

人工知能（AI）は 医療に何をもたらすか

東京医科歯科大学 名誉教授
東北大学 東北メディカル・メガバンク機構
機構長特別補佐

田中 博

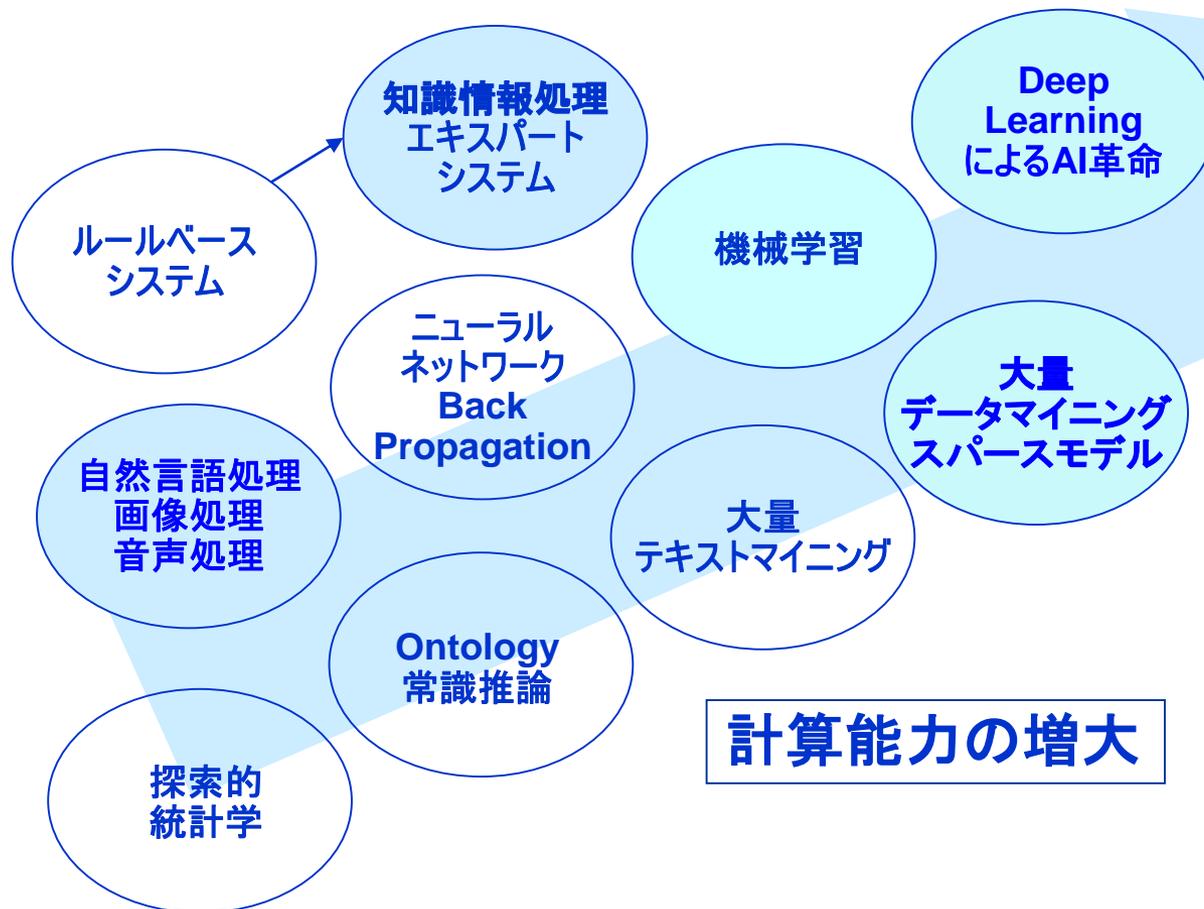


人工知能への期待

人工知能 (AI) の分野

データの増大

ビッグデータ
人工知能による
知的処理



人工知能の最近の話題

- 「**アルファ碁**」 (Google DeepMindによるコンピュータ囲碁プログラム) が2016年3月に数多くの世界戦優勝経験のあるプロ棋士李世石 (Lee Sedol : 九段) に挑戦し、**4勝1敗と勝ち越した**
 - チェス : IBM 「Deep Blue」 が1997年に当時の世界 champion, カスパロフ氏 (ロシア) に勝利
 - 将棋 : ボンクラーズ, 2012年米長永世棋聖に勝利
- 人工知能が1000万枚の画像を与えて「**猫**」を認識するニューロンをできたと2012年に発表



