Tutorial for Artificial Intelligence

Revision 3.3

2016/12/9

有限会社 プロスペクト

PROSPECT
INNOVATIVE CONSULTING

目次

1.人工知能、人工汎用知能とは

人工知能とは

- 人工知能（Artificial Intelligence）とは：人工的にコンピュータ上などで人間と同様の知能を実現させようとする試み、そのための一連の基礎技術。またはエージェント。
  - 記号処理での知能の記述を主体とする情報処理やアプローチの意味でも使われる。
- 2つのAI学派
  - 「従来からのAI」（記号的AI、論理的AI、正統派AI、古き良きAI（GOFAI）などと）：数学の世界
    手法例：エキスパートシステム、事例ベース推論（CBR）、ベイジアンネットワーク、ふるまいに基づくAI
  - 「計算知能（Computational Intelligence：CI）（非記号的AI、美しくないAI、ソフトコンピューティング）：　開発と経験に基づく学習を基本とする。
    手法例：ニューラルネットワーク、ファジイ制御、遺伝的アルゴリズム、群知能。
- 強いAI（Strong AI）と弱い（Weak AI）：
  - 弱いAI：人間が全知能能力を必要としない程度の問題解決や推論を行うソフトウエアの実装や研究。特定問題解決器：ALPHAGO（碁）、WATSON（会話）もその例
  - 強いAI：人間の仕事をこなせる、または幅広い知識と何らかの自意識（人工意識、クオリア、魂）を持つ。
- 人工汎用知能（Artificial General Intelligence：AGI)
  - 人間レベルの知能の実現を目指す。強いAI。

弱いAIの例

- 弱いAI：特定問題解決器としての例
  - 画像処理・音声認識
  - 乗換案内・道案内
  - チェス、将棋、碁などのプレイヤ
  - クイズ・質問に答える：WATSON、Siriなど
  - 検索エンジン・投資エンジン
  - （あえて言えば）IME、自動電子レンジ

- コグニティブ・コンピューティング（Cognitive：経験的知識に基づく、認知の：Computing）
  - IBMが提唱する、これまでの「Tabulating（作表） Computing」「Programmable System」時代に対し、コンピュータが自ら学習し考え大量のデータを様々な情報源から収集、統合、瞬時に分析する時代を表す言葉。
  - 従来を「左脳型」と呼び、CCを「右脳型」コンピュータとも呼ぶこともある。
  - WATSONはその例

強いAI/人工汎用知能の機能

- 強いAIは以下の機能を単一のシステムで実現する。
  - できごとを記憶する。
  - 会話する、ジョークや比喩を理解する。
  - 高度な推論を行う。
  - 道具を使う。
  - 嘘をつく。
- 強いAIに必要な能力（松尾）
  - 自己意識（自我の芽生え）
  - 世界をモデル化（原理・法則の発見）
  - 行動と結果の予測（分析能力）
  - 目標を定め最適の行動を選択（意思決定）
- 結果として創造力を持つ。
  - 小説・詩などの創作、新しい技術の発明などが可能に。

技術的特異点 Technological Singularity

- 人工（汎用）知能が人間の能力を超えることでおきる出来事。
  - テクノロジーが急速に変化し、その結果人間の生活が後戻りできないほど変容してしまう。
  - 未来研究においては、その先は推測できないとされる。
  - AGIが自らを少しでも超えるAGIを産み出せるようになったとき一気に発散する（松尾）。
- ハードな離陸とソフトな離陸
  - ソフトな離陸：カーツワイルは「技術的特異点へむけての変化は緩やかである」とする
  - ハードな離陸：ヴィンジらは「自己改造する超知性が産まれ急激な変化が生まれる」とする

技術的特異点：時期の予測

- ニックボストロム（2015）：100年以内にはAIが人間を超え、人間と敵対する可能性がある。
- ヴィンジ（1993）：2005年～2030年の間：「30年以内に私たちは超人間的な知性を作成する技術的な方法を持ち、直後に人の時代は終わるだろう」
- カーツワイル（2005）：2045年：「SCハードウエアは2015年までにPCサイズでは2025年までにハードが人間の知能レベル（10PetaFLOPS）に達し、ソフトウエアが2025年頃までにチューリングテスト（機械と人間が区別できないかをテストする）に合格する。
  2030年代始め人間の総知能容量に到達、
  2045年には＄1000のコンピュータが10^{16}(10PFLOPS)の10^{10}倍（＝10^{26}FLOPS：100 Yotta FLOPS）の能力を持ち、技術的特異点に至るだろう。」
- ホーキンズ（1993）：
  「マインドステップ」：方法論または世界観に起きた劇的で不可逆な変化。間隔は短くなってきている。次のマインドステップは2021年、その後2つのマインドステップが2053年までにおきる。

技術的特異点の否定論と危険性

- 批判（内容省略：Wiki参照）
  - 否定論から批判、物理的観点からの批判、社会経済学的観点・生物学的観点からの批判がある。
  - さらに人工知能研究者の一部も無理としている。
- 危険性: AIが人間を排除するようになるのでは？
  - ニックボストロム（2015）
    - 100年以内にはAIが人間を超え、人間と敵対する可能性がある。
    - 170名の研究者の18%が人類に脅威を、13%が不利益を与えるようになるといっている。
  - ホーキング（2014）
    - 人口知能の進化は人類の終焉を意味する。
    - フェールセーフな仕組みが必要。
  - イーロン・マスク（2014）
    - 人工知能は核兵器より潜在的な危険をはらむ。最新の注意が必要。
- DeepMind：「AIの暴走を止める緊急停止ボタンの仕組みを開発」（2016/6/6記事）

人工汎用知能：計算能力の規模は？

- 脳：
  - 約10^{11}（1000億）個の神経細胞を約10^{14}（100兆：2兆程度という文献有）のシナプスで結合
  - 神経細胞情報更新：約10^{14}個/秒。神経細胞励起：200回/秒（約5ms）。
  - 神経細胞間信号伝達速度： 150m/s。消費電力：約20W。
- 技術的特異点実現時の計算能力予測値：
  - カールワイツの見積：人間の脳の計算能力を10 Peta（10^{16}）FLOPSと想定（SCでは既に到達）。
    技術的特異点がおきる値：その10^{10}倍（＝10^{26}FLOPS：100 Yotta FLOPS）
- 最適化していない脳のシュミレーションに必要な計算能力：10^{18}FLOPS（1 Exa FLOPS）
- IBM Blue Brain Project：
  - Blue Gene（世界TOP3のSC：Rpeak～20 PetaFLOPS）上に約6万個（$6x10^4$）の神経細胞と全長5kmのシナプシスからなる大脳新皮質シミュレート中。
  - 最終目標は脳全体のシミュレーション。
- ALPHAGO Hardware規模（ファンフィ対戦時）
  - 約CPU 1200個＋GPU 約170枚： 単純計算 3.8PetaFLops（実際は分散系でこれよりずっと低い）
    XEON 3.1GHz（198.4GFLOPS）、nVIDEA Tesla P100（21.2TLOPS：FP16）と仮定

http://wired.jp/2016/01/31/huge-breakthrough-google-ai/ から：但し後述するTPUが使われたという記事がありどちらが真が不明

（Deep Learningでは半精度浮動小数点FP16（13E5)で十分）

強いAI/AGI まとめ

- 強いAIはまだまだ研究が必要。
  - 方向性すらわかっていない。そもそも「知能」とは何かすらわかっていない。
- Deep Learningというソフトウエア技術が、弱いAIの機械学習の方法に50年来のブレークスルーを起こした。 が、
  - 人工汎用知能ができたとき、そのごく一部にDeep Learningが使われるかもしれない。（使われないかもしれない）
  - （今の）Deep Learningは人間が介在して（ニューラルネットの階層構造と）調節・工夫しないと使えない。これを人口汎用知能が代行できるようになる時期は見えていない。
- 松尾意見：モラベックスのパラドックス（子供のできることほどAIには難しい。高度な推論より認識や運動スキルの獲得が難しい。）の解決にはDeep Learningが不可欠
  - この3年ぐらいで「画像認識が人間の精度を上回った」「運動の習熟ができるようになった」：Deep Learningが現実世界から特徴量を抽出する（人の）作業を不要にした。
  - 「大人の人工知能」は専門家が裏で作りこんでいる。
    販売、マーケティング、医療、金融、教育向けか。
  - 「子供の人工知能」は人間の発達と同じ進化過程をたどる（認識⇒運動⇒言語の順）。
    モノづくりにはこれが必要。

2. 弱いAIとDeep Learning

人工知能をめぐる動向

- 第1次AIブーム（1956～1960年代）：探索・推論の時代
  - ダートマスワークショップ（1956）
    - 人工知能（Artificial Intelligence）という言葉が決まる
    - 世界最初のコンピュータENIAC (1946)のわずか10年後
  - 数学の定理証明、チェスを指す人工知能等
- …冬の時代

考えるのが早い人工知能

- 第2次AIブーム（1980年代）：知識の時代
  - エキスパートシステム
  - 医療診断、有機化合物の特定、…
  - 第5世代コンピュータプロジェクト：通産省が570億円
- …冬の時代

ものしりな人工知能

- 第3次AIブーム（2013年～）：機械学習・ディープラーニングの時代
  - ウェブとビッグデータの発展
  - 計算機の能力の向上

データから学習する人工知能

松尾：人口知能の未来-ディープラーニングの先にあるもの（総務省資料　P.4 ）

人工知能をめぐる動向（補足）

- 「脳の再現」（いいすぎ、単純化しすぎ）：第１次ブーム（1956年～1974年)
  - 1958年：「10年以内にデジタルコンピュータはチェスの世界チャンピオンに勝つ」
    「10年以内にデジタルコンピュータは新しい重要な数学の定理を発見し証明する」
    1965年:「20年以内に人間ができることは何でも機械でできるようになるだろう」
    1967年:「一世代のうちに（中略）人工知能を生み出す問題のほとんどは解決されるだろう」
    1970年:「3年から8年の間に、平均的な人間の一般的知能を備えた機械が登場するだろう」
- 初期のニューラルネット「2層構造：形式ニューロンを2重にしたパーセプトロン」では「線形非分離問題がとけない、XORがとけない」。
  - だから「つかいものにならない」「だめだ」：1969年　ミミンスキーとパート

  ⇒　第一次冬の時代：1974年～1980年
- 1986年 Backpropagation（3層構造ニューラルネット）発表。数学的に「うまく設計すればどんなものも学習できることが」証明され、第2次ブーム（1980年～1987年）始まる。
  - BPも多層化すると性能が下がる。
  - 精々3~4層が限界。NNではないSVMの方が性能が良い。
  - 3つの難題（次ページ参照）：特徴量設定・フレーム問題・シンボルグラウンディング

  ⇒　第2次冬の時代：　1987年～1993年

  ⇒　「なんとかして、多層ニューラルネットをうまく学習させられないか」

  ⇒　Deep Learningの発明

第2次AIブームの壁

- 難しい問題1：機械学習における特徴量の設計（Feature engineering）
  - 機械学習において、変数（特徴量）の設計が難しかった。
  - 人間が対象をよく観察して設計するしかなかった。
- 難しい問題2：フレーム問題
  - 人間が知識を記述することで、人工知能を動作させる。
  - そのときに、いくら知識を書いても、うまく例外に対応できない。
- 難しい問題3：シンボルグラウンディング問題
  - シマウマがシマのある馬だと、計算機が理解することができない。
  - シンボル（記号）がそれが指すものと接続（グラウンド）しておらず、シンボルの操作ができない。

結局のところ、いままでの人工知能は、
人間が現実世界の対象物を観察し、「どこに注目」するかを見ぬいて（特徴量を取り出して）、モデルの構築を行っていた。
その後の処理は自動で行うことができたが、モデル化の部分に人間が大きく介在していた。
それが、唯一にして最大の問題であった。

7

松尾：人口知能の未来-ディープラーニングの先にあるもの（総務省資料　P7）

将棋電王戦

IBM ワトソン

http://venturebeat.com/2011/02/15/ibm-watson-jeopardy-2/, http://weekly.ascii.jp/elem/000/000/207/207410/

ディープラーニング(2007-)

ILSVRCでの圧勝(2012)
Googleの猫認識(2012)
ディープマインドの買収(2013)
FB/Baiduの研究所(2013)

車・ロボットへの活用
自動運転
Pepper

機械学習
ウェブ・ビッグデータ

統計的自然言語処理
(機械翻訳など)
検索エンジンへの活用

タスクオントロジー

MYCIN(医療診断)
DENDRAL
エキスパートシステム
オントロジー
ワトソン(2011)

LOD(Linked Open Data)

Eliza
対話システムの研究
Caloプロジェクト
Siri(2012)
bot

探索
迷路・パズル
プランニング
STRIPS
チェス(1997)
Deep Blue
将棋(2012-)
電王戦
囲碁

1956
1970
1980
1995
2010
2015

第一次AIブーム
(推論・探索)

第二次AIブーム
(知識表現)

第三次AIブーム
(機械学習・ディープラーニング)

松尾:人口知能の未来-ディープラーニングの先にあるもの(総務省資料 P.5)

深層学習Deep Learningの登場（1）

- 第3次AIブームの火付け役。AIにおける50年来のブレークスルー。
  - ILSVRC（Large Scale Visual Recognition Challenge）2012
    （画像中の物体の認識率を競うベンチマーク）エラー率26%⇒17%の飛躍的認識率向上を記録。
  - 2012年 Google Brain：猫の学習の成果を発表

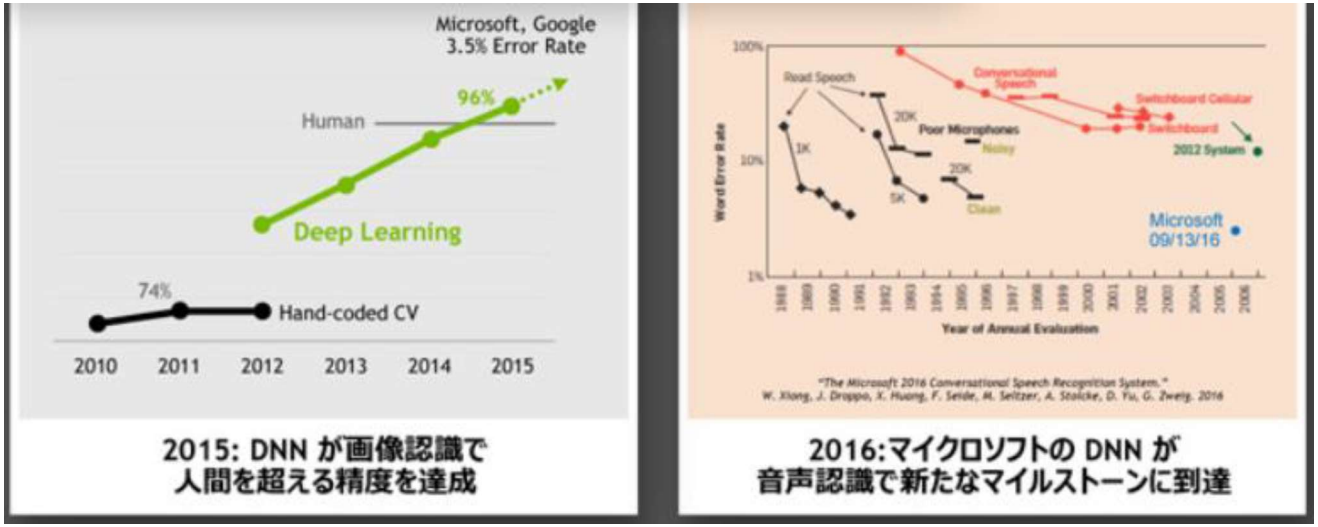
    （1000台のサーバの16000個のコアを使用：3日間で実現）
  - 2013年 MIT Tech. Review「10 Break through technology」に選出される。DL元年。
    - 2012年をDLビックバンだとする説もあり（nVIDIA）
- Deep Learningは、三層以上のニューラルネットの機械学習を実現させるソフトウエア技術
  - 人間が特徴量のモデル化を行なっていたのが、これまでの問題の根源だった。DLでは、モデル化せずに学習に使えるデータ量を膨大にして、これを膨大な計算量でじっくり機械学習を進める。
    - Internetが必要なデータ量を与え、NETWORK接続されたGPUが計算能力を提供できたことが背景。
  - 「Auto Encoder（自己符合化器：後述）」の発明が、DL実現のきっかけ。（CNNではもはやあまり使われない）
  - 最終段出力を正解と比較するか、報酬を与えて、誤差逆拡散法（後述）で（ランダムに選んだサブセットで）逆伝搬させる。さらに入力にノイズを与えデータ数を増やしたり、ニューロンの一部（50%など）を停止（ドロップアウト）させるなど、膨大な数の学習により収束させる。
  - 実際に内部でどのようなメカニズムが働いてうまくいっているのは解明できていない。

