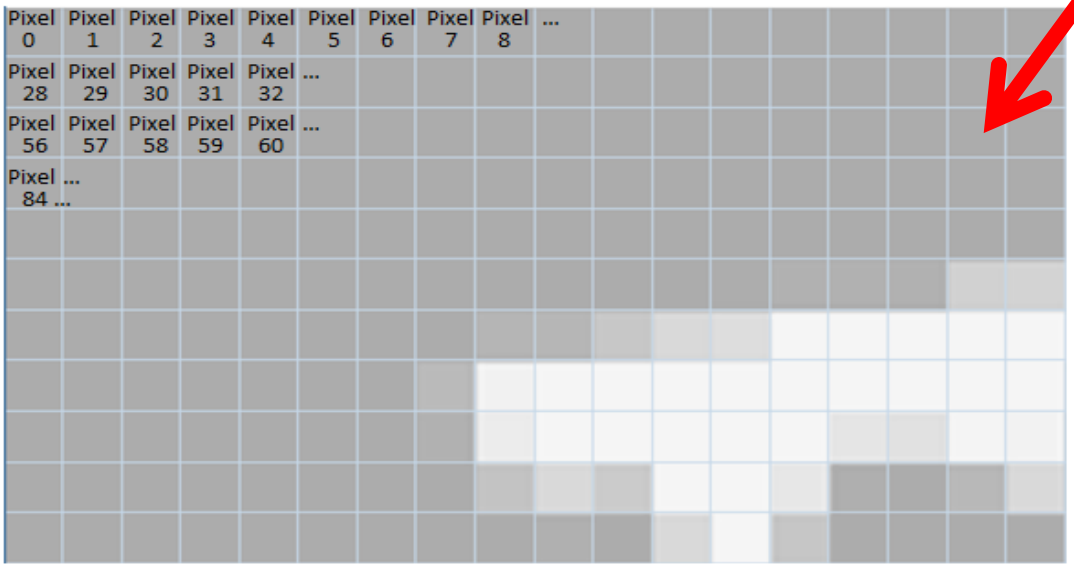
①構成：Encoder, Decoder, Representation
②学習：InputとOutputは同じにし（**恒等写像**）、2つの誤差を最小にする
③効果：Inputを適切な修正するAutoEncoder