アルファ碁

- 「アルファ碁」にはニューラルネットワーク (Deep Learning) が使われた。評価経験則が人間によってコードされていない
- 最初、棋譜に記録された熟練した棋士の手と合致する手をさすように訓練され、次に、ある程度の能力に達すると、強化学習を用いて自分自身と多数の対戦 (3000万回) を行うことで上達した。
- 2017年初頭、その改良版が日中韓のトップ棋士を相手に60戦無敗というな驚異的戦績を挙げた。

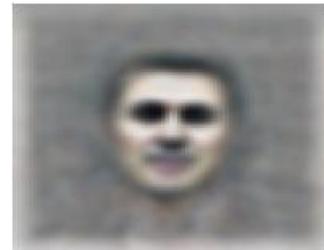


Googleの猫

- YouTubeにアップロードされている動画から、ランダムに取り出した200x200ピクセルサイズの画像を1000万枚用意し、これを用いてDeep Learningを行った（3%前後の画像に人間の顔が含まれていた。猫が含まれる画像もたくさんあった）
- ニューラルネットワーク（Deep Learning）の最初の層の入力は各画素（ $200 \times 200 = 40,000$ ）のRGBの値で、9つの階層を構築した。
- その結果、人間の顔、猫の顔、人間の体の写真に反応するニューロンができた
- ニューロンが最も強く反応する画像を作り出した



猫のニューロンが最もよく反応する画像



人の顔のニューロンが最もよく反応する画像

人工知能とは

定義：人工的にコンピュータ上など知的な振舞いを行うシステムを実現すること、あるいはそのための技術

2つのアプローチ

記号的AI：人間の知識を明示的に表現し、それを用いた推論などの知的課題を解決する。「人間を真似る」
「古いAI」。知識準拠型（エキスパート）システムなど

非記号的AI：人間の用いる思考・知識内容とは関係なく最良の計算方式により知的課題を解決する。「人間を超える？」
「計算AI」ニューロネットワークなど

人工知能研究の推移

	人工知能の置かれた状況	主な技術等	人工知能に関する出来事
1950年代			チューリングテストの提唱 (1950年)
1960年代	<p>第一次人工知能ブーム (探索と推論)</p>	<ul style="list-style-type: none"> • 探索、推論 • 自然言語処理 • ニューラルネットワーク • 遺伝的アルゴリズム 	ダートマス会議にて「人工知能」という言葉が登場 (1956年) ニューラルネットワークのパーセプトロン開発 (1958年) 人工対話システムELIZA開発 (1964年)
1970年代	<p>冬の時代</p>	<ul style="list-style-type: none"> • エキスパートシステム 	初のエキスパートシステムMYCIN開発 (1972年) MYCINの知識表現と推論を一般化したEMYCIN開発 (1979年)
1980年代	<p>第二次人工知能ブーム (知識表現)</p>	<ul style="list-style-type: none"> • 知識ベース • 音声認識 	第五世代コンピュータプロジェクト (1982~92年) 知識記述のサイクプロジェクト開始 (1984年) 誤差逆伝播法の発表 (1986年)
1990年代	<p>冬の時代</p>	<ul style="list-style-type: none"> • データマイニング • オントロジー 	
2000年代	<p>第三次人工知能ブーム (機械学習)</p>	<ul style="list-style-type: none"> • 統計的自然言語処理 • ディープラーニング 	ディープラーニングの提唱 (2006年)
2010年代			ディープラーニング技術を画像認識コンテストに適用 (2012年)

(出典) 総務省「ICTの進化が雇用と働き方に及ぼす影響に関する調査研究」(平成28年)

医療分野の人工知能の歴史

記号（シンボル）的知識処理

ニューロネットワーク処理

1970

問題解決の一般探索手法 **GPS**
解決木の高速探索（ゲーム）

ニューロネットワーク
3層の学習機械 **Perceptron**
入力層、隠れ層、出力層

1980

推論システム（if-thenルールシステム）
知識の表現と利用（専門家システム）
医療診断システム（Mycin, Internist-I）
大ブーム 医療から産業応用の期待波及