深層学習Deep Learningの登場（2）

- Deep Learningの段数
  - ALPHAGOのNNは14層、100層を超えるNNも試されている。
- Deep Learningの概念は古くからある：
  - 実はみんな思っていた。1980年　ネオコグニトロン（福島）、1990年　野田、他
  - 十分なデータ量が必要膨大な計算能力が必要：ノイズを加えたり接続をきったりする
  - WEBデータの蓄積とコンピュータ技術の向上（今のGPUでも100台以上）、クラウド化など、があって初めて実現できた。
- DeepMind（2014年Googleに買収された）
  - スコアを報酬としてゲームをプレイするAIをDeep Learningを使い強化学習。
  - 最初は下手、試行錯誤で習熟。インベーダゲームでは最後は「なごやうち」も身に着ける。
  - 「全く同じプログラム」で異なるゲームを学習できる。（DQN）
  - 開発したALPHAGOがプロ棋士を破った（2016/3）。（10年以上先とみられていた）

AIが人間を超える

- 画像認識（ILSCRC）：1000種類120万枚のタグ付き画像で教師あり学習。ランダムに5万枚を検証に、10万枚を判定に使用。
  - 2015年に人の認識率（約4.5%)を超えた。Microsoft/Google 3.5%。
- 音声認識：英単語誤り率
  - Microsoft 2016/10/19： 5.9% （6%以下に挑戦が続けられていた。下図9/13 6/3%）

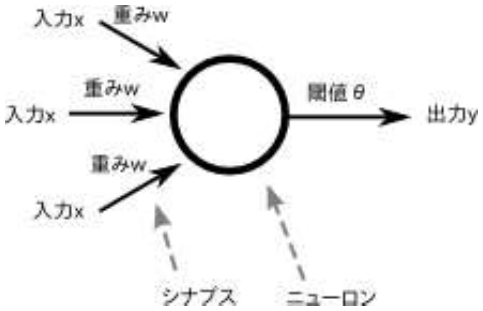


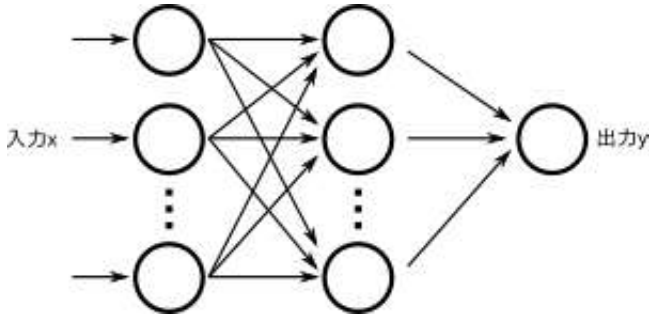
3. Deep Learningの技術解説と例

https://drive.google.com/file/d/0B04ol8GVySUubjVsUDdXc0hla00/view?pref=2&pli=1 に非常に詳しい解説があるので参照されたい。

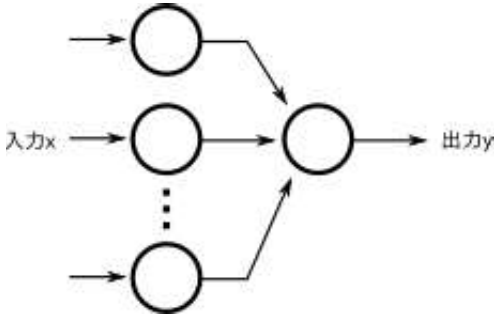
技術解説：パーセプトロン/ニューラルネットとは

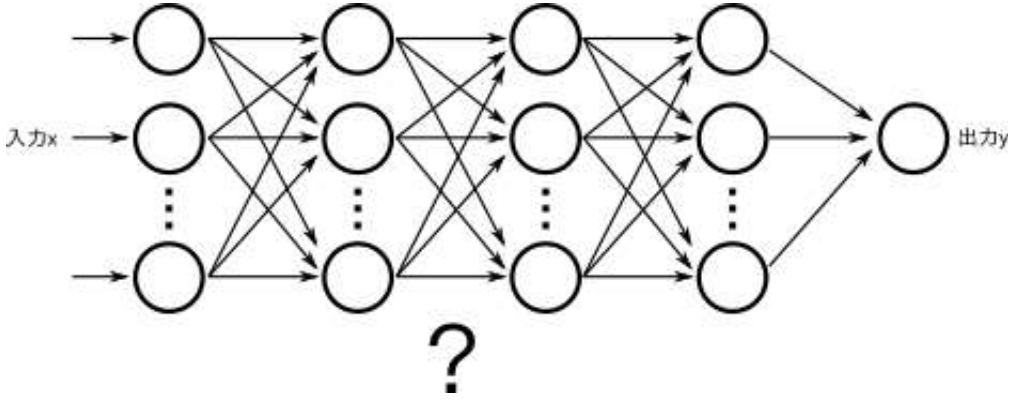
形式ニューロン　　バックプロパゲーション（3層構造）：隠れ層





パーセプトロン（2層構造）　　ニューラルネット（多層）





技術解説：ニューラルネットの学習方法

- 人工知能の学習（機械学習）は3つに大別される。
  - 「教師あり学習」：学習例とそれに対する正解を示し、NNの出力をこれに一致するようにNN内部の重みを調整する。
    - 一個の例による場合と、複数の例による場合がある。
  - 「教師なし学習」：入力（学習例）に対する正解を示さないで、コンピュータ自身が何らかの基準に基づき、NN内部の重みを調整する。
    - 例：クラスタリング（K-meansなど）、主成分分析、ベクトル量子化など
  - 「強化学習」：正解ではないが何らかの報酬を決め、それが最大になるようにNN内部の重みを調整する。（これを「教師なし学習」とする場合もある）
- NN出力の分類
  - 分類問題：対象が目的の物（事象）かどうかを判定
    - 全体を加えると1になる関数（Softmax）で、次元数だけに分類することも行われる。
  - 回帰問題：数値出力 （例えば株価）

技術解説：誤差逆伝搬法　Back Propagation（1）

- ニューラルネットのニューロン（神経細胞）は、右図1のように、加算器と活性化関数のカスケード接続で構成される。
- 加算器の出力（活性化関数処理の前）を内部状態という。（右図1のVk）　X_0はバイアスを与える。
- 活性化関数は、形式ニューロンでは、ステップ関数（入力がゼロ以上の時だけ1を出力）である。次の世代では、微分可能であることからシグモイド関数（1/2*(1+tanh(x))）が使われた。最近はReLU（Recitified Linear Unit）（ランプ）関数（max(0,x)：劣微分可能）が多く使われる。
- 3層の例（右図2）で説明する。〇はニューロン、矢印は各ニューロンに与える入力で矢印上の記号はニューロンでの加算の重みである。
- 活性化関数Sはシグモイド関数と仮定する。このとき微分 S'(x)＝x(1-x)になる。（活性化関数、コスト関数は微分可能でなくてはならない）
- ここで　入力層入力値　X_i ：i =1~3
  隠れ層出力値　Y_j ：j =1~4
  Xiに対する出力層出力値　Z_k ：k =1~3
  正解（教師値）　T_k ：k =1~3
  y, zは各ノードの内部状態、とする。

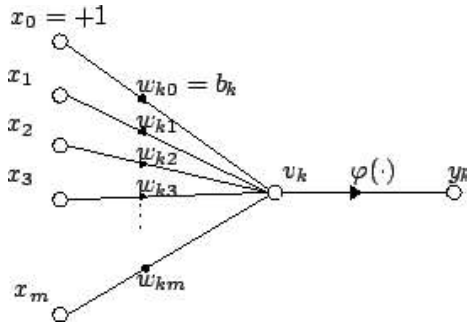


図1. ニューロンの構成

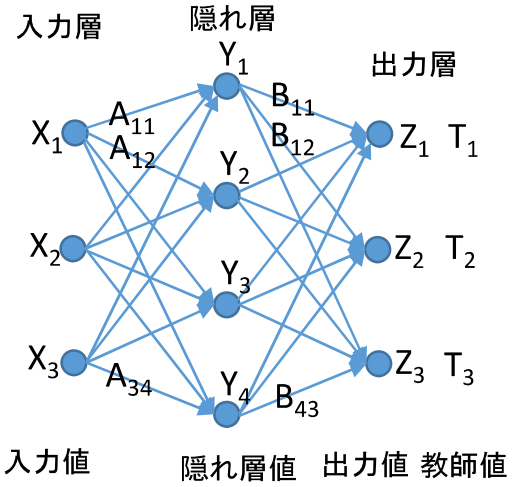


図2. 誤差逆伝搬法の例

技術解説：誤差逆伝搬法（2）

処理手順：

ある学習例のセットを入力（X_i）に与えたときの誤差「Z_kとT_kとの差」のk=1~3の二乗和（の1/2）を最小にすることを目的とする。このとき「隠れ層と出力層の結合係数」さらには「入力層と隠れ層の結合係数」に遡って分配するが（このとき一回で学習させる量を学習係数αで調節するとして）数学的に下記で厳密な解を得ることができる。

① 学習前の各ノード間の重み（層間の結合係数$A_{ij,}$ B_{jk}）はランダムに与えられる

② この結合係数での、X_i 入力時のZ_kを計算する。教師値T_kとの誤差$(Z_k - T_k)$を求める。

③ 隠れ層と出力層の結合係数B_{jk}は一回の学習で下記ΔB_{jk}だけ調整される。

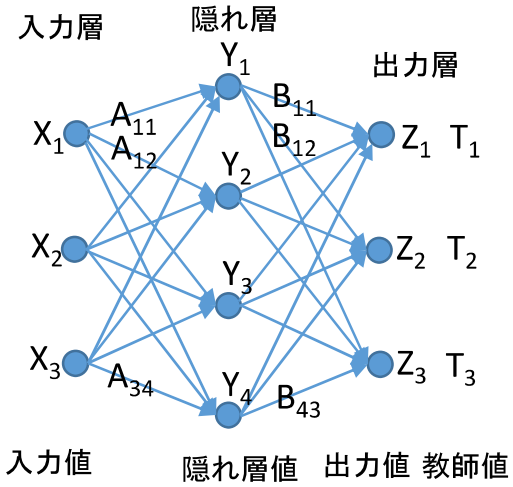
$$\Delta B_{jk} = -\alpha \cdot (Z_k - T_k) \cdot S'(Z_k) \cdot Y_j = -\alpha \cdot (Z_k - T_k) \cdot Z_k (1 - Z_k) \cdot Y_j$$

④ 入力層と隠れ層の結合係数A_{ij}は一回の学習で、
隠れ層出力Yjがつながる
全ての出力先（この例では3個）の誤差を各枝に
寄与させた合計値W_jを使って、下記ΔA_{ij}だけ調整される。

$$W_j = (Z_1 - T_1) \cdot Z_1 (1 - Z_1) \cdot B_{j1} \cdot z_1 +$$
$$(Z_2 - T_2) \cdot Z_2 (1 - Z_2) \cdot B_{j2} \cdot z_2 +$$
$$(Z_3 - T_3) \cdot Z_3 (1 - Z_3) \cdot B_{j3} \cdot z_3$$ （要再確認）

$$\Delta A_{ij} = -\alpha \cdot W_j \cdot S'(Y_j) \cdot X_i = -\alpha \cdot W_j \cdot Y_j (1 - Y_j) \cdot X_i$$

⑤ 複数（多数）の学習例で②~④を収束するまで繰り返す。



技術解説：Auto Encoder

- 人工知能の学習（機械学習）は3つに大別される。2006年ジェフリー・ヒントンらによって提案されたNNの学習方法。
  - 最初の層に入力を与え、次元数を減らした隠れ層を元の層と同じ次元に戻して、最初の層と隠れ層の間の重みを決定する。教師なし学習である。
  - 圧縮・伸長処理に似ている。これを層の数だけ繰り返す。

図3-16
一章で紹介したオートエンコーダ

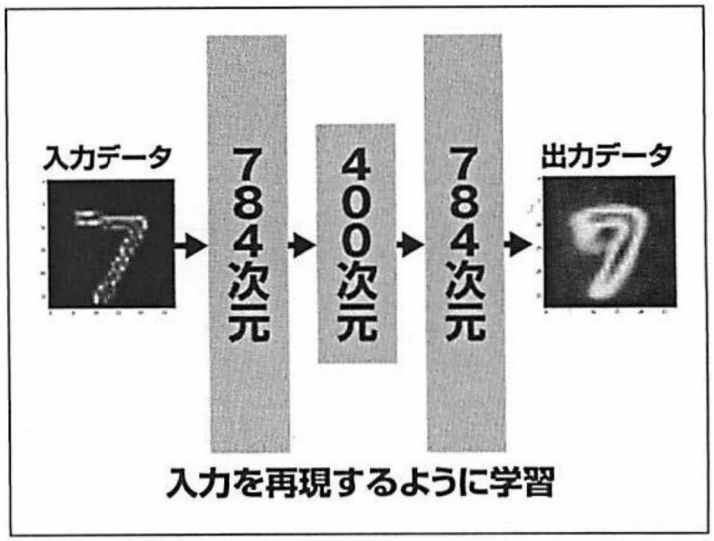
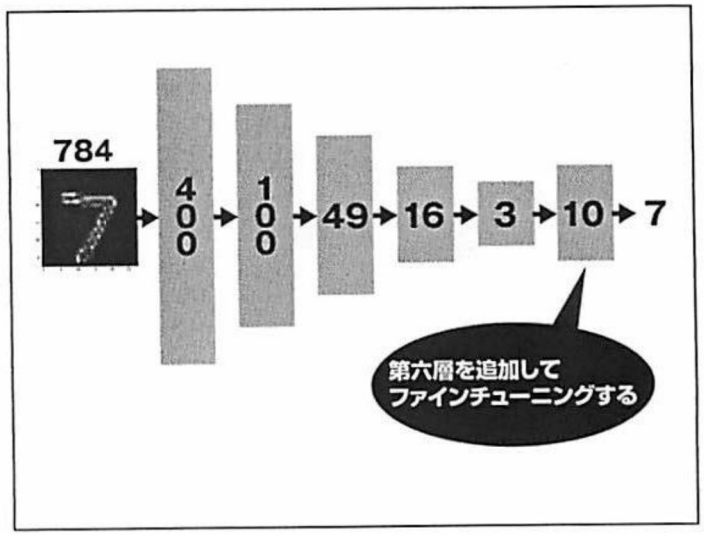