Stacked Denosing AutoEncoder：
Inputにノイズを入れることで学習効果を高める

手書き数字のデータ

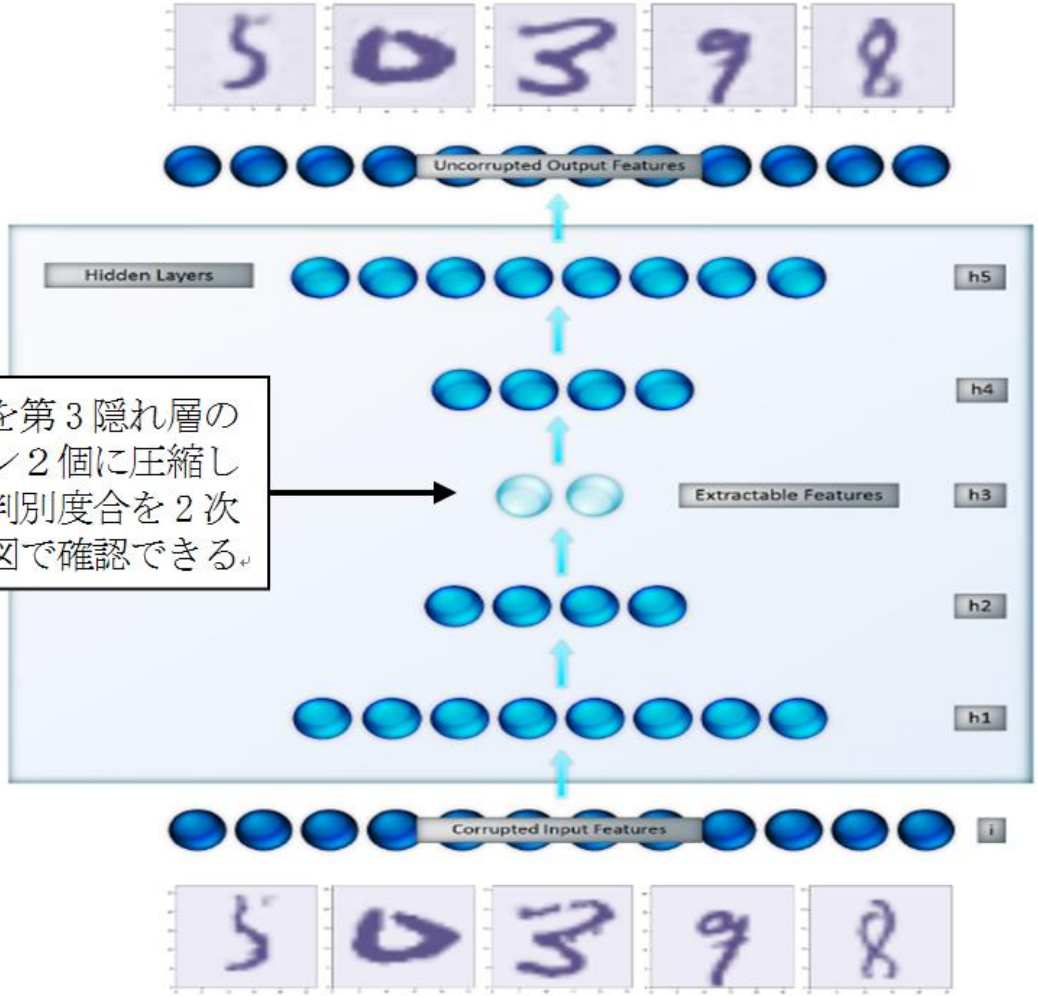


各Pixelの色の輝度の値をpixel0～pixel783（28×28）までに格納



	label	pixel203	pixel204	pixel205	pixel206	pixel207	pixel208	pixel209
201	3	0	217	252	252	253	252	252
202	6	0	0	133	251	180	6	0
203	1	0	0	0	0	0	0	0
204	1	0	0	0	0	0	0	32
205	1	0	0	0	0	0	0	0
206	3	95	179	9	0	0	0	0
207	9	0	0	0	9	130	236	252
208	5	0	136	253	235	174	199	222
209	2	0	0	0	50	158	124	47
210	9	0	0	0	10	59	132	248
211	4	144	253	142	0	0	0	0
212	5	0	0	0	0	0	0	0
213	9	62	152	152	254	172	152	112
214	3	0	0	4	9	10	99	224
215	9	0	0	15	121	221	252	252
216	0	0	0	0	150	252	254	253