多層型ニューロネット
後方伝播 **Back Propagation**
結合係数修正アルゴリズム

1990

期待消滅！

知識発見 機械学習
Machine Learning, KDD
診断知識のDBからの学習

しばらく停滞！

2000

知識準拠診療支援（DSS）
医療ターミノロジー
医療オントロジー

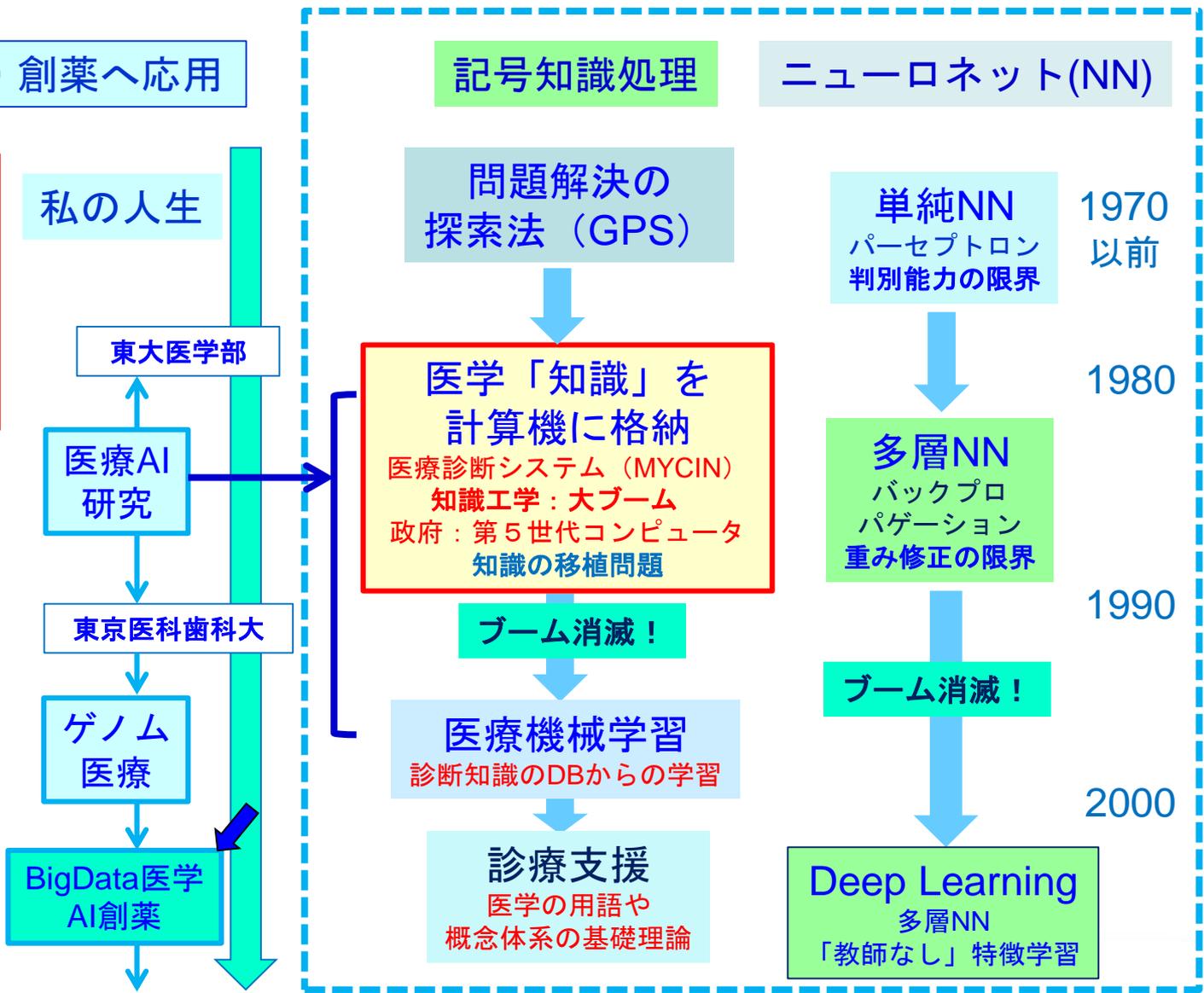
ニューロネットワーク型
多層型ニューロネット
深層学習 Deep Learning
結合係数修正アルゴリズム
画像処理から創薬まで

自己紹介と医療人工知能の歴史

人工知能(AI)を医療・創薬へ応用

田中 博
 東京医科歯科大学
 生命医療情報学
 東北大学
 東北メディカル・
 メガバンク機構

1980から1995
 第1期の
 AIブームの時
 医療AI研究に従事



私の人生

東大医学部

医療AI研究

東京医科歯科大

ゲノム医療

BigData医学 AI創薬

記号知識処理

問題解決の探索法 (GPS)

医学「知識」を計算機に格納
 医療診断システム (MYCIN)
 知識工学: 大ブーム
 政府: 第5世代コンピュータ
 知識の移植問題

ブーム消滅!