図3-20
ファインチューニングの仕組み



技術解説：確率的勾配降下法

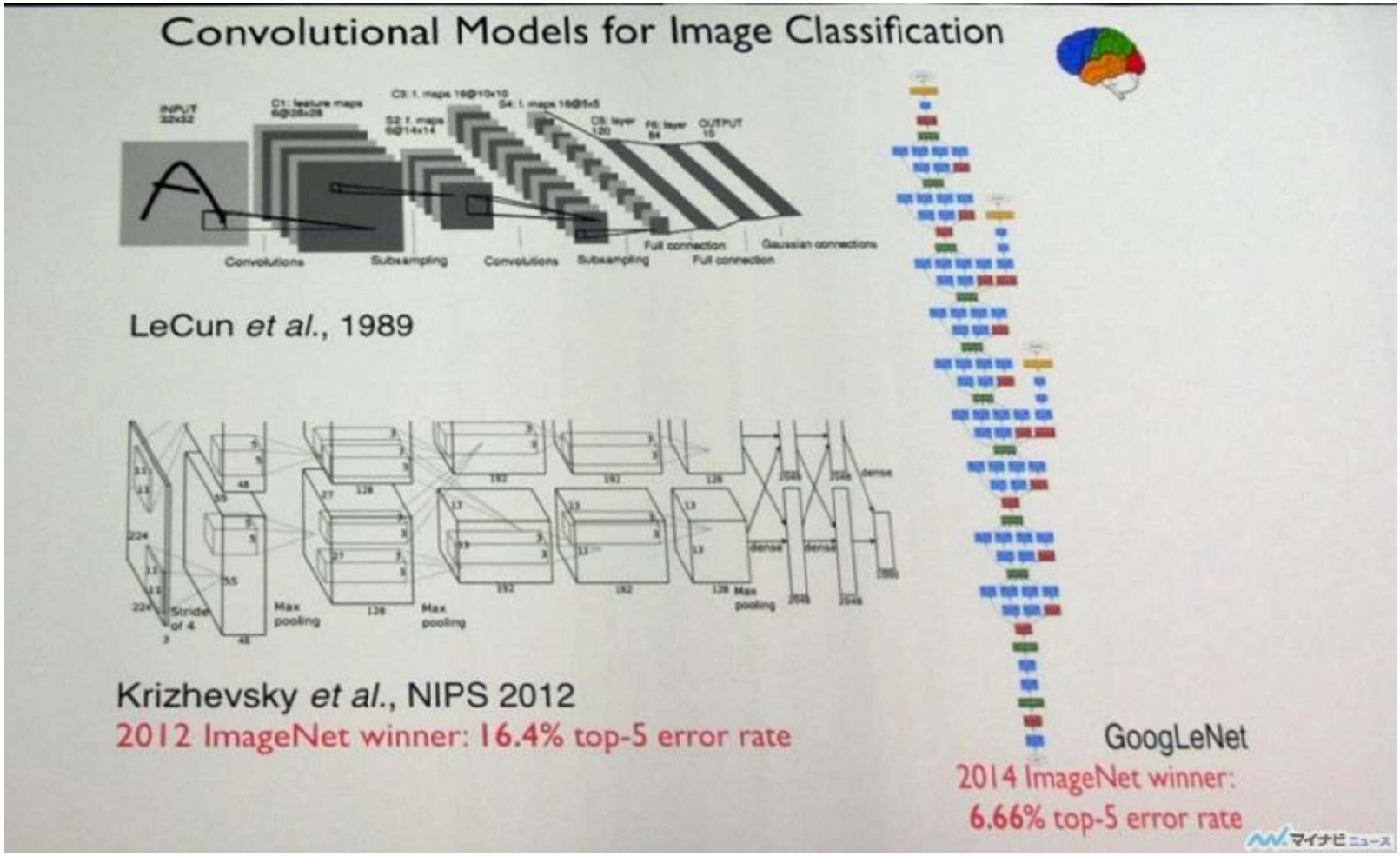
- 最急降下法による誤差逆伝搬は（3層から4層が限界で）、ユニットが多くなり隠れ層が多くなると、収束が遅い・しない、不安定になる、過学習が起こるなどの問題が発生する。
- この問題を解決するために、深層学習では、確率的勾配降下法（Stochastic Gradient Descent：SDG）が使われる。
  - コスト関数（誤差関数）を（例えば）平均二乗誤差（二乗和を学習例のセットの数の2倍で割ったもの）として、これを最小にするように重みやバイアスの調整が行う。このときn個の全学習例からm個（n＞＞m）がランダムに選ばれる。 m個の学習例が選ばれる度に重みの修正が行われる。
- さらに①学習係数の減衰、ドロップアウト（ランダムに隠れ層のユニットを削除し、あとで重ね合わせる）、②L1/L2正規化（マトリックスの対角線に情報を集中させる）、③人工的な学習例の拡張（Data Augmentation）、④クロスエントロピーなどコスト関数の改良、⑤モーメンタム・AdaGrad、Adamなどの正則化により収束を助ける。
- 以上のように、入力、出力をベクトルとして扱うと全体の計算はアダマール変換の拡張になり、GPU等による高速化が可能となる。（これがDL発展の一つの背景）

技術解説：NNの構造

- NNの構造には以下のような種類がある。（代表的例）
  - CNN：Convolutional Neural Network
    - 画像データの認識エンジンがルーツ。通常、全結合層を持つ。
    - Fully Convolutional NNでは全結合層を持たない。元のデータに戻して比較可能。
    - 生成敵対学習（GAN：Generative Adversarial Network）の一つDCGAN（Deep Convolutinal GAN）では画像を別の画像におきかえるため逆畳込み演算を行う必要がありFully CNNを使う。
  - RNN：Recurrent Neural Network
    - 音声データの認識エンジンがルーツ。
    - 連続的なデータを連続的に（一個前を順に記憶しながら）処理する。
    - RNNにもDeep（Bidirectional）RNNがある。LSTM（Long short-term memory）もRNNの拡張。
- CNN構造の図表現
  - 各段の構成は、要素数（縦）x要素数（横）x次元数（特徴量）で表現される。3階層Tensor。
  - Convolution（畳込み）：4角錐で表現。底面内の全次元のデータに（最初はランダムな値から学習させる）重みを掛けて（活性化関数を通して）次段に出力する。
    - 5x5：底面の一辺のデータ数が5個であることを表す。次元数は1024などという例もある。
  - Pooling：データを間引く（Max Pooling：最大値、Average Pooling：平均値）、データ数が縦横半分（場合によっては1/4、それ以上）になる。
  - 全結合層：通常１次元ベクトルで表現される

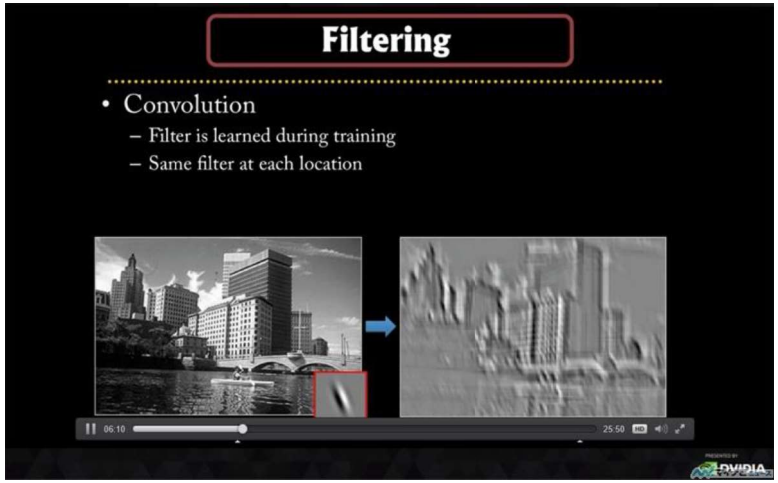
深層学習の例：画像認識（1）

左上は、1989年にLeCanが開発した手書き郵便番号の読み取りシステム。
左下は、2012年のImageNetコンペで優勝したAlexNet。
そして、右側は2014年のImageNetコンペで優勝したGoogleNetである。

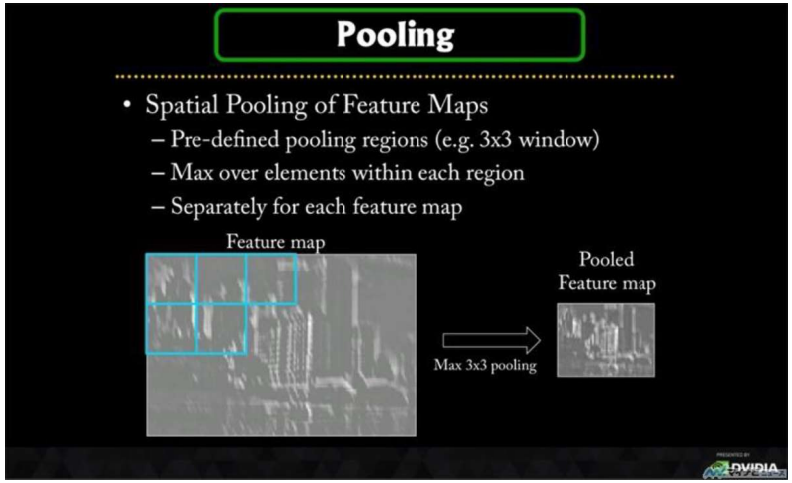


深層学習の例：画像認識（2）

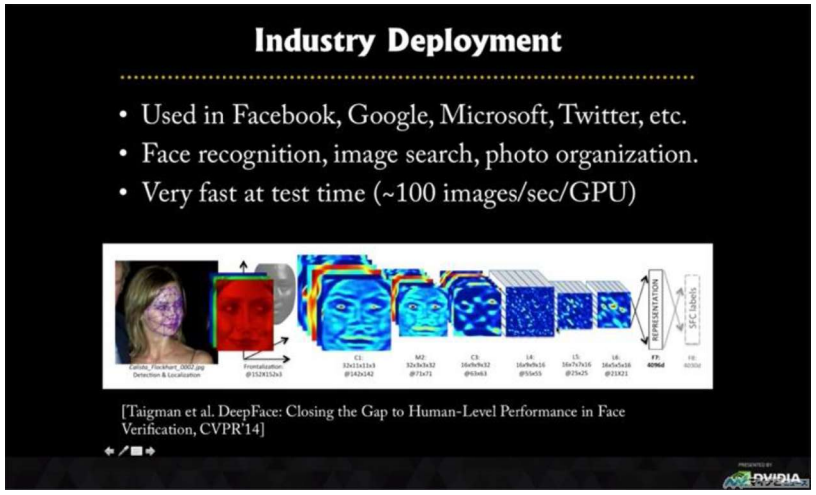
第一層　CONVOLUTION FILTER



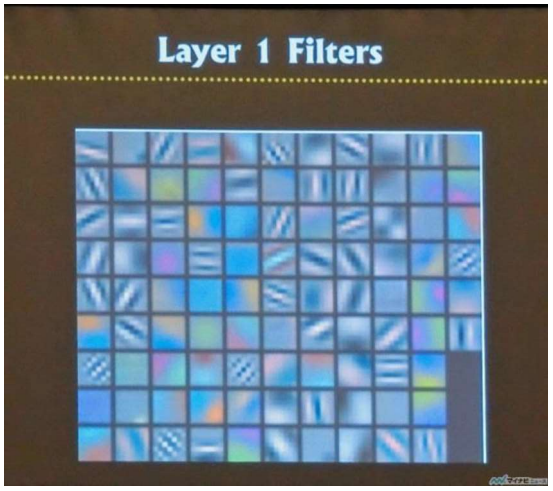
左＋ Pooling



C1/M2/C3/L4/L5/L6

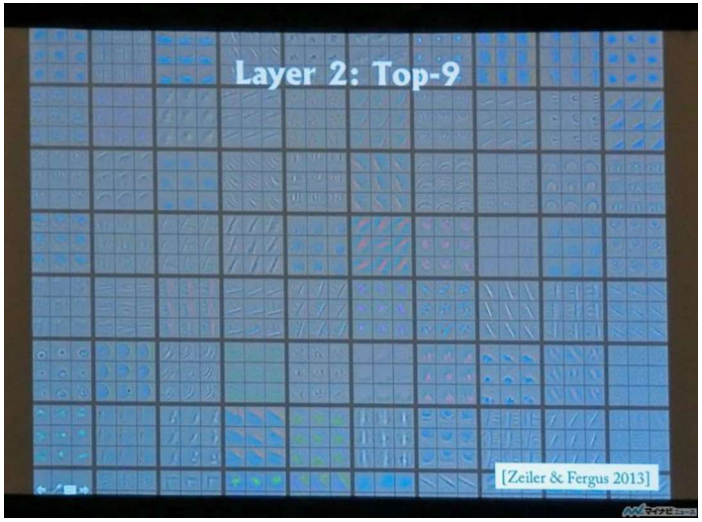


第一層のフィルタが適合するパターン

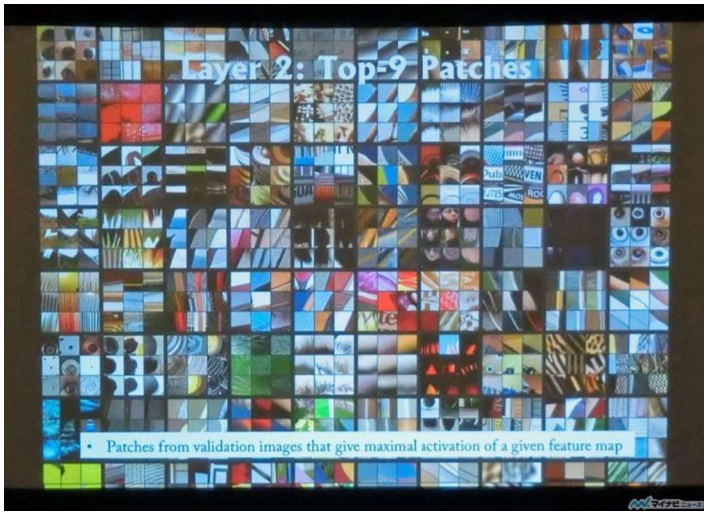


深層学習の例：画像認識（3）

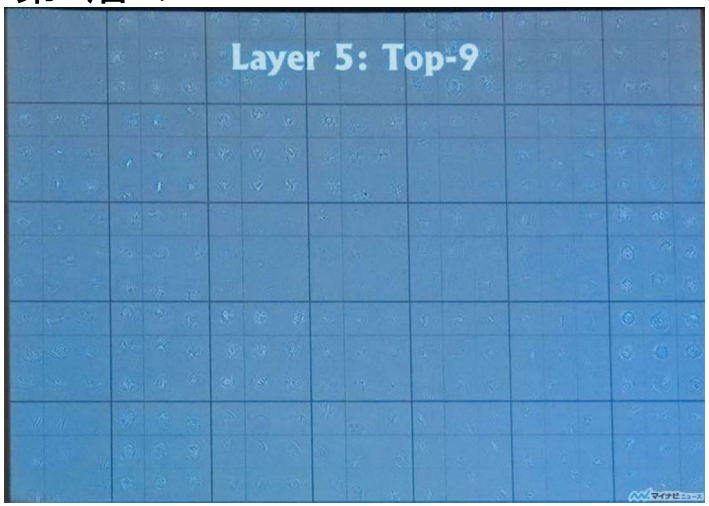
第二層に適合するパターンの各TOP9
縞や大きな点などを抽出している



第2層の適合するTop9の元画像



第5層のTOP9



第5層の適合するTop9の元画像

深層学習の例：ALPHAGO（論文発表時）（1）

- ニューラルネット構造
  - 入力層：19x19のマス目に対して49個の0/1（白、黒、空、手数、、、、）：ポリシィでは48個
  - ポリシィネット隠れ層：（19x19x192）x12層（ReLU関数）
  - バリューネット隠れ層：（19x19x192）x12層（ReLU関数）＋（19x19）x1層（1x1フィルタ＋ ReLU関数？）＋256 x1層（全結合＋ReLU関数）
  - 共通：第1層：23x23に対して（5x5）の畳み込み。ReLU関数。
  - 共通：第2層から第12層：21x21に対して（3x3）の畳み込み。ReLU関数。
  - ポリシィネット出力層：（1x1）フィルタ？、座標毎に異なるバイアス、Softmax関数（19x19の総和=1）。
  - バリューネット出力層：全結合＋tanh関数。
- ニューラルネット（主にGPUに実装）のほかに、40並列（推定）のCPUが使われ下記を処理。
  - 対戦時（意味理解できていない）：各CPUは、並列で、現在の盤面（ルート）から評価値が高い手を選びながら、探索木を下がり、リーフ（探索木の端）まで来たらバリュー計算を依頼しロールアウト（最後までランダムに打ってみること）させる。勝敗がついたらルートまで伝搬。リーフの評価回数が閾値を超えたらリーフを展開しポリシィ計算を依頼。

深層学習の例：ALPHAGO（論文発表時）（2）

- ポリシーネットワーク（これから打つ手を想像する）の教師あり学習
  - 16万棋譜2940万盤面中2840万盤面（x対称性8）を訓練データとして確立的勾配降下法を実施。訓練データは16個をランダムに選択。学習率：初期値0.003（8000万回毎に1/2）
    - 50GPUで3週間
  - さらに学習済みネットワーク同士を対戦させて強化学習。128回対戦の結果を報酬とし確立的勾配降下法を実施。これを1万回繰り返す。：50GPUで1日
- バリューネットワーク（どちらが優勢か判断する）の強化学習
  - ポリシーネットワークの結果を使ってランダム打ちでロールアウトさせ強化学習する。
  - 訓練データは一様乱数（1～450）で手数Uを決定。U手目の打つ場所を一様乱数できめる。強化学習済みのポリシィをつかってロールアウトし勝ち負けを決定。
  - 32個の訓練画面で確率的勾配降下法を実施。これを5000万回繰り返す：50GPUで1週間。
- 対戦（論文時：48CPU＋8GPU、探索スレッド40：推定）
  - 大局観と学習済み局面の両方を実現
  - 現在の盤面をルートとして探索木を構成。自分の手番でシミュレーションを繰り返し最良の手を出力。
  - ファンフィとの対決時には「CPU　1,200　と　GPU　約170枚」が使われた。20倍強の規模。