Stacked Denosing AutoEncoder



特徴抽出を第３隠れ層のニューロン２個に圧縮したため、判別度合を２次元の散布図で確認できる。

補正後の手書き数字

出力層　784 個

第５層ニューロン数 300 個

第４層ニューロン数 100 個

第３層ニューロン数２個

第２層ニューロン数 100 個

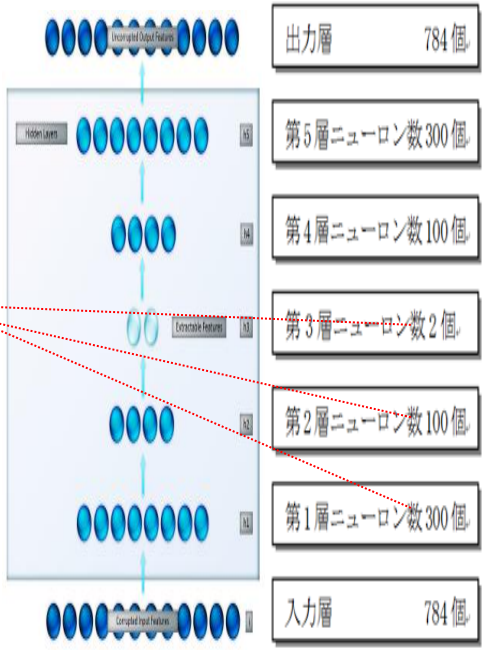
第１層ニューロン数 300 個

入力層　784 個

補正前の手書き数字

Deep Learning プログラム1

```
proc neural          /* DeepLearning のアーキテクチャは neural プロシジャーで構築する        */
    data= autoencoderTraining  dmdbcat= work.autoencoderTrainingCat;
    performance compile details cpucount=4  threads= yes; /* ENTER VALUE FOR CPU COUNT */
                                                          /*DO NOT EXCEED NUMBER OF PHYSICAL CORES */

    /* DEFAULTS: ACT= TANH COMBINE= LINEAR */ /* IDS ARE USED AS LAYER INDICATORS - SEE FIGURE 6 */
    /* INPUTS AND TARGETS SHOULD BE STANDARDIZED */
    archi MLP hidden= 5;              /* 隠れ層は5層を設定                    */
    hidden 300 / id= h1;              /*第1 隠れ層は300ニューロンを設定      */
    hidden 100 / id= h2;              /*第2 隠れ層は100ニューロンを設定      */
    hidden 2   / id= h3 act= linear;  /*第3 隠れ層は  2ニューロンを設定      */
    hidden 100 / id= h4;              /*第4 隠れ層は100ニューロンを設定      */
    hidden 300 / id= h5;              /*第3 隠れ層は300ニューロンを設定      */
    input &inputs   / id= i level= int std= std;
    target &targets / act= identity id= t level= int std= std;

    /*BEFORE PRELIMINARY TRAINING WEIGHTS WILL BE RANDOM */
    initial random= 123;
    prelim 10 preiter= 10;

    /* TRAIN LAYERS SEPARATELY */ /* 全各層を一度固定する */
    freeze h1->h2;
    freeze h2->h3;
    freeze h3->h4;
    freeze h4->h5;
    train technique= congra maxtime= 10000 maxiter= 1000; /* 学習 */
```

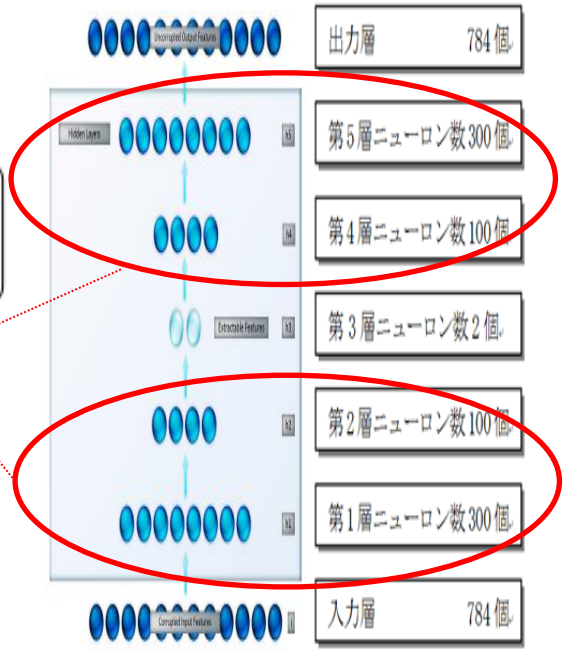


SAS の Deep Learning の実装では freeze（ネット層の固定）と thaw（ネット層の解放）のコマンドの使い方が鍵となる。

Deep Learning プログラム2

```
freeze i->h1;            /* 入力層から第1隠れ層を固定する      */
thaw h1->h2;             /* 第1隠れ層から第2隠れ層を開放する */
train technique= congra maxtime= 10000 maxiter= 1000; /* 学習 */

freeze h1->h2;           /* 第1隠れ層から第2隠れ層を固定する */
thaw h2->h3;             /* 第2隠れ層から第3隠れ層を開放する */
train technique= congra maxtime= 10000 maxiter= 1000; /* 学習 */

freeze h2->h3;           /* 第2隠れ層から第3隠れ層を固定する */
thaw h3->h4;             /* 第3隠れ層から第4隠れ層を開放する */
train technique= congra maxtime= 10000 maxiter= 1000; /* 学習 */

freeze h3->h4;           /* 第3隠れ層から第4隠れ層を固定する */
thaw h4->h5;             /* 第4隠れ層から第5隠れ層を開放する */
train technique= congra maxtime= 10000 maxiter= 1000; /* 学習 */

/* RETRAIN ALL LAYERS SIMULTANEOUSLY */ /* 各階層の学習を終えたあと、全層を開放する */
thaw i ->h1;
thaw h1->h2;
thaw h2->h3;
thaw h3->h4;
thaw h4->h5;
```

AutoEncoder を積み上げていく部分



Deep Learning では最後に微調整のために再度、学習を行う。ここがミソ。

```
train technique= congra maxtime= 10000 maxiter= 1000;  /* 学習*/

code file= 'D:¥sas_mat¥mnist¥code.sas'; run; /* ENTER SCORE CODE FILE PATH */
                                             /*  各隠れ層のニューロンの反応率をファイルへ書き込む*/
```