医療機械学習
 診断知識のDBからの学習

診療支援
 医学の用語や
 概念体系の基礎理論

ニューロネット(NN)

単純NN
 パーセプトロン
 判別能力の限界
 1970 以前

多層NN
 バックプロパゲーション
 重み修正の限界
 1980

ブーム消滅!

Deep Learning
 多層NN
 「教師なし」特徴学習
 2000

知識ベース型医療人工知能

知識ベース型医療AIシステムの興隆

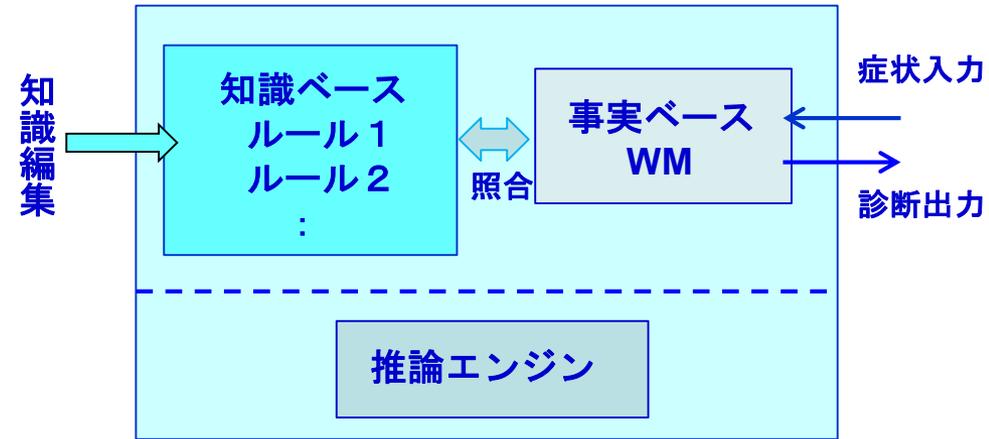
知識型人工知能の流れ（1970年代後半から）

- **知識原理**：問題解決能力を高めるのは、すべての問題に共通するような一般的な推論戦略ではなくて、対象領域に固有な「知識」である。
- **エキスパート（知識ベース）システム**
 - 専門家の問題解決と知識、スタンフォード大学計算機科学科
 - 医師の知識や診断活動に注目
- **感染症の診断システム「MYCIN」**
 - 最初の医療エキスパートシステム。
 - 医療だけでなく産業界に知識情報処理として多大な影響を与えた
 - 細菌感染症の診断と抗生物質による治療支援する
 - 医学知識 <IF（前提）THEN（結論）>ルール形式
 - <IF（症状）THEN（病名）>など
- **同時代の医療エキスパートシステム**
 - 腎疾患診断システムPIP（MIT）
 - 緑内障診断システムCASNET（ラトガース大）
 - 内科疾患全体Internist-I（ピッツバーグ大）など

知識ベース型医療AIシステム

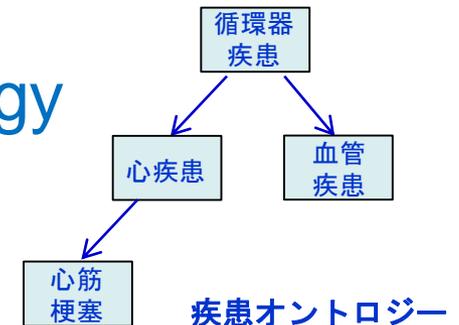
知識ベースシステムの構造と機能

- **知識ベース**
 - 推論ルールの集合
 - IF-THENルールなど
 - MYCINで400程度
- **推論エンジン**
 - ルールを連鎖させて診断を行う
 - 前向き推論 $\langle A \Rightarrow B \rangle, \langle B \Rightarrow C \rangle$ なら $\langle A \Rightarrow C \rangle$
 - 後向き推論 $\langle C \leftarrow B \rangle, \langle B \leftarrow A \rangle$ ならAを検証
- **事実ベース**
 - 作業メモリ
 - 事実を書込みルールを適用して推論結果に書き変える