4. Deep Learningまとめ

Deep Learning：専門家の介在の必要性

- Deep Learningは機械学習の自動化をもたらしたが、学習させるニューロネット自体の構造は専門家が考えCut&Tryする必要がある。まだまだ「たくさん試したもん勝ち。」
  - 構造：畳み込みNN（CNN）、再帰型NN（RNN)、Deep Belief Networkなどなど。
  - 以下のような設計項目がある。さらにこれらにハイパーパラメータがあり自由度がある。

  ①学習させるデータをどう定義するか、

  ②何層にするか、

  ③各層のコンボリューション範囲（サイズ）、どれだけPoolingする（次元を減らす）か、

  ④どんな活性化関数をつかうか：次段におくるデータを制約する
  古くはシグモイド関数、最近では、ReLU（ランプ関数）、Maxoutなど：

  ⑤変数選択の方法、規模

  ドロップアウト、スパースコーディングなど：次元を減らす。

  ⑥確率的勾配降下法の選択数、事前学習の方法など

  ⑦学習データの選択、ノイズなどランダムデータの導入等

- Deep Learning特有の問題（勾配消失、過学習など）がおきたときには構造、ハイパーパラメータ、変数選択などを見直す必要がある。

Deep Learning：今後の発展

① 画像
画像から、特徴量を抽出する
画像認識の精度向上

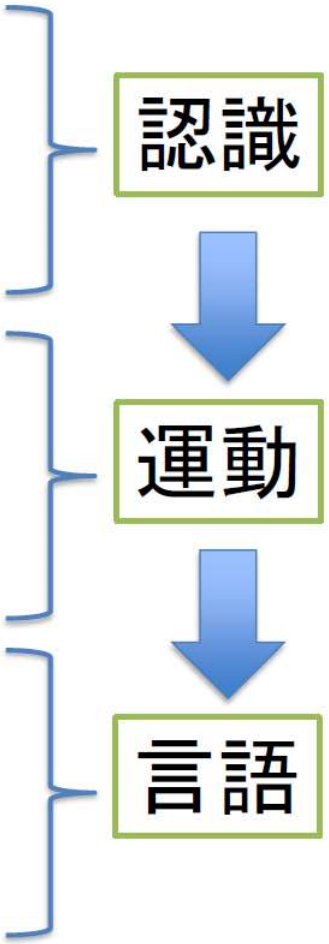
② マルチモーダル
映像、センサーなどのマルチモーダルなデータから
特徴量を抽出し、モデル化する
動画の認識精度の向上、行動予測、異常検知

③ ロボティクス（行動）
自分の行動と観測のデータをセットにして、特徴量を抽出する。
記号を操作し、行動計画を作る。
プランニング、推論

④ インタラクション
外界と試行錯誤することで、外界の特徴量を引き出す
オントロジー、高度な状況の認識

⑤ 言葉とのひもづけ（シンボルグラウンディング）
高次特徴量を、言語とひもづける
言語理解、自動翻訳

⑥ 言語からの知識獲得
グラウンディングされた言語データの大量の入力により、さらなる抽象化を行う
知識獲得のボトルネックの解決



ディープラーニングがすごいというより
その先に広がる世界がすごい

松尾：人口知能の未来-ディープラーニングの先にあるもの（総務省資料　P.16 ）

5. Deep Learning Hardware/Software

Deep Learning：ハードウエア

- 機械学習時には（膨大な学習条件をこなすため）膨大なリソースを必要とする。
  - 実時間ではないのでクラウド化・分散化も不可能ではない。
  - 実際には（分散化された枝の処理全部の終了をまち相互依存データを交換する必要があるので）ネットワークで密に結合された（一か所におかれた）リソースが使われる。
  - Deep Learningでは、（TOP500でのFP32のような）処理能力は必要なく、FP16（さらに固定小数点化も可能）で十分。専用ハードウエアも登場してきている（後述）。
- 実行時（推論 Toleranceという）には（短時間での反応のために必要なだけの）リソースが必要。
  - ただし学習済みのNNの反応を得るだけなので、データ量も演算速度も膨大ではなく、エンベット化も可能。
  - （後述するように）マルチスレッド・マルチNN対応能力が必要なる（はず）。
- 「AIエンジン・AIチップ」は2つに分化するという見方がある。
  - クラウド側：（メインフレーム時代のような）垂直統合型のソリューションへ
  - エッジ側：（低消費電力化された）GPU、CPU、DSP、ASSP、MPU、専用（脳型）チップなどに水平分業化される。

AI エンジン/チップ：サーバソリューション(1) nVIDIA

- 新GPUアーキテクチャ PascalでDeep Learning対応 （2016/4/5発表）
  - 半精度浮動小数点演算回路をCUDA COREに追加
    - 性能はFP32の2倍（但し2 Word PACKでSIMD処理）。
  - Tesla P100 搭載 GP100 ：15.3G Transistors。610㎜^{2}(≒25mm□) 16nmFINFET
    - 性能 21.2TFLOPS（FP16） ＝ CUDAコア3584個x4（FP16積和4個/Core）
      x1480MHz （パイプラインスループット Boost時：実売品では1328MHz）
    - HBM2（16GB（4個x4枚スタックx8Gb）/GPU＠720GB/s）とGP100をSiインタポーザで積層
- DGX-1 （Deep Learning 専用ラック型コンピュータ） （2016/4/5発表）
  - 169.6 TFLOPS（FP16 Peak） ： Tesla P100 を8台搭載
  - NVLINK Hybrid Cubu Mesh＋DUAL 10GbE＋Quad InfiniBand 100Gb 7TB
  - 3U Power/3200W 1400万円
- GPU推論用アクセラレータカードTesla P40 /P4 （2016/9/13発表）
  - P100 ：3584 CUDA CORE 21.2 TFLOPS/FP16 16GB（HBM2）250W
  - P40 ：3840 CUDA CORE 12TFLOPS/FP16 47 TOPS/INT8 24GB（GDDR5）250W
  - P4 ：2560 CUDA CORE 5.5TFLPOS/FP16 22 TOPS/INT8 8GB（GDDR5） 50W
- 次世代VOLTA CORE開発中：10nm世代 対PASCAL性能比 1.8倍

AI エンジン/チップ：サーバソリューション(2)

- Google：Tensor Processing Unit（TPU）：2016/5/18 Google IOで発表
  - サーバラックのHDD SLOTに挿入して使う。Jupiterネットワーク（10GbEx100KPort収容可能、合計1.2Pbpsトラフィック）に接続する（想像）。
  - FPGA BASE。Wあたり機械学習実行効率を最適化。ただしチップ販売の計画無。
  - Deep Learning TensorFlowアルゴリズム（後述）に特化。
  - 半精度浮動小数点（それ以下？）に最適化（8bit MACHINEだという報道もあった）。
  - Google音声認識サービス、Cloud Machine Learningサービス、ALPHAGOに利用。
  - ASIC化中との情報有。（EDGE用にもつかえる？）
- Microsoft：自社FPGA BOARD（Catapult）に実装。
- Nervana Systems：
  - 独自学習アクセラレータNervana Engine開発中（TSMC28nm GPUサイズ）2017年初頭
  - Nervana社クラウド（GPUとNervana Engineを搭載）でサービス提供。
- Deep Insites（PEZY子会社）：AI ENGINEを2017年末までに開発する。
  - 単精度（32bit）FP、半精度（16bit）FP：8bit、4bit、2bit、bit演算も視野内に。
  - 1積層チップ：100万コア、100TB/s、DRAM一体、100W：サーバ側、SCチップとは別。
  - チップ間伝送を磁界結合で実現しシナプス結合の重みづける。アナログ結合も？
  - ノイマン型。（別に非ノイマン型脳型チップの開発を考えているらしい。）

AI エンジン/チップ：エッジ側

- Movidius：組込みアプリケーション用チップ独自開発中 （Deep Learning on a stick）
  - Googleと提携。既存の画像処理用CHIP「Myraid2」がNNの実行に向く。
- TeraDeep：組み込み用FPGA SOLUTIONを提供。
- Cadence：Tensilica（DSP）の命令セットを拡張。
- nVIDEA：GPUの大幅な低消費電力化にMITと取り組み中。
  - CNNを低電力で実行できる専用エンジンを開発中
- Qualcomm：Snapdragon820向けSDK提供予定（2016年後半）
  - Zeroth NPU（Nerural Processing Unit）の（機能を）プラットフォーム化
- RENESAS：R-INエンジンにクロスコンパイル・インテリジェンス開発のコアを導入。
  - クロスコンパイル・インテリジェンス：東工大発ベンチャー。7人。

この他に：

- SBのARM買収：「AIチップを開発するため。」ではないか？という観測がある。
- Samsung：米国オースチンの研究所を拡張して人工知能チップ開発に注力？

nVIDIAのEDGE側SOLUTION

- JETSON TK1：組み込み向け 開発者用KIT
  - TegraK1 SoC搭載：Quad Core ARM A15 、KeplerコアGPU（192 CUDAコア）
  - 10W。 1 TFLOPS。
- Parker SOC（HotChips2016で発表）
  - CPU：64BIT ARMライセンストコア（Denver）x2個＋Cortex-A57 x4個搭載（120 SPECMark）
  - GPUコア：Pascal x2（256 CUDAコア）
  - 16nm FINFET TSMC。Safety Engine。
- Drive PX2、PX2 Autocruise ：車載向け自律走行ソリューションプラットフォーム。
  - PX2: Parker SOC 2個搭載。80W。
  - PX2 Autocuise ： Parker SOC 1個搭載。10W。1.3TFLPS。
- Xavier：Parkerの後継SOC。VOLTA CORE開発中。
  - PX2の消費電力80Wを20Wに低減（8CPU、512CUDA GPU）
- ソフトウエア・フレームワーク群を提供
  - TensorRT：16bit/32bitで学習されたNNを8ビット演算に「最適化」して推論実行できる
  - DeepStream：最大93本のHDビデオストリームをリアルタイム分析
  - DriveWorkAPLHA：自立走行ソフトウエアプラットフォーム

脳型チップ/Neuromorphic chip/ニューロチップ

- 脳型チップは、脳を模倣しニューロンとシナプスを（模擬的に）実装。消費電力を大幅に減らせる。IBMが開発した「TrueNorth」（2014年発表）を契機に各社の開発が本格化。
  - IBM「TrueNorth」：STDP（Spike-Timing-Dependent Plasticity）を実装。
    - DARPA：SyNAPSE（Systems of Nueromorphic Adaptive Plastic Scalable Electronics) Project。「TrueNorth」実装システムNS16eをLawrence Livermore National Laboratoryに納入済。
    - 16個で1600万個のニューロンと40億個のシナプスを実現。2.5WTyp。
  - 中国清華大「Tanji Chip」
    - TrueNorthによく似る。2015年。
  - Zurich大/Zurich工科大：CMOSチップでReconfigurable On-Line Learning Spikingを実現。CNNとしても動作する。
  - 東芝：SolidMind。Altera上に実現。（2016/11/8 LSI化の発表があるらしい。by日経）
  - DARPA：Cortical Processor Project。CMOSプロセスでXeonの100万倍の省電力化目標。
- 次のステップとしてシナプスの機能を受動素子で実現する試みが進んでいる。
  - DARPA UP-SIDE Projectもこの領域。
  - メモリスター（第4の受動素子）：電圧をかけるとヒステリシスが出現。
    - POSTECH/SK Hynix社他（TiN/PCMOメモリスター）
    - UCSB/DENSO他（Al2O3/TiO2メモリスター）：1㎠角に2500万個ニューロンと2500億個のシナプスを実装
  - 相変化メモリ（PCM：Phase Changing Memory）
    - IBM（2015年）GeSbTeによるPCM
    - パナソニック（2013年）FeMEM。CMOSチップで「シナプス144個/ニューロン9個」

Deep Learning処理に必要な精度

- 論文によれば、Deep Learningに必要な演算精度はFP16で十分
  - 丸めを統計処理すると（スケーリング付）固定小数点8bit～14bitでも可能。
- nVIDEA/Pascalアーキテクチャ
  - 半精度浮動小数点演算回路を追加
  - 推論処理アクセラレータ P4/P40 ：INT8対応
- TPU（Tensorflow Processing Unit）も（たぶん） FP16対応
  - 但し内部処理で8bitでRelu処理するという記述もある。
- PEZY/AI ENGINE
  - 単精度（32bit）FP、半精度（16bit）FP：8bit、4bit、2bit。bit演算も視野内に。

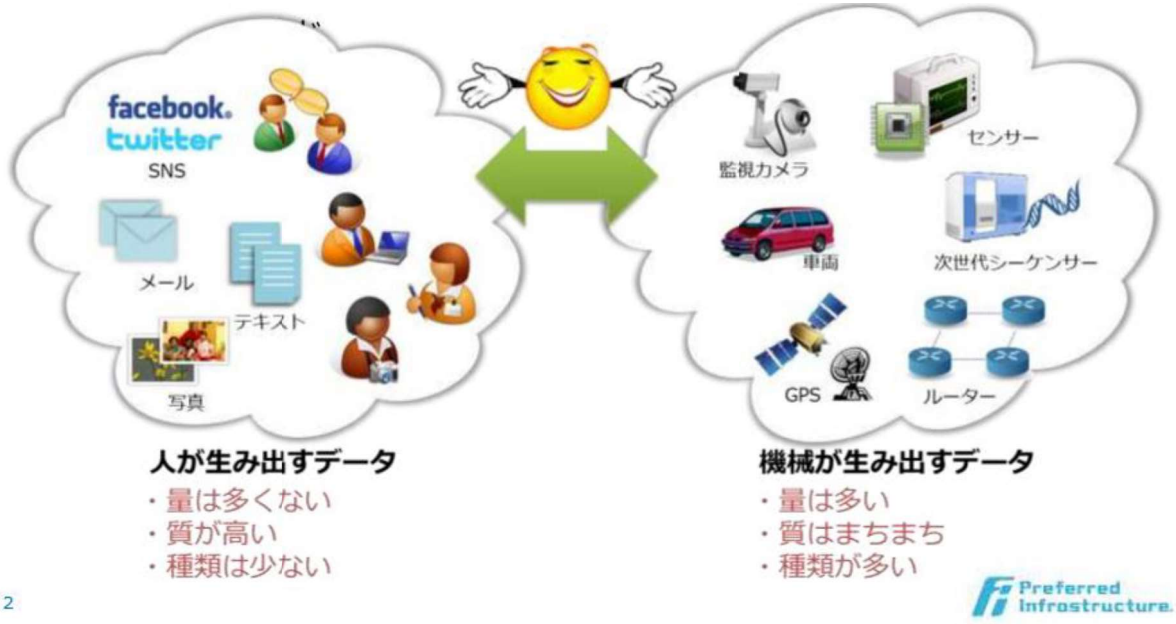
Deep Learning：ソフトウエアフレームワーク

- TensorFlow
  - Googleが2011に発表したオープンソース・フレームワーク。Apache2.0ライセンス。
  - 2015/11公開。CPU、GPU、TPUで利用可能。複数台GPU対応未公開（2016/6時点）
  - 多次元配列（Tensor）の処理フローを直観的に記述できる。
  - 学習をGPUサーバで行い、Android上で実行も可能。
- Chainer
  - Preferred Network社（日本）開発。2015/6公開。Pythonライブラリ。GPU対応。
  - あらゆるNN構造に対応可能。
  - 動的計算グラフによる直観的コード体系。複数GPUによる学習にも直観的記述可能。
- Caffe
  - UCB/BVLC開発のC++ オープンソースライブラリ。C++/Python/MATLABで利用可能。
- CTNK：Computational Network Toolkit （現Version 1.5：2016/6発表）
  - Microsoft提供。2016/1公開オープンソース（MITライセンス）。マルチGPUサポート有。
- DSSTNE
  - Amazon提供のDLライブラリ。2016/5公開。マルチGPUサポート有。スパースデータで早い。
- 他に「Torch7」（Faithbook）、「Theano+Pyleams」（Montreal大）」「Deeplearning4J 」（Skymind社）、「H_2O」、「∞ReNom」（GRID：日本）
- 半導体ベンダFW/SDK：「Intel Deep Learning Framework」、「TensorRT」（nVIDIA）、「Snapdragon Neural Processing Engine」など