Deep Learning プログラム3

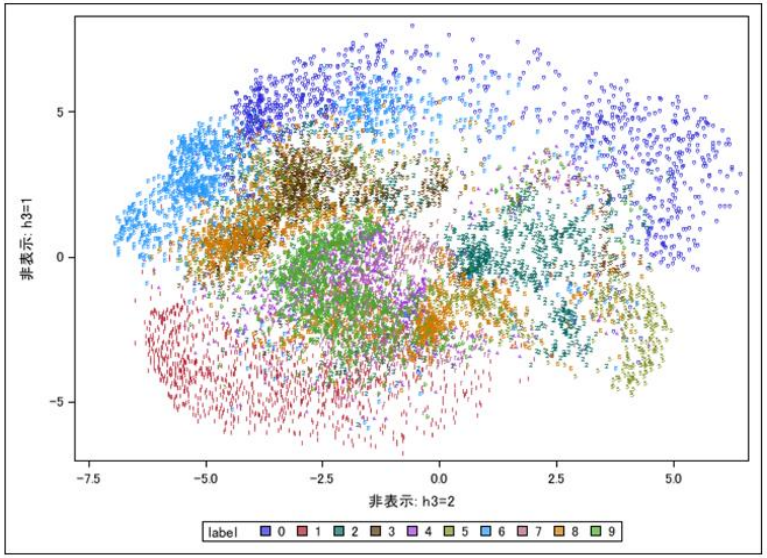
```
data extractedFeatures;     /* 手書き数字に反応するニューロンを表示する  */
    set autoencoderTraining;
    %include 'D:¥sas_mat¥mnist¥code.sas';          /* ENTER SCORE CODE FILE PATH */
    keep label h31 h32; run;     /*手書きの正解（数字）と第3隠れ層の2ニューロンの反応率を残す */
proc sort data= extractedFeatures; by label;
proc sgplot data= extractedFeatures;
      scatter x= h32 y= h31 /    /* 第3隠れ層のニューロン2個(h32とh31)の反応率で散布図を作成 */
            group= label groupdisplay= cluster clusterwidth= 0
            markercharattrs= (size= 3.75pt)   markerchar= label transparency= 0.3; run;
```

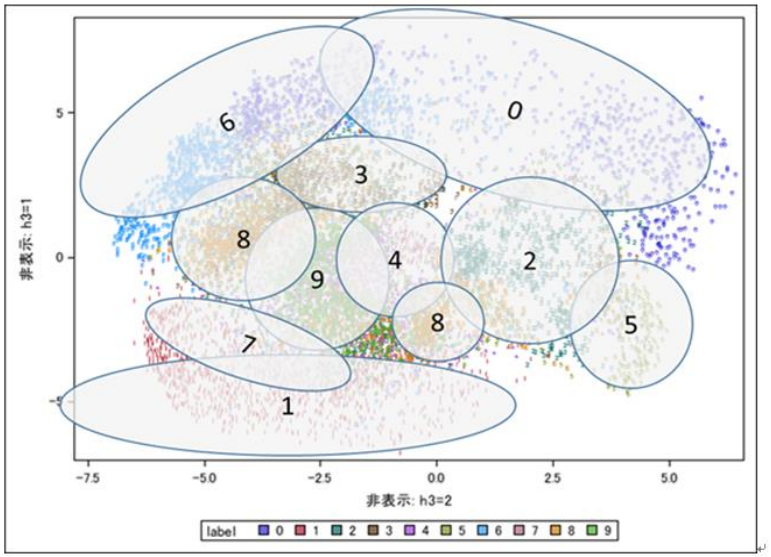
主成分分析プログラム

```
    /*  下記１行が主成分分析のプロシジャ。n=2 の指定により第１および第２主成分スコアのみ出力する*/
proc PRINCOMP data=autoencoderTraining out=PRCFEATURE n=2;  RUN;
    /*  出力した第１および第２の主成分スコアを sgplot する  */
proc sgplot  data= PRCFEATURE  scatter x= Prin1 y= Prin2 /group= label groupdisplay= cluster
    clusterwidth= 0  markercharattrs= (size= 3.75pt)  markerchar= label  transparency= 0.3;run;
```