知識ベース型医療AIのその後の展開

- 知識ベース作成の困難性より沈滞
 - 知識獲得問題 (knowledge acquisition)
 - 医学診断の**全面的AI化は不可能**
- 診療支援システムへ
 - 診療コンテキストへの知識AIの部分的適応
 - 計算機が得意な知識の広汎性など支援
 - 電子カルテと連動してPOCで警告,代替案
- 医学概念の電子化の基礎学へ
 - 概念間の体系的関係を表すontology
 - 医学概念の相互運用性のためterminology

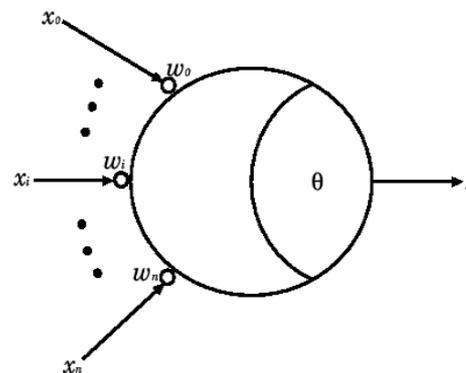
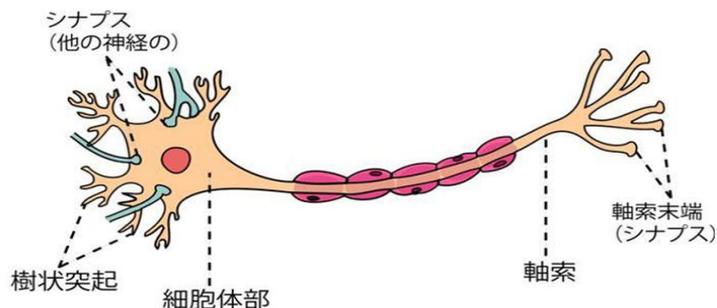


ニューロコンピュータと Deep Learning 人工知能の 革命性

ニューロ・コンピュータの歴史

- マッカロー・ピッツ (MCP) モデル
- 神経細胞 (ニューロン) の機能
- 信号処理素子モデルとして提案
- 1943年の神経生理学者マッカローと数学者ピッツにより提案される