6. Edge Heavy Computing

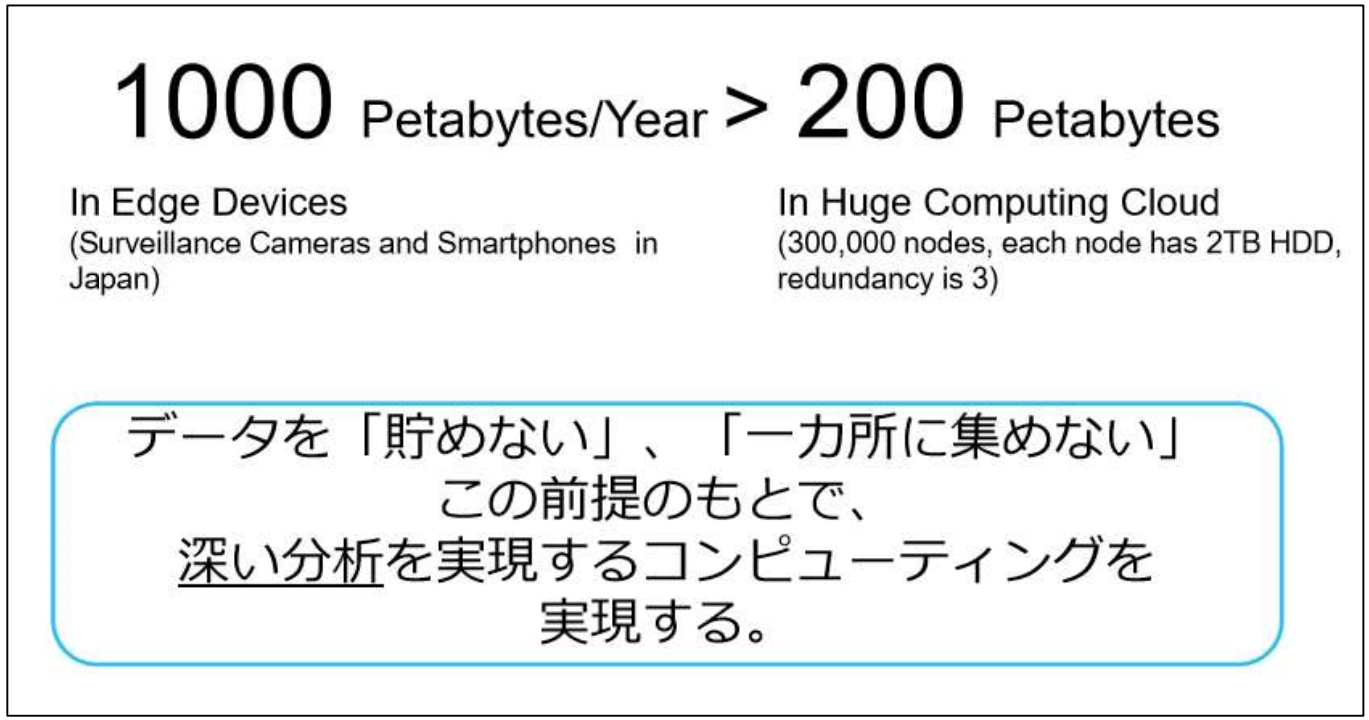
人が生み出すデータvs機械が生み出すデータ

- 今まで:(テキスト、画像、音声など)人が生み出す。
- これから:機械がエッジでデータを生み出す。
- 量が多く、種類が多く(爆発する)、質はまちまち。人が見ない、触らないデータも集められる



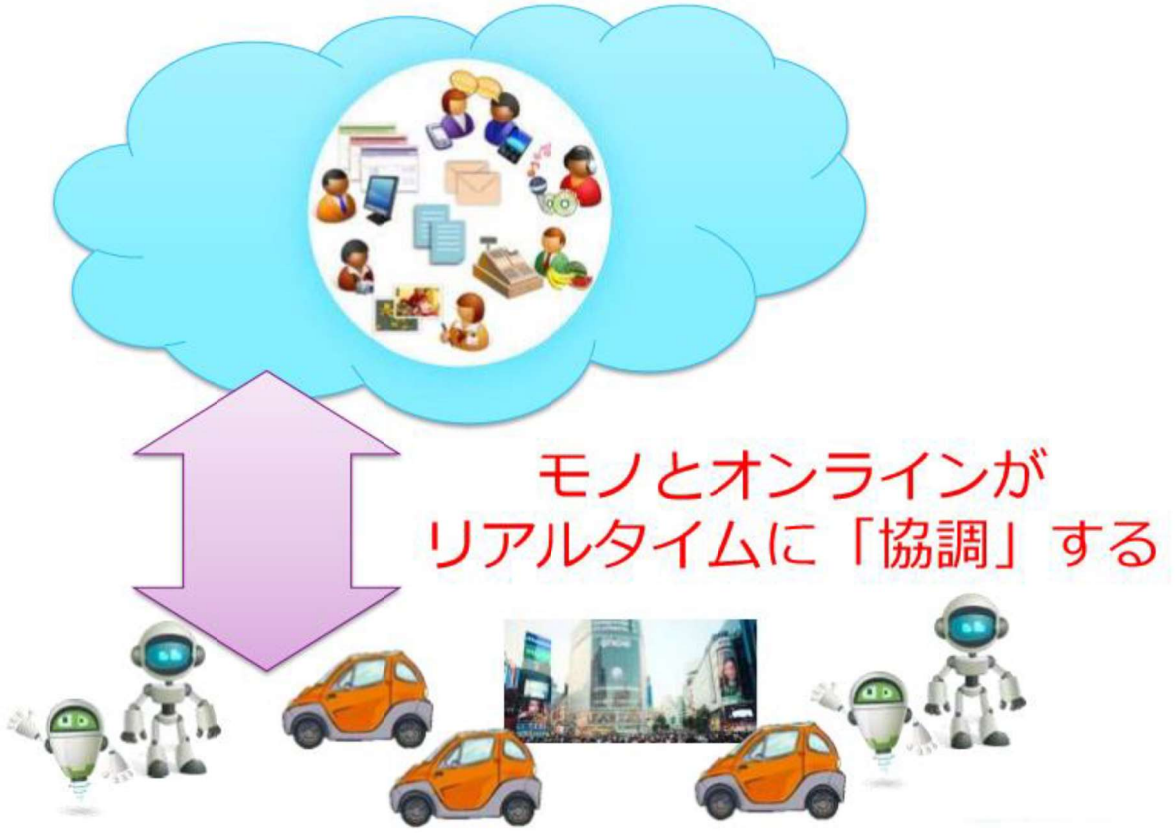
エッジヘビーコンピューティング

- （量が多すぎて）データを一か所に集めて貯める（Googleのビジネスモデル）ことはできなくなる。
- 特徴抽出も学習で獲得、生のデータから直接認識。
- 人を超えるような認識、識別、判断、予測を実現できる。
  （スーパーコンピュターが世界を変えるのではない）



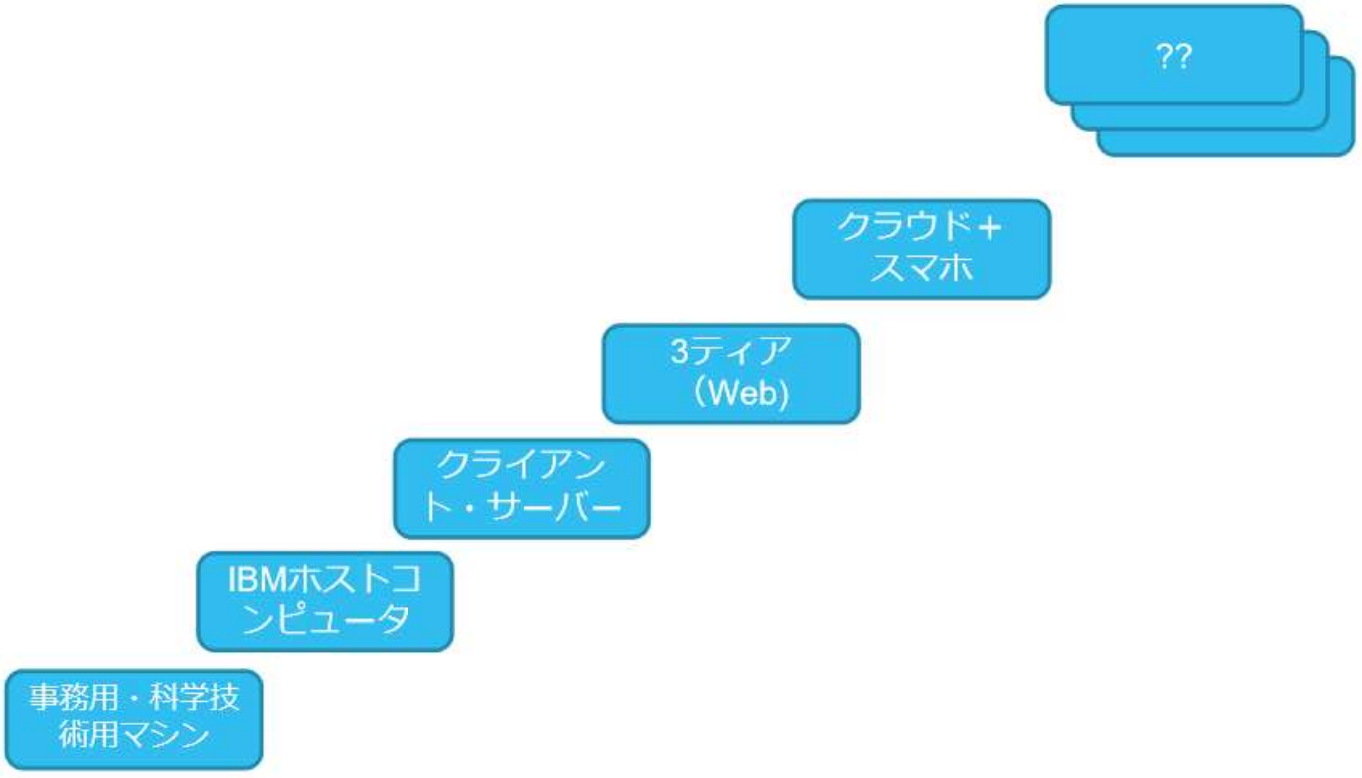
オンラインとオフラインを「リアルタイム」でつなぐ

- モノとオンラインがリアルタイムに協調



IoT時代のアーキクチャは何か？

- クラウドの次はエッジヘビーコンピューティングだ。

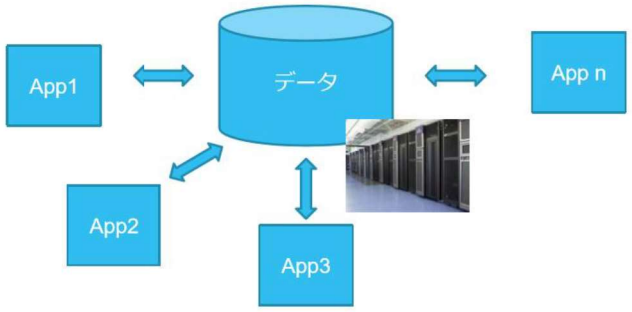


データにこそ価値がある

- ビックデータは排気データ。ビット当たり価値単価は下がっている（集めて事前分析することに意味がなくなる）。
- エッジにデータ価値の多くはエッジに存在するようになる。
- センサデータの増加⇒データ総価値の増加⇒価値密度の低下。
- 多くのデータは収集されるが利用されない。
- 利用を見込んで投機的に事前処理することは割に合わなくなる。

⇒多くのデータが収集された地点で格納・処理されるようになる。

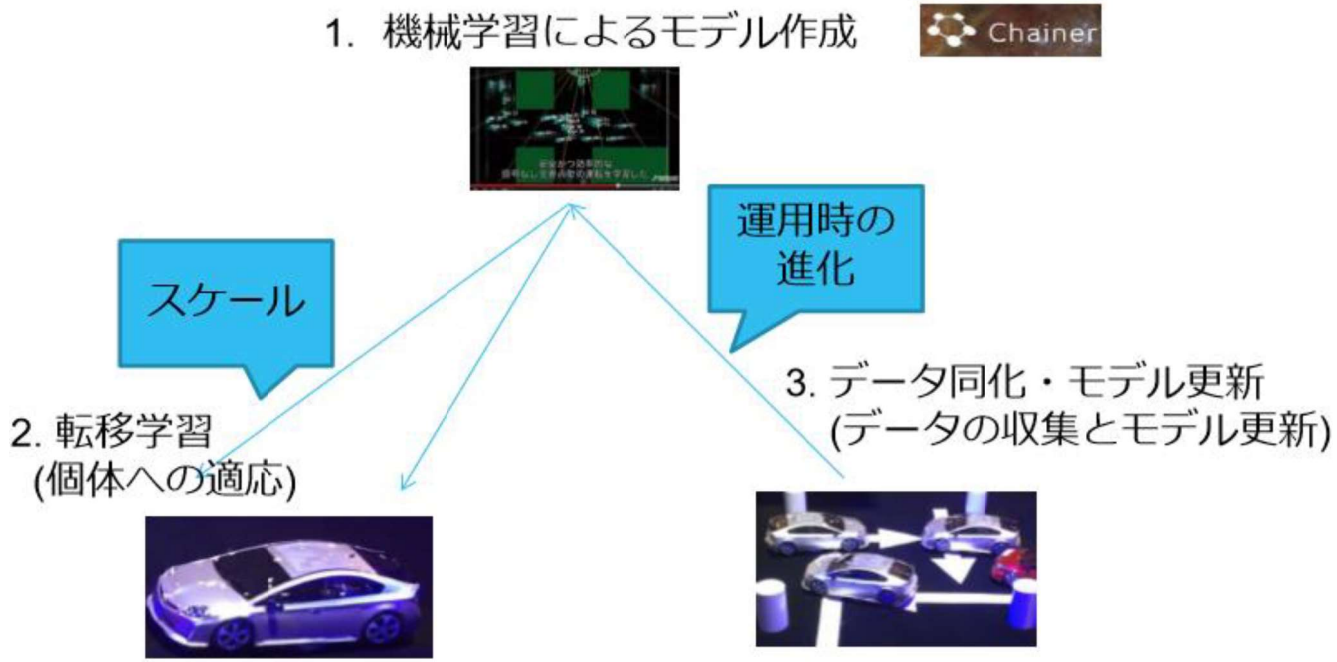
コンピュータ科学の教え： データ独立（Codd, 1970)



データこそが価値！

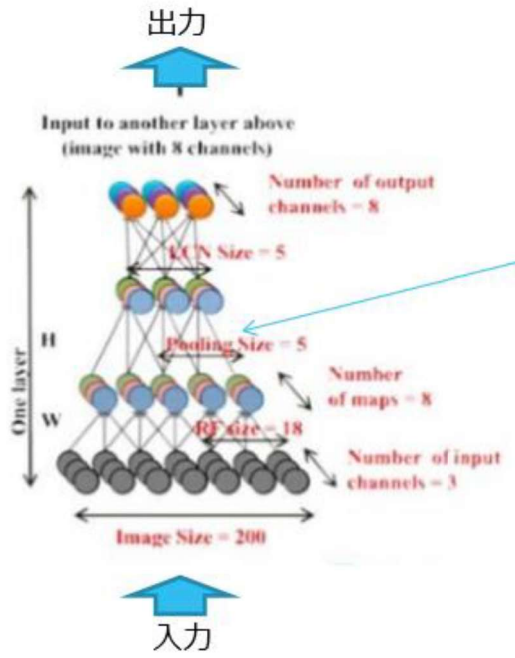
モデルを進化させる

- IoT＋MLでは、機械学習によってモデルが作成される。
- これを転移してスケールアウトする。
- 「データ同化モデル更新」がモデルを運用時に進化させる。



「学習済みモデル」に知財価値がある

- 機械学習の結果得られた「学習済みモデル」（構造＋重み）に知財価値がある。
- モデル検索、リコメンデーション、トレーサビリティ、匿名化・加工、計算資源の提供など、モデル流通の秩序形成に価値がある。