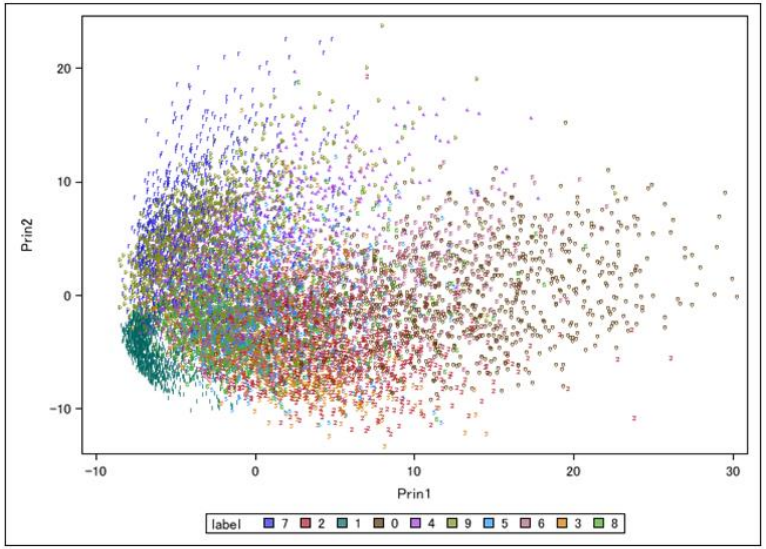
手書き数字の判別力

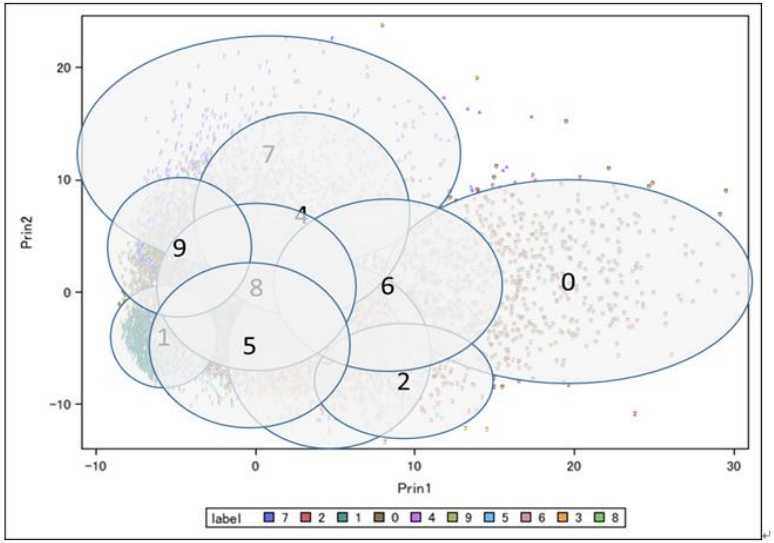
DeepLearning





主成分分析





Deep Learningの弱点

① 理論が不明
② 学習時間が長～い、メモリ不足⇒Errorとの戦い
③ 専門家の不足（世界で50人？）
④ 米国との技術格差（巨額のコンピュータ投資）

```
10825        /* 7階層*/ /* TRAIN LAYERS SEPARATELY */
10826       freeze h1->h2;
10827       freeze h2->h3;
10828       freeze h3->h4;
10829       freeze h4->h5;
10830       freeze h5->h6;
10831       freeze h6->h7;
10832       train technique= congra maxtime= 10000 maxiter= 1000;
ERROR: 例外が発生しました。   SAS タスク名は[NEURAL]です
NOTE: PROCEDURE NEURAL 処理(合計処理時間):
      処理時間          13:05:53.96        CPU 時間          35:31:27.12
ERROR: 十分なメモリを割り当てられません。少なくとも 11435K バイトが必要ですが、
1658K しか割り当てられません。メモリを増やすか、 プログラムを変更してください。
```

まとめ

- Deep LearningはSASver9.4で実装可能
  - 2012年の技術水準まで行けかどうか未知数？
  - 誰か確かめて！！！
- 研究には巨額なコンピュータ投資が必要
- 生物に近づけるには、さらなるブレークスルー
- SASユーザーの皆様が、機械学習に興味をもって頂けば、このチュートリアルは大成功！！