入力信号 0,1



AND回路 $\theta = 1.5$

OR回路 $\theta = 0.5$

NOT回路

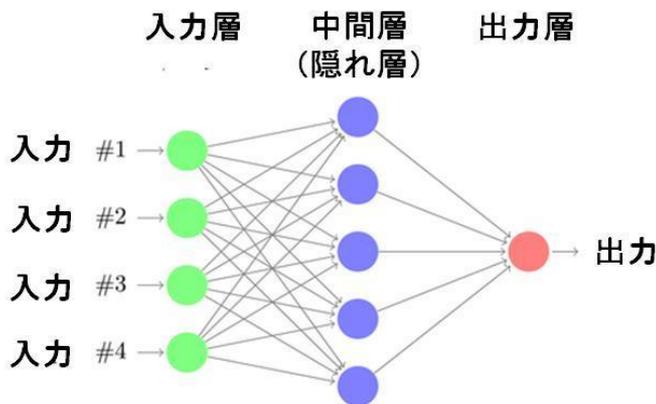
$w = -1.0$

$\theta = -0.5$

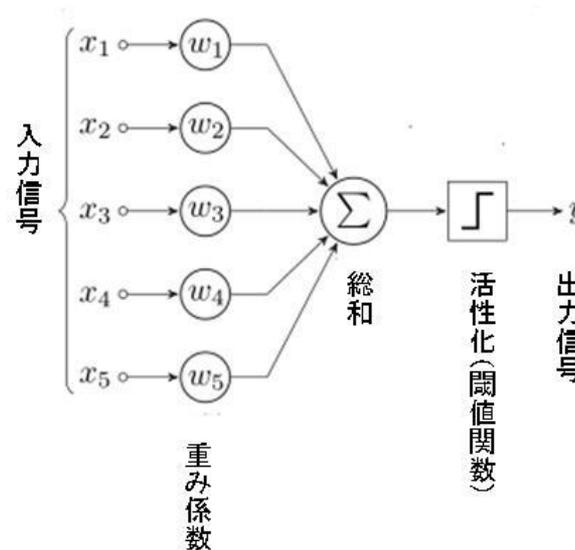
マッカロー・ピッツのニューロン・モデル

パーセプトロン・モデル

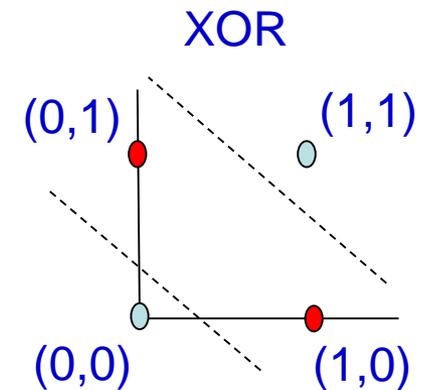
- パーセプトロンモデル
 - ローゼンブラットが1957年に提案 (58年論文)
 - MCPの神経モデルを用いて脳の神経網を表すニューラルネットワークモデルを構築
 - 学習法：出力誤差に合わせて w_i を大きさに比例して更新
- MITのMinskyによって線形分離問題しか解けないことを明示される⇒熱意の消失



(a) 3層表現のパーセプトロン



(b) 単純パーセプトロン

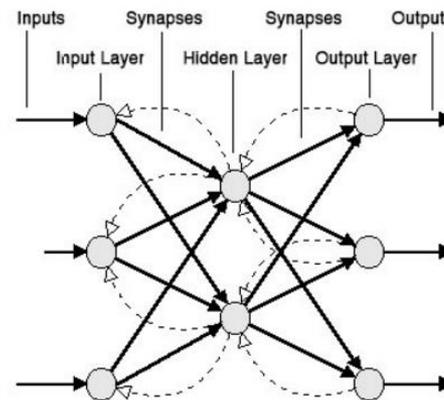
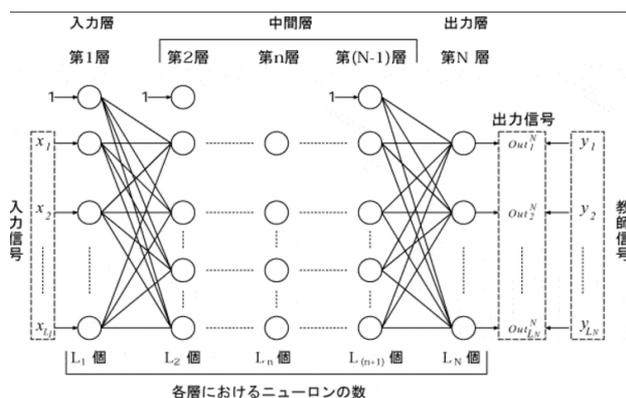


(c) 線形分離できない



多層パーセプトロンと逆伝播法

- 多層にすれば線形判別だけでなく非線形判別が可能になり、ニューロネットワークの識別能力も増大
- 多層にしたとき、加重ネットワーク結合の重みの更新法を1986年にルーマルハートが提案
- 出力層から誤差伝播—逆伝播法（backpropagation法）
- 多層が深くなると入力層まで誤差修正が届かない
 - ブーム鎮静化