【学習済みモデル】 =
ネットワークの構造
＋
各リンクの重み

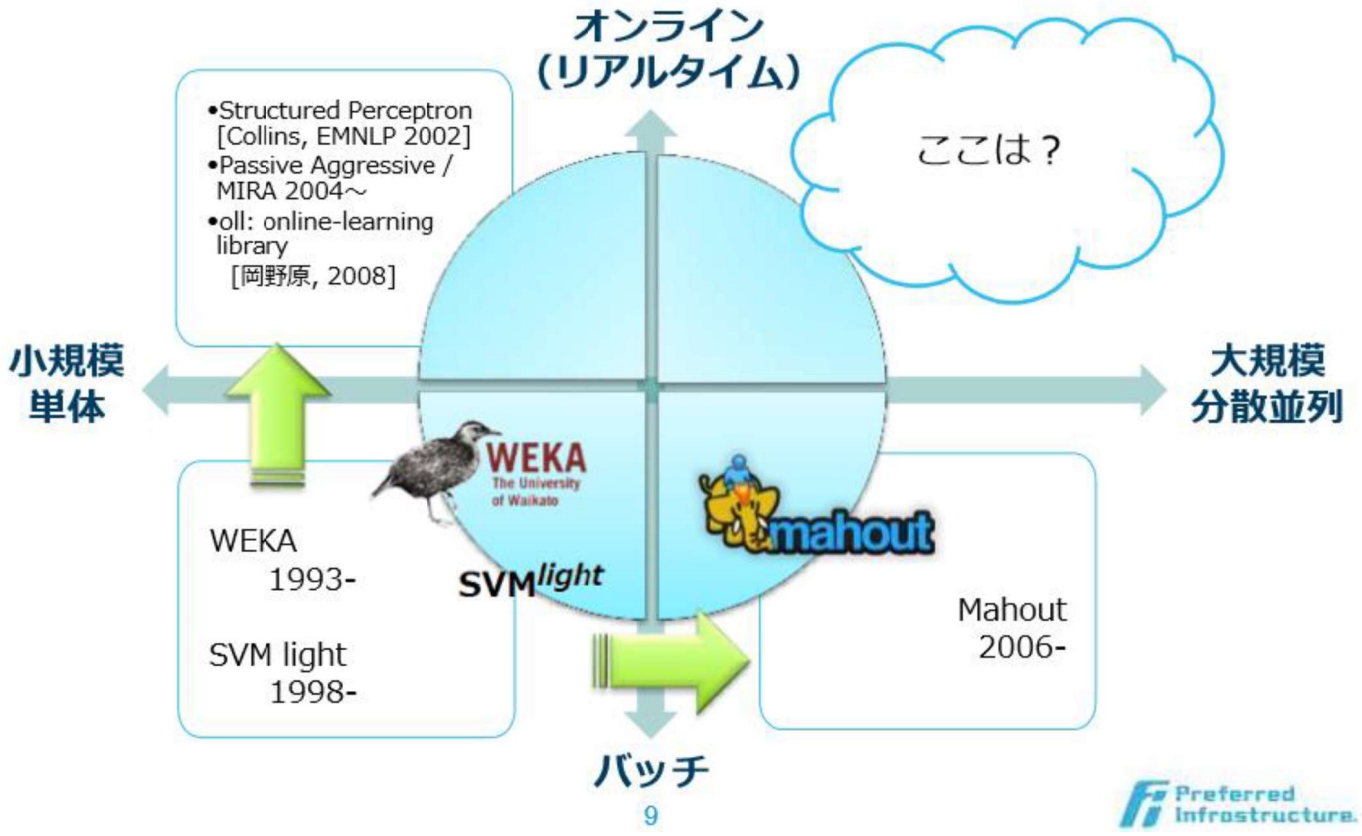
大量の数値の列

- モデル検索・リコメンデーション
- モデルトレーサビリティ
- 来歴管理（provenance）
- 決済
- 計算資源の提供(アルゴリズム、シミュレータ、計算パワーなど)
- 匿名化、加工

represent the real world. Digital Reasoning, a cognitive computing company based in Franklin, Tenn., recently announced that it has trained a neural network consisting of 160 billion parameters—more than 10 times larger than previous neural networks.

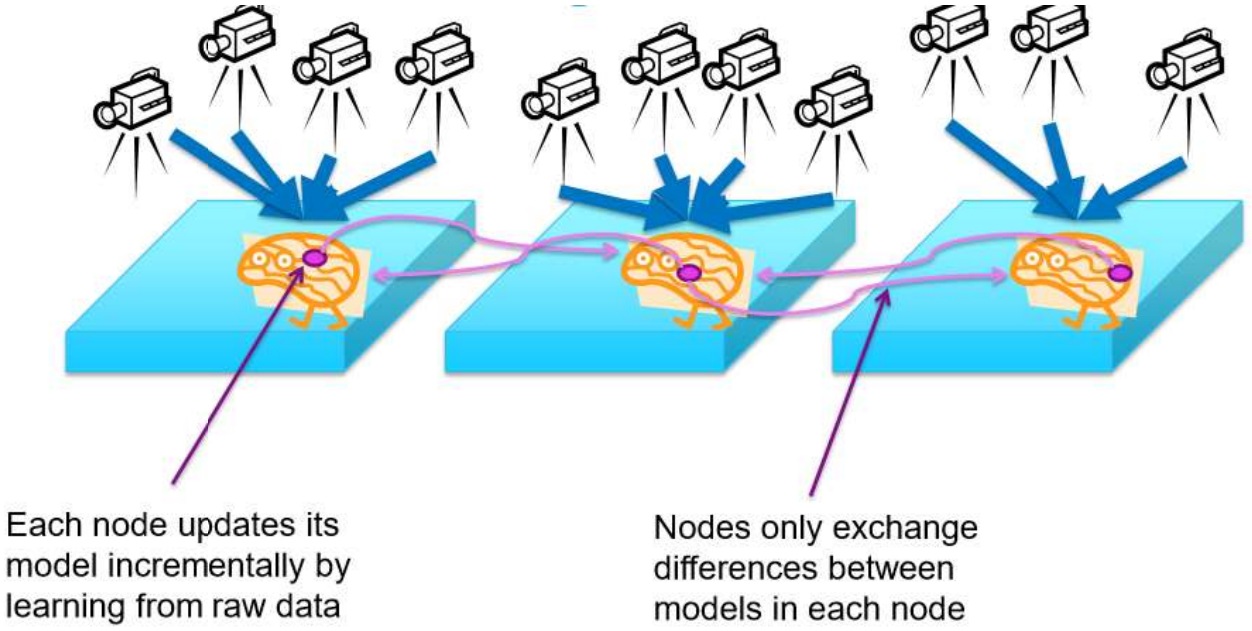
リアルタイム大規模分散処理

- リアルタイムで大規模な分散並列処理技術が必要



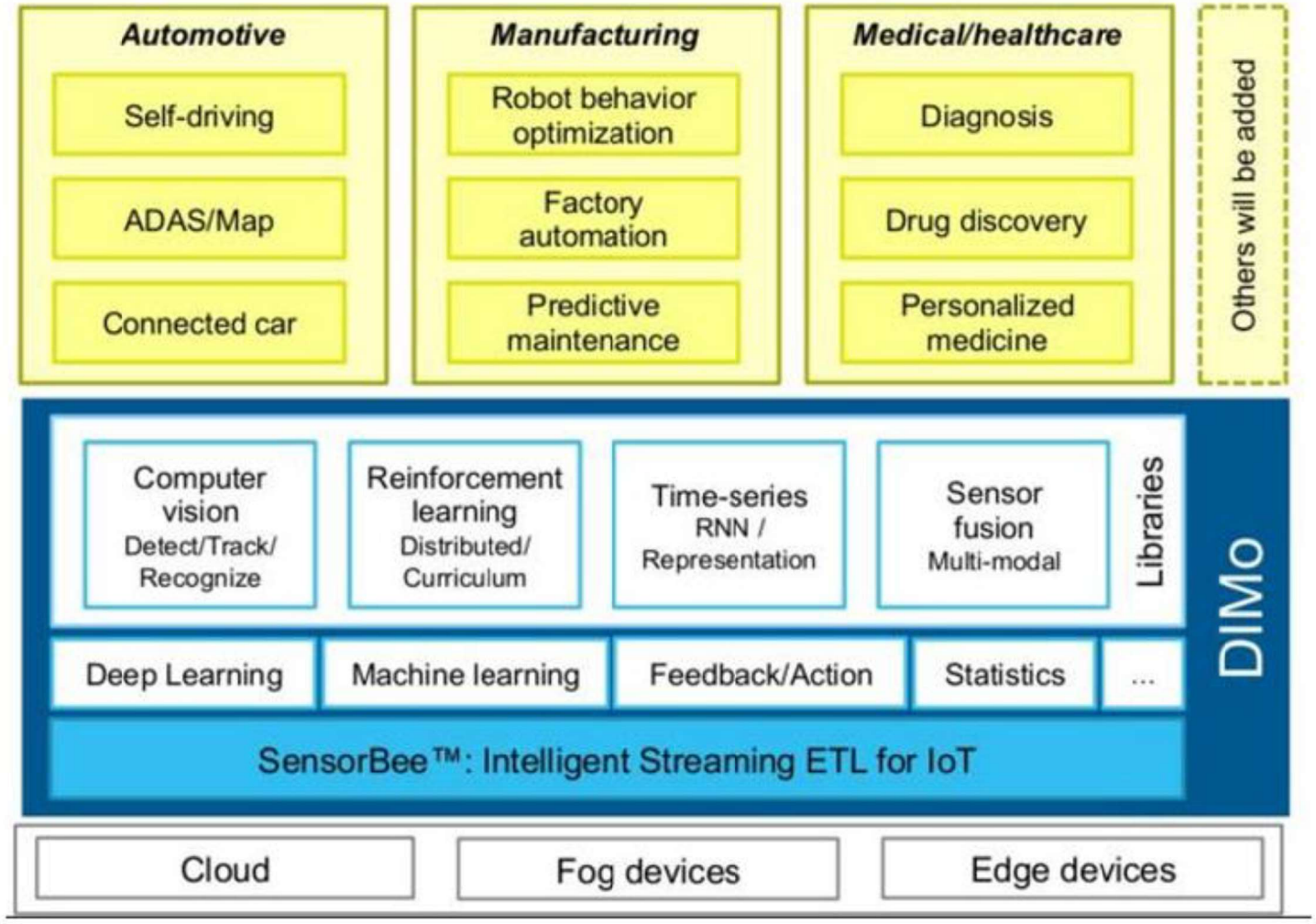
分散ストリーミング処理SOLUTION

- PFNの解：エッジノードで分散・ストリーミング処理しながらノード間で差分を交換する。



PFNの提案：分散ストリーミング処理アーキテクチャ

- Krill、SensorBee、DIMo、Jubatus



②P.10

エッジヘビーデータの例

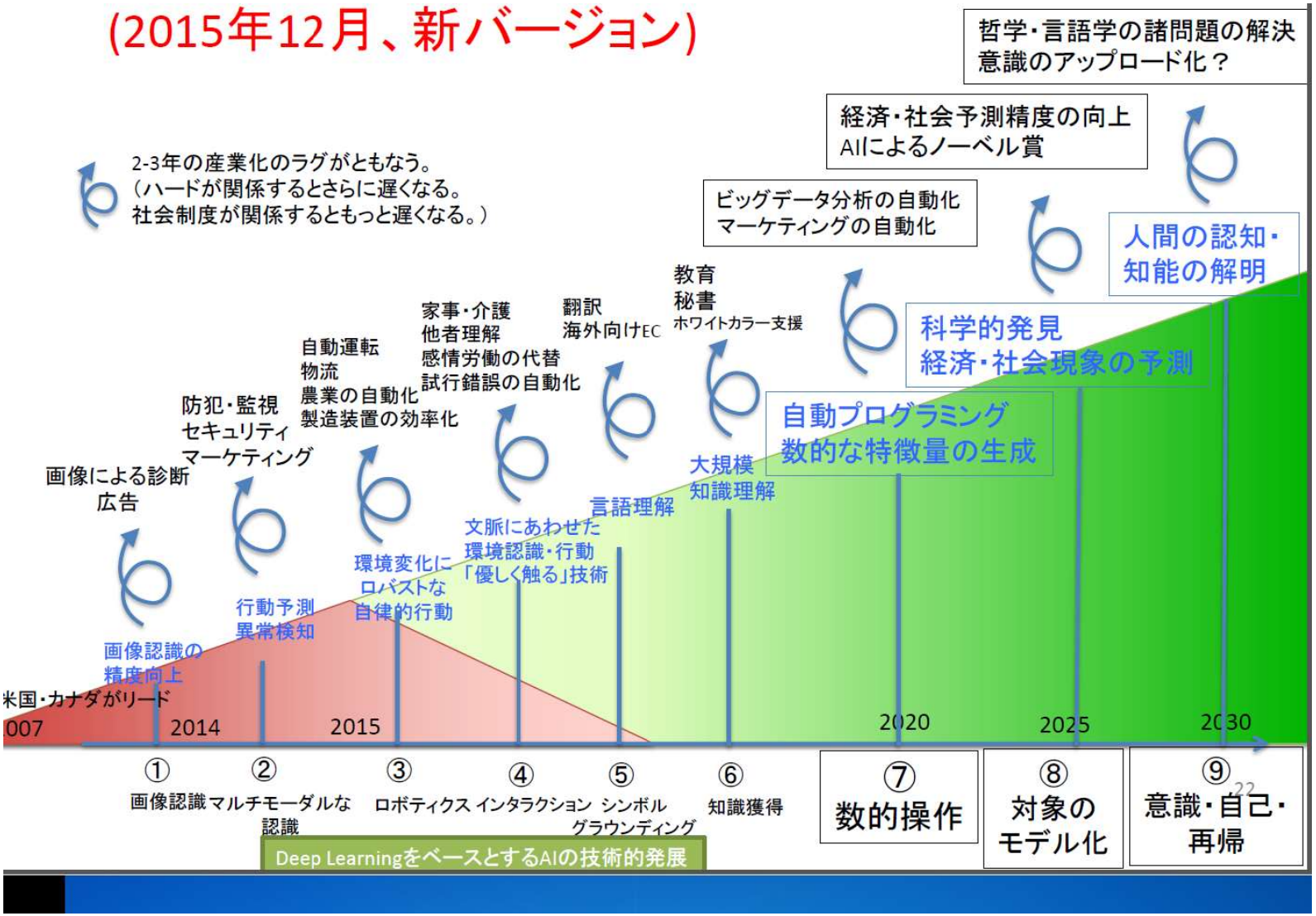
- 典型的利用シナリオ
  - 監視カメラ
    - デジタルレコーダにLOCALに保存。ネットワーク経由でサーバに送られることは通常ない。
    - 750万台x100GB/台=750PB
    - 例「周囲のカメラが連動。特定の人を追跡する際周囲のカメラにデータを送り、局所的に人物の追跡を行う」
  - 生体センサ
    - 血圧、脈拍、体温、呼吸などのデータを疾患予測や健康増進、創薬に生かす。
    - プライバシー保護上データをクラウドに集めるのは危険。
  - パーソナルモビリティ
    - 一人乗り自動車。
    - リアルタイム処理が必要。レイテンシー問題でエッジで処理要
- スマホはエッジヘビーデータ・インフラを構成できる。
  - 10GB台（少なめにみて）x2000万台（国内）＝400PB蓄積可能

EDGE HEAVY COMPUTINGまとめ

- 機械がデータを生む⇒DLで学習させる
  ⇒量が膨大⇒サーバに集められない
  - 新しいビジネスモデル！
  - 情報を一元管理することが最大の武器であるGoogleのビジネスモデルからはずれるのではないか。
- エッジでリアルタイム処理するシステムを考えるべき
  - そのための仕組みはPFNが開発している。
- エッジでの処理⇒プライバシー保護
- DLモデルを所有
  - 進化させることで差別化が可能。

7. 社会へのインパクト

技術の発展と社会への影響



松尾：人口知能の未来-ディープラーニングの先にあるもの（総務省資料　P.22）

消える職業・なくなる職業

- Oxfordの研究(2013)
- 10年で消えそうなもの
- 702業種
- 職業を性質に分解
- 9つの特性から
  - 手先の器用さ、芸術的な能力、交渉力、説得力など
- 機械学習で判定

- → なくなるのではなく、質が変わる。

主な「消える職業」「なくなる仕事」
銀行の融資担当者
スポーツの審判
不動産ブローカー
レストランの案内係
保険の審査担当者
動物のブリーダー
電話オペレーター
給与・福利厚生担当者
レジ係
娯楽施設の案内係、チケットもぎり係
カジノのディーラー
ネイリスト
クレジットカード申込者の承認・調査を行う作業員
集金人
パラリーガル、弁護士助手
ホテルの受付係
電話販売員
仕立屋(手縫い)
時計修理工
税務申告書代行者
図書館員の補助員
データ入力作業員
彫刻師
苦情の処理・調査担当者
簿記、会計、監査の事務員
検査、分類、見本採取、測定を行う作業員
映写技師
カメラ、撮影機器の修理工
金融機関のクレジットアナリスト
メガネ、コンタクトレンズの技術者
殺虫剤の混合、散布の技術者
義歯制作技術者
測量技術者、地図作製技術者
造園・用地管理の作業員
建設機器のオペレーター
訪問販売員、路上新聞売り、露店商人
塗装工、壁紙張り職人

(註)オズボーン氏の論文『雇用の未来』の中で、コンピューターに代わられる確率の高い仕事として挙げられたものを記載

Frey, Carl Benedikt, and Michael A. Osborne. "The future of employment: how susceptible are jobs to computerisation?." Sept 17 (2013): 2013. 53

松尾：人口知能の未来-ディープラーニングの先にあるもの（大学講演資料　P.53 ）

仕事の変化の予想

- 短期（5年以内）
  - 各分野でのビッグデータ、AI化が少しずつ進む
  - 特定の分野（法律、医療、会計・税務）で比較的急に進む

- 中期（5年～15年）
  - 「監視系業務」はほとんどいらなくなる。監視員、警備員。
  - 店舗の店員や飲食店の従業員、工場の作業員でも「監視系業務」はいらなくなる。
    - 「なにかおかしいことに気づく」のは、表現学習を備えたAIのお得意なところ
  - 商品の数を数える、売上をまとめてエクセルを作るなどのルーティーンも人工知能に。
  - 顧客の例外対応や、クリエイティブな分野、あるいはデータ分析・予測に基づく判断は人間の仕事

- 長期（15年以上）
  - 2極化する。ますますAIでできる分野が広がる。
  - 経営や一部の「大域的判断を必要とする仕事」は人間
  - 営業、店員、マッサージ師などの、「対人間」の高級なインタフェースは人間に。

松尾：人口知能の未来-ディープラーニングの先にあるもの（大学講演資料　P.54 ）

未来の社会と産業の構造変化

- 1995年のインターネット
- Googleにあたるものはなにか？
- Amazonにあたるものはなにか？
- Facebookにあたるものはなにか？

- キープレイヤーは？プラットフォーマーはどのように出現する？
- 新たな産業は？産業構造の変化は？
- 競争力はどう変化する？
- 社会はどう変わる？

松尾：人口知能の未来-ディープラーニングの先にあるもの（大学講演資料　P.58 ）

できるかもしれないこと

1. 大規模な一般画像認識(1)
2. コツを学ぶマシン(3):Deep Mindのリアル版
3. 動画から不審者を検出する(2):RNN系か特徴抽出→マルコフモデル
4. 表情、感情を読み取る(2):特徴抽出を教師データの方向に引っ張っていく
5. 試行錯誤して行動を組み立てる(3):Deep Mindの状態遷移獲得→プランニング版
6. 言葉をしゃべると絵を書く(5):SHRDLUのDLによる画像認識版
7. 不正検知システム(2):ハックしようとしても「何かおかしい挙動」を見つける
8. 不快のないように持ち上げる(4):人間からの教師信号の獲得＋試行錯誤
9. 設備保守システム(4):叩いて音の異常を見る
10. 人の事故や危険を察知する(2):表情、感情を教師信号としての学習＋予測
11. 特徴利用の制限スキーム(1-3):プライバシー保護
12. ゲームと実世界を結ぶ。モデル化、シミュレーション、リアル。(4):シミュレーションが特徴抽出で高度化
13. 画面から情報を読み取り入力する(3):情報システムのつなぎこみ
14. 仮説出し(3):実験結果から特徴量抽出→仮説生成
15. 文章を読んで意味を理解する(6)

- 工業用ロボットでラインエの代わりになる
- 運転者のモデルで乗りやすい車をシミュレートする
- パズルゲームを作る

59

松尾:人口知能の未来-ディープラーニングの先にあるもの(大学講演資料 P.59)

変わりゆく社会と産業

- 仮説生成と試行の高速化（デザイン、製薬）
- なんでも市場化する？
- 情報システムが全部つながる？
- 監視・防犯：犯罪は非常に減る？
- 心を持っているように見えるAIが生まれる？禁止される？：恋愛
- 製造者責任と保険
- 物流の完全自動化
- 軍事：権力者を倒す、心を操る
- 知財の新しい形
- 忘れられる権利、いいところだけを見せる権利、悪いことをする権利、大目に見られる（警告を受ける）権利、好きになる権利、…
- 遺伝子の多様性を維持する必要性
- 「人間」の定義：人権、投票権

松尾：人口知能の未来-ディープラーニングの先にあるもの（大学講演資料　P.61 ）

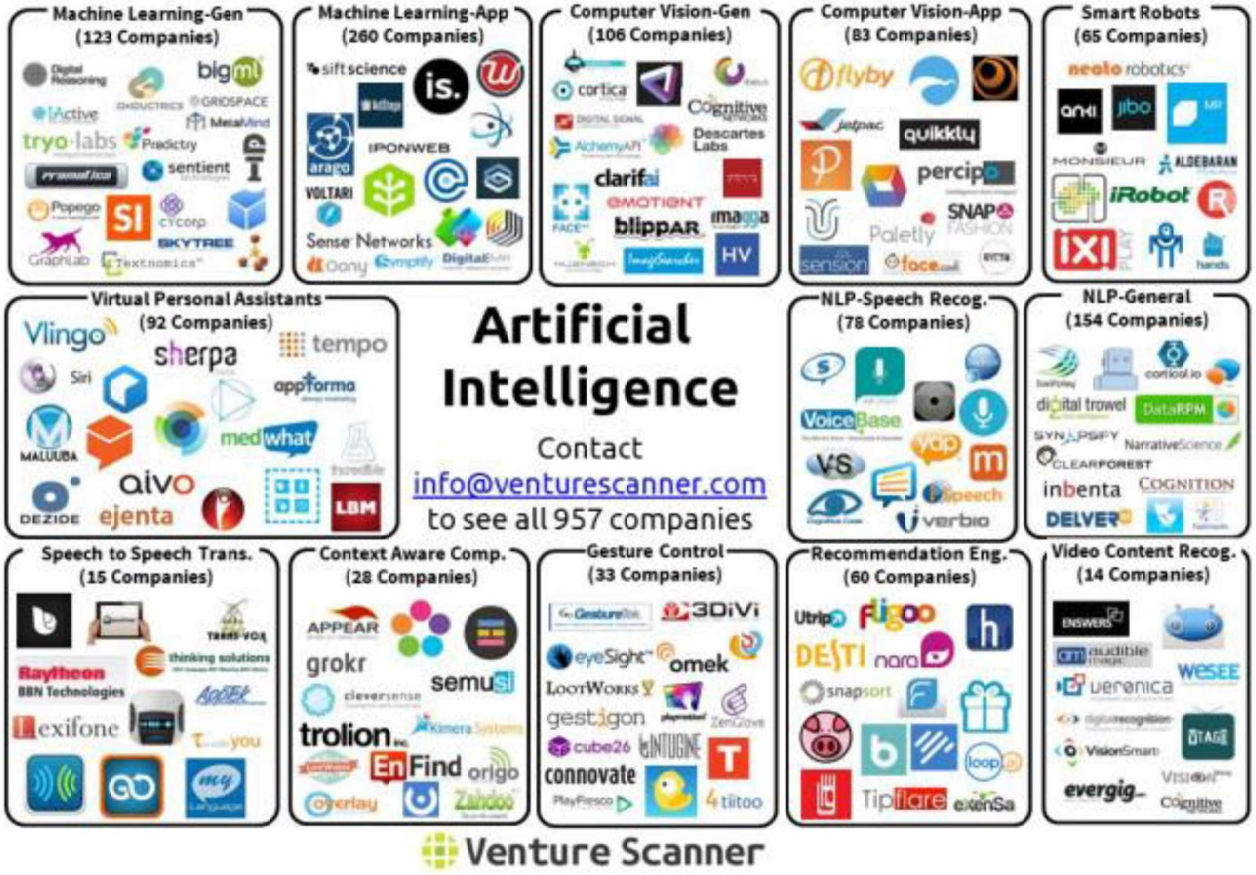
日本の未来

- 少子高齢化する日本のなかで、人工知能を切り札として産業競争力を再び高めたい。
- 日本にもチャンスが
  - 人工知能研究者の人数、人工知能に興味をもつ人数
  - 世代を通じた理解
  - 「賢さ」と「真面目さ」が重要な領域
  - 言語があまり関係ない
- まずはやってみること！
- 人工知能で変化する産業と社会。未来社会を描きたい

松尾：人口知能の未来-ディープラーニングの先にあるもの（大学講演資料　P.63 ）

8. アプリケーション例
ビジネス化例

Venture Scanner ：AI ：957 社(春)⇒1422(10月)



13カテゴリー： 左から右へ

1. 機械学習アルゴリズム/モデル/プラットフォーム
2. 特定分野機械学習アプリ
5. コンピュータビジョン/画像認識（汎用）
6. 同特定分野アプリ
9. スマートロボット

8. パーソナルアシスタンス

4. 自然言語認識
3. 自然言語処理

12. 会話翻訳
11. コンテクスト・アウェア
7. ジェスチャー制御
10. 推薦エンジンと協調フィルタ
13. ビデオ自動内容分析

18 Deep Learning Startups (1) (VentureRader2016/1/9)

- Deepomatic：コンピュータ・ビジョン
  - 製品の属性の自動認識とパターンと色の比較を組合せることにより、e-コマース上で同じ、または同じような製品を探す。
- Clarifai ：静止・動画画像認識
  - Imagenet2013での好成績の後、速度、語彙数、メモリサイズなどを大幅に改善し画像以外の知識抽出にも踏み出している。その肝は新世代の高性能深層学習APIであり、誰もが機械学習を使えるようにする。
- Descartes Labs：画像認識・衛星・農業
  - リモートセンシングに深層学習を適用。最初に穀物生産量を増やすための大量の可視・非可視衛星画像に適用される。
- HyperVerge ：コンピュータビジョンと画像認識エンジン
  - クラウド上の画像とビデオを処理する。独自の特許化された画像技術により、顔認識、顔検知、場面認識、非適切画像、複製画像、写真クラスタ化、写真アルバム化、顔強調などを提供する。
- Tractable：コンピュータビジョン
  - コンピュータビジョンに特化し、ラベル付けされていないデータと教師あり学習の両方をつかう。保険支払、産業検査、リモート監視などに適用。
- Indico：自然言語処理と画像解析
  - 自動認識の目的別モデルを提供し、複数のモデルを組合わせて適用することでデータ解析を助ける。「TEXTタグ」モデルは文節内のトピックを見つけ、「顔画像同定」は同じ顔を見つける。

18 Deep Learning Startups (2) (VentureRader2016/1/9)

- Metamind：自然言語処理と画像解析
  - 自然言語処理、画像認識、知識ベース解析のためにAIプラットフォームを提供。医療画像、食物認識、その他カスタム用の製品を提供中。
- Synapsify：文書解析
  - 文書から意味を読み取り、発見・洞察・推論を助けるアプリを構築。文書解析を機械でおこなうAffectivaのに技術的な知識も経験を必要としない。
- Skymind：文章解析、不正利用検知、スパム検知
  - メディア、画像、音を解析してビジネスに影響するパターンを見つけ出し定量化するエンタプライズソフトウエアを提供。会社は、深層学習スペシャリスト、ロボットと商用に耐える分散深層学習フレーワークの開発者で構成される。
- Atomwise：薬探索
  - 深層学習で新薬を探索する。結合親和力予測と毒性検知の組み合わせで正解最高の結果を得たという。
- Deep Genomics：ゲノム精密医薬
  - ゲノム生物学と精密医薬で世界をリード。自然・治療によるゲノム変化によりDNAの変化がおきるときに細胞内で何が起きるのかを予測する。
- Quantified Skin：健康管理
  - 肥満による慢性病を軽減させることを目的に、人の活動内容を分析、適用、推奨して、人の行動を変化させる。

18 Deep Learning Startups (3) (VentureRader2016/1/9)

- Trustingsocial：クレジット格付け
  - ソーシャル、モバイル、WEBデータにビックデータと深層学習を適用して新規市場向けに顧客のクレジット信用情報提供する。短期収入と長期収入と信用度を学習しFICOスコア（米国でローンの利率の基礎）を補足する。

- Idibon：自然言語処理とソーシャルメディア解析
  - クラウドベースの自然言語サービスを提供。ビジネス上重要な文書の質問に自動で応答する。テキスト抽出、感情分析、テキスト・コンテンツ分類、言語検知/同定などに使える。
- Rapjar：ソーシャルメディア分析
  - ビジネス用ビックデータ分析アプリを提供。ソーシャルメディア、ニュース配信、ブログ、WEBページなどの外部情報と内部情報を結びつけ、ビジネスの重大決定を助ける。
- Trak.io：データ解析
  - 顧客データを追跡するクラウドベースSaaSプラットフォームを提供。使用、支払、チケット、イーメールの履歴を集めて分析し、顧客を分類、動向を予測するモデルを構築できる。

- Affectiva：人の表情の分析
  - 視覚刺激に対する感情反応をとらえて分析する技術。代表製品Affedexは、Webカメラがあれば他には何も必要ない（クラウド上ソフト）。
- Enltitic ：医療診断
  - X線、MRI、CTなどの画像を学習し、診断したり異常をスポットして医者の診断を助ける。

大企業の例

- Apple : Siri
- Google：47種のサービスに適用中（2015/3時点）
  - 検索エンジン
  - Google Speech Recognition：Google Cloud Speech API
  - Google Car
- Baidu 百度：検索エンジン・音声認識・画像認識・コンテンツ連動サービスなど
- Amazon：検索エンジン・推奨エンジンに使用中（未確認）
- Microsoft：Bing検索、音声認識
- Facebook：画像認識他（未確認）
- IBM: WATSON
  - クラウドサービス：Watson Analytics無償版提供中
  - 有償版30日無料トライアル可能。（US版）
- トヨタ
  - TRI（AI研究所）をパロアルトに設立。5年で10億＄。元DARPAギル・プラット氏が代表
- ホンダ
  - 2016/9 赤坂にAI研究所HONDA INOVATION LABO TOKYOを開設 （2016/6/2発表）
  - 本田技研はソフトバンクと協力、ソフトバンクGr CoCoRo SBのAI技術（Pepper向け感情エンジン）を車に応用（2016/7/21）

9. まとめ

まとめ

- AIとDLについての基礎知識を集めた。専門家ではないので間違いが含まれる。
  - かなりの部分が東大松尾先生資料からの写しである。
  - 参考論文は別表を参照されたい。
- DLがなぜうまくいくのかは誰にもわかっていない。やったもの勝ち。
  - やらないでいる手はないのでは？とりあえずフレームワークでとっつくのは難しくない。
- 日進月歩の世界。昨日の常識が翌日には非常識。

2. 「常識」が何度も覆る → 技術が枯れにくい

- 1980年代：LeCunら「画像認識には**CNN**が極めて有効」
  - CNNの要素技術（MaxPoolingなど）を使わず同等の精度を達成する手法が登場
- 2006年：Bengioら「**事前学習**を使えば深いNNを学習できる」
  - NNの最適化技術が発達 → 事前学習なしでも同等の精度を達成できるタスクも増えている
- これまでの情報科学にない速度感での発展
  - 「取材して**2週間**経つと新しい技術が現れる」
  - 論文が公開されると**1, 2ヶ月**でフレームワークに取り込まれる

付録（予備）

Deep Learning：OS

- ほぼすべてのDeep Learning System、大規模なSuper Computerは、LINUXを拡張したものを使っている。
  - TOP500の10位以内は全てLINUX（拡張）。
  - 次ページに、EXAスケールSuper Computer用OSの要件についてまとめた。
- 一般のLINUXとの違い
  - 学習と実行の計算部分はハードウエアに直接アクセスする（らしい）。
  - OSは外部通信のプロトコル対応と全体のスケジューリングを行っている（だけ）。
  - （たぶん）本質的にマルチタスクを想定していない。
- 実行時には（必要なレベルの）短時間での反応のためのOSが必要。
  - 例えば、音も画像も関していているロボットを想定すれば、音声処理と画像処理では別のNNが必要になるはず。実際の処理では、少なくともアクセラレータを複数系統もつかNNを定期的に入れ替える必要がある。（NNのパラメータ自体は数MB程度に収まるケースが多い。）
  - このため、マルチタスク、マルチスレッド、（マルチユーザ）の処理能力が必要なる（と推定）。