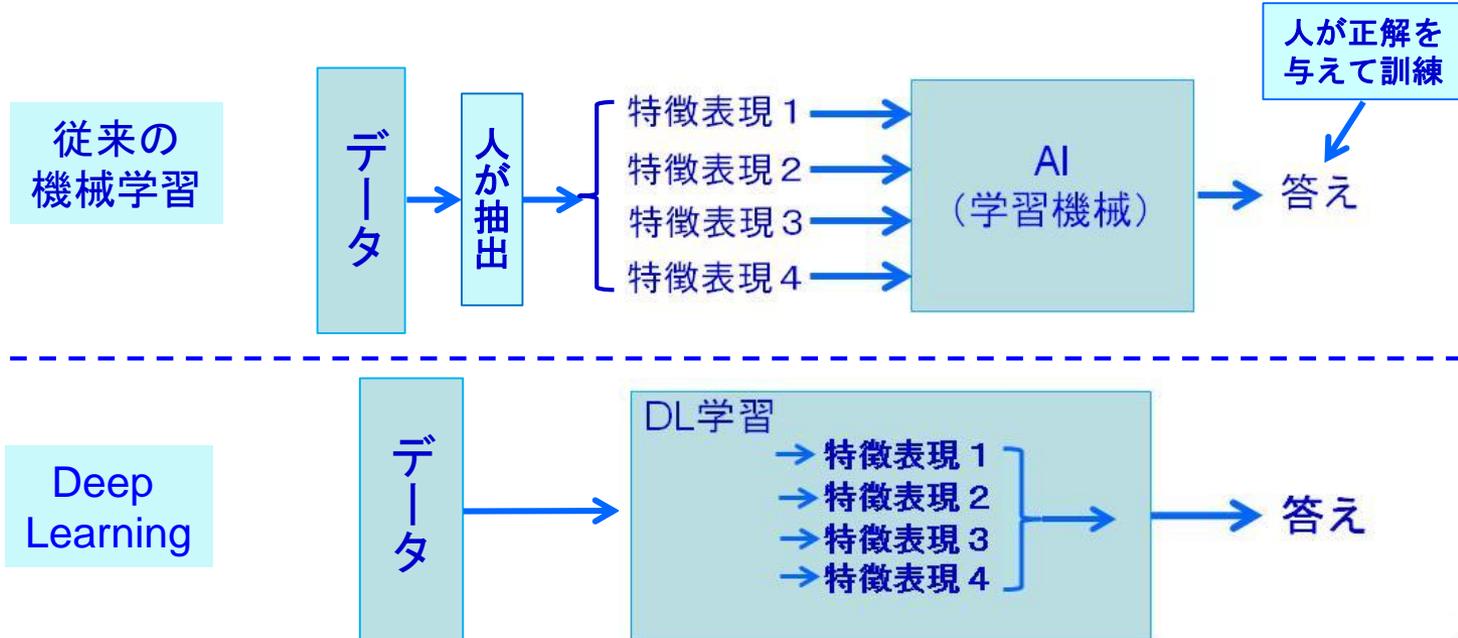


Back Propagation (1986 Rumelhart)
望ましい出力との誤差を教師信号として与える事により、逆方向に結合係数を変化させ、最終的に正しい出力が得られるようにする。結合係数を変える事を学習と呼ぶ。この学習方法には、最急降下法（勾配法）が使われる。出力層へ寄与の高いノードの重みの変更。

多層にわたる逆伝搬で修正感度減衰

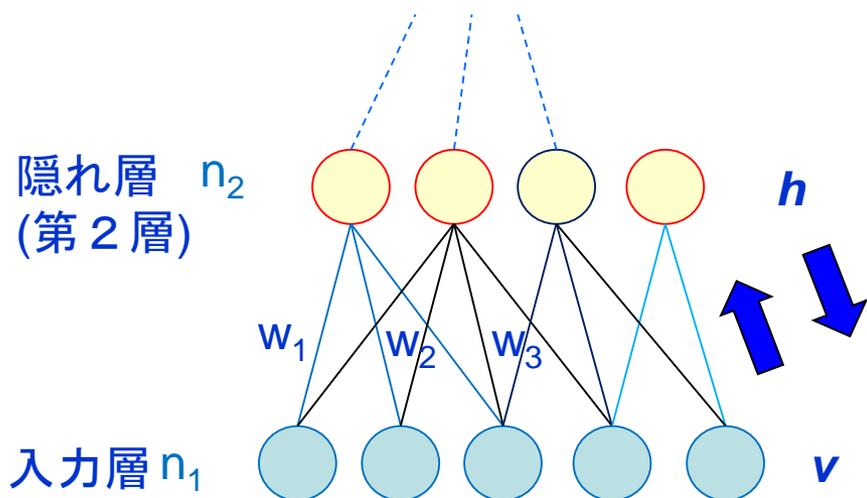
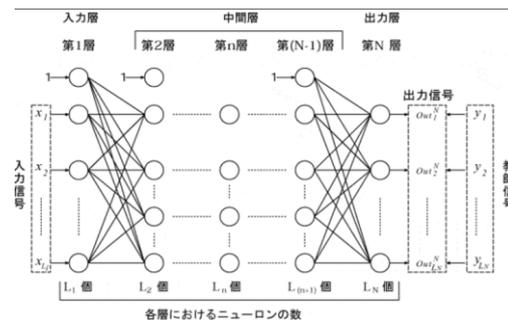
Deep Learning による 人工知能革命

- 機械学習のこれまでの限界
 - 「教師あり学習」
 - 分類対象の特徴と正解を与え学習機械（AI）を構築
- Deep Learningの革命性
 - 「教師なし学習」
 - 対象の特徴表現や対象の高次特徴量を自ら学ぶ