ExaスケールSCのOS要件

各ノードがCPU＋GPUs（ヘテロジニアス・アーキテクチャ）で構成されると仮定：

① プログラミングモデルの提供：
- 高水準プログラム言語、フレームワーク、ライブラリ。現在はMPI/OpenMPレベル。
- ソフトウエア生産性・可搬性・移植性の確保：既存アプリケーションの継続性

② 大規模並列性の提供：O（10^9）ノードの管理が可能なスケーラビリティ

③ 耐故障性：ExaではMTBF 40分 ！ ？
- 故障予測（事前対応）、故障判定（OSで正しく動いているか定期的に検証、ECC）、Fault Resillience化（Check Point+リスタートの高速化、冗長実行など）。

④ 低消費電力化：O（10^9）ノードの電力管理とスループットの極大化。

⑤ デバック環境と性能プロファイルの提供。

- さらに：「複雑なメモリ階層構造を扱える構造」「CPUとGPUのリソース管理・スケジューラを一体化」「通信の隠蔽」などへの展開要。
  - Deep Learning：SIMULATIONアプリほどレイテンシ依存にならずスループット重視。
  - マルチスレッド化、マルチユーザ時のメモリ管理（ゼロクリアなど）

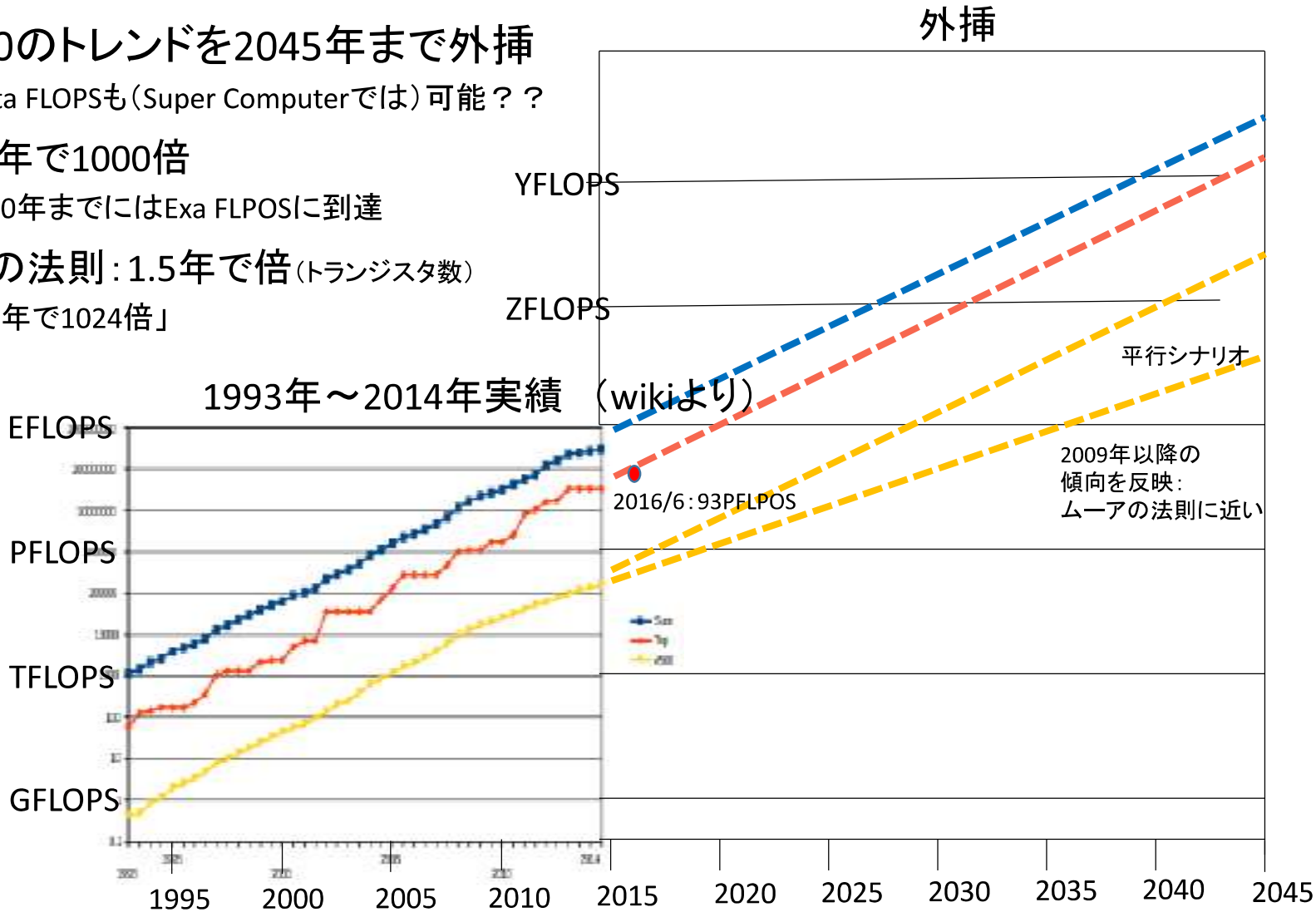
Supercomputer：Top500 2016/6 の10位まで

順位	場所	システム	会社	コア数	Rmax（Tflops/S）	Rpeak（Tflops/s）	電力（kW）
1	中国：NSCC 無錫	Sunway TaihuLight　：Sunway MPP, Sunway 1.45GHz, Sunway	NRCPC	10,649,600	93,014.6	125,435.9	15,371
2	中国：NSCC 広州	Tianhe天河-2（Milkeyway-2）：IVB-FEP Cluster, Xeon E5 2.2GHz, TH Express-2, Xeon Phi	NUDT	3,120,000	33,862.7	54,902.4	17,808
3	米国：オークリッジ国立研究所	Titan：Cray XK7, Opteron 2.2GHz, Cray Gemini, Nvidea K20x	クレイ	560,640	17,590.0	27,112.5	8.209
4	米国：LLNL	Sequoia：BlueGene/Q、PowerPC 1.6GHz, Custom	IBM	1,572,864	17,173.2	20,132.7	7,890
5	日本：理研　AICS	京：SPARC64　2.0GHz, Tofu	富士通	705,024	10,510.0	11,280.4	12,660
6	米国：アルゴンヌ国立研究所	Mira：BlueGene/Q、PowerPC 1.6GHz, Custom	IBM	786,432	8,586.6	10,066.3	3,945
7	米国：SNL	Trinity：Cray XC40　Xeon E5 2.3GHz, Aries	クレイ	301,056	8,100.9	11,078.9	
8	スイス：CSCS	PizDaint：Cray XC30　Xeon E5 2.6GHz, Aries, Nvidea K20x	クレイ	115,984	6,271.0	7,788.9	2,325
9	ドイツ：HLRS Stuttgart	Hazel Hen：Cray XC40　Xeon E5 2.5GHz, Aries	クレイ	185,088	5,640.2	7,403.5	
10	サウジアラビア：アブダラ国王大学	Hazel Hen：Cray XC40　Xeon E5 2.3GHz, Aries,	クレイ	196,608	5,537.0	7,235.2	2,834

100PetaFlopsが目の前に。1位は電力効率を従来の3倍に。

TOP500　トレンド

- TOP500のトレンドを2045年まで外挿
  - Yotta FLOPSも（Super Computerでは）可能？？
- 12～13年で1000倍
  - 2020年までにはExa FLPOSに到達
- ムーアの法則：1.5年で倍（トランジスタ数）
  - 「15年で1024倍」



PEZY連合の戦略（1）

- ZettaScaler-2.0で100Petaを実現（液浸槽70個）。2017年。
- ZettaScaler-3.0で1Exaを実現。2019年末。25MW目標。
  - 2016/6 Sunway TaihuLight：93PFOLPS
  - 各国の目標：2018 US Summit：200P/2020 日本ポスト京：1Exa下回る？ /2020 中国天河3号：1Exa
- PEZY-SCxはソフト開発容易性を捨てている。
  - 齊藤氏はAI/DLには不向きと言い切っている。DL FWも全く動かない。
  - スクラッチパッドメモリを採用（TaifuLightと同じ）。キャッシュ間コヒーレンシも取らない（GPU内部も取っていない）。分岐予測もしない。
  - Shobuで利用者公開している：電通大 山崎氏はOpenCLでGPUからShobuに小脳シミュレーションを移植は可能だがTUNINGが大変といっている。10億ニューロン実装済み。
  - （齊藤氏談）日本では人海戦術に頼りずらいので人工知能を活用して開発できるようにする。

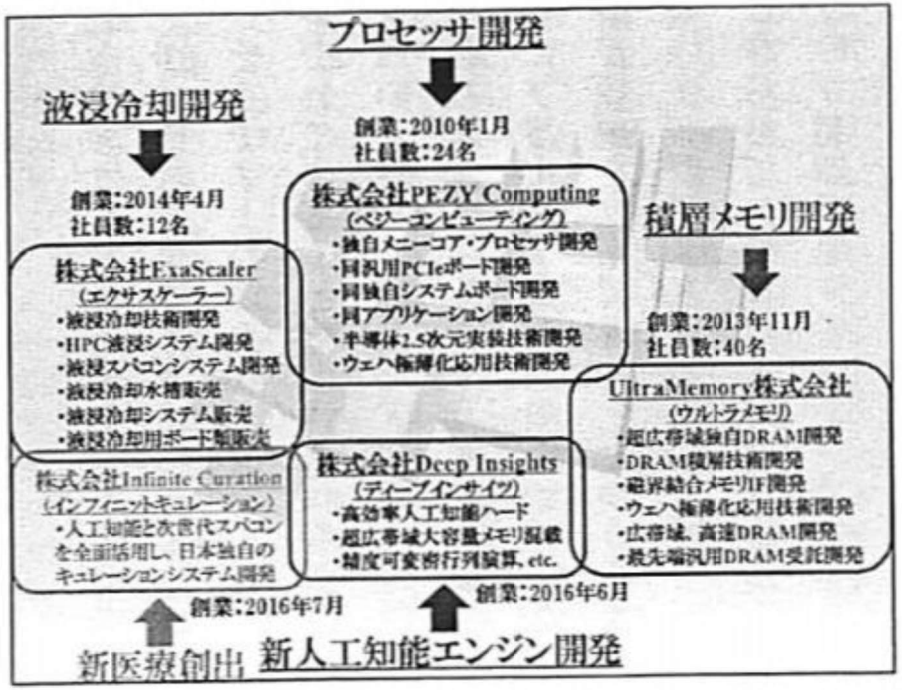
表1　2019年にも1ExaFLOPSに
PEZYグループのスパコン開発ロードマップ。2018年には深層学習用プロセッサーの投入も予定している。

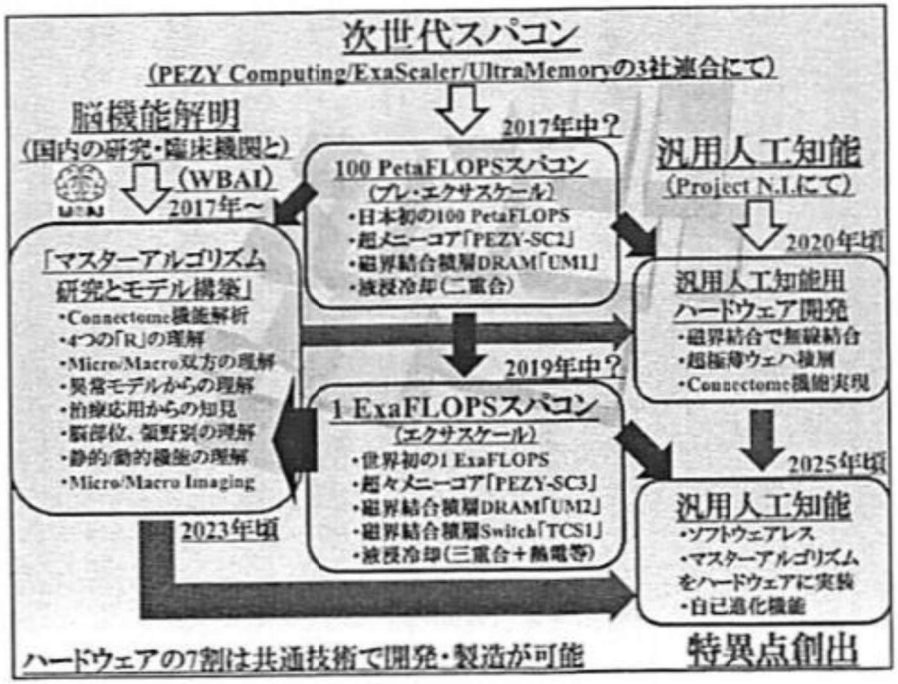
時期	2016年6月	2017年6月		2019年11月
システム名称	ZettaScaler-1.6	ZettaScaler-2.0		ZettaScaler-3.0
液浸槽数	5	7	70	125
LINPACK性能(FLOPS)	1Peta	10Peta	100Peta	1Exa
消費電力当たりの性能(GFLOPS/W)	6.67	15～20		40
プロセッサー	PEZY-SCnp*1	PEZY-SC2		PEZY-SC3
MIPS64コア数	なし	12		未定
PE(Processor Element)数	1024	4096		8192
ピーク演算性能(倍精度、FLOPS)	1.5T	8.2T		20T
動作周波数(Hz)	733M	1G		1.25G
メモリー容量(Gバイト)	32(DDR4)	16(磁界結合方式)+256		未定
メモリー帯域(バイト/秒)	最大150G	最大4.1T		最大10T～20T

*1 「Xeon E5-2618L v3」を併用

PEZY連合の戦略（2）

- Infinite Curation、Deep Insitesは設立時メッセージ後ほぼ情報なし。
- Ultra Memory UM1開発状況？？
  - 2016/末 ES（2015/8時点、出資時点もこの認識）。試作中のはず。
- PEZY-SC2開発状況？？
  - TSMC 16nm FinFET　2017/2 ES（2016/4時点）（2015/8時点では2016/末ESだった）





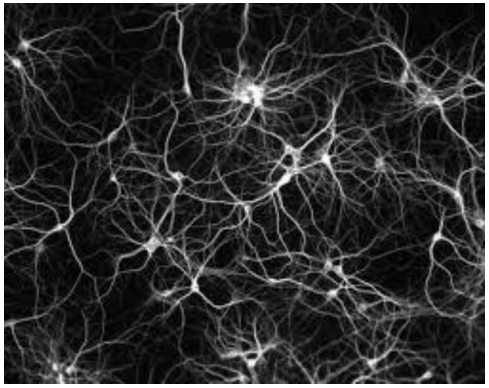
The Master Algorithm

- 人工知能がシンギュラリティにたどり着くためにはハードウエア「AIエンジン・AIチップ」とソフトウエアの大幅な進化が必要。
  - Master Algorithmが解明され実装される必要がある。
    - 大脳新皮質で働いているアルゴリズム。
    - HTM（Hierarchical Temporal Memory）とPredictive Codingが有力視されている（らしい）。
  - 特定の目的別のモデルではなく人の管理が入らないで機械学習ができる「マスターアルゴリズム」（Pedro Domingo：2015　The Master Algorithm）が必要とされる。
- HTMは、機械学習の一つのオンライン・モデル
  - 実装はC++/PYTHON BASEのフリーライセンスでNuPICにより研究用にリリースされている。
  - SVMやConvolutional Networkと違い、大脳皮質構造を参考にした階層的構造を持ち入力された情報の流れと時間を統合して記憶することができるとしている。
  - 小職は以前から「AIチップが脳構造をモデル化するものであるなら、深い（古い）記憶と中間のすぐに思い出せる記憶と現在の（浅い）情報を最低3段に管理すべきではないか」と考えている。これに共通する点がある？

Connectome（調査中）

- Connectome
  - 生物の神経回路の地図全体を示す言葉。Genomeの次の解明ターゲット。
- Micro/Macro
  - Micro Connectome：ニューロン間シナプス結合、領野内の接続（蜜結合）
  - Macro Connectome：領域間の接続 （疎結合）
- 四つのR
  - 不明
- 異常モデル
  - 不明

Macro Connectome



Micro Connectome