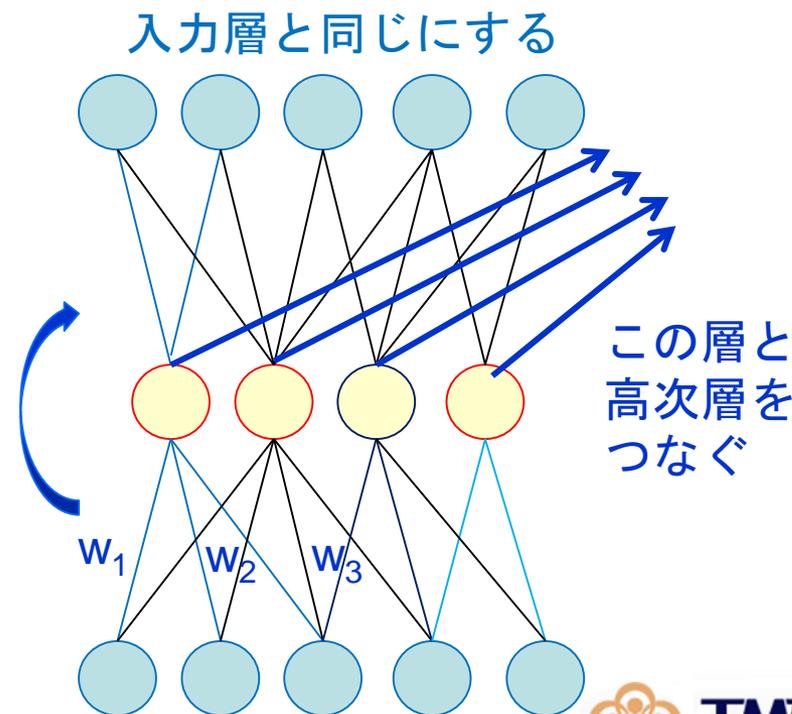


DLの革命点 Autoencode 1

- 対象に固有な**内在的特徴**を学ぶ自己符号化の原理
- 格段ごとに入力を、少ない中間層を介して復元できるか、最適な重みを見付ける
- 次元を圧縮されても可及的に復元する
 - できるだけ**復元に効果的な特徴量**を探索する
 - 内在的な特徴量**を見出す

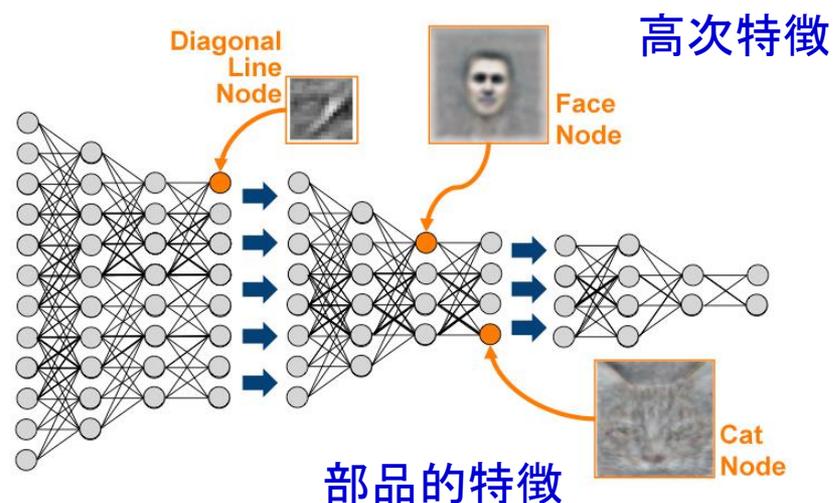
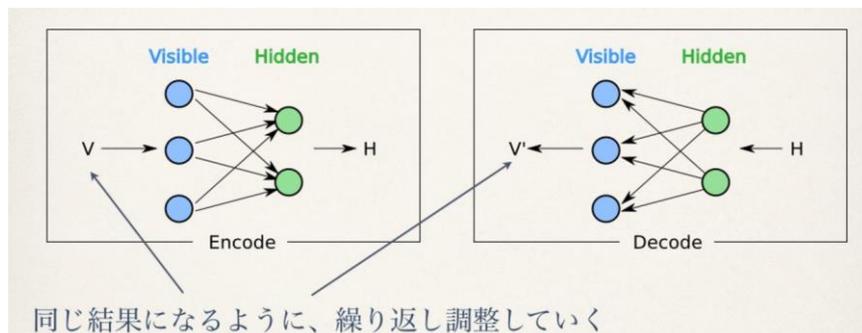


$$n_1 > n_2$$



DLの革命点 Autoencoder 2

- 各層ごとに自己符号化を行うので**何層でも組める**
 - 各層間で「自己符号化」の積上げ (autoencoder stack)
- 第一層で学習した特徴量を使って次の階層を作るので**高次の特徴量**が作られる
- 特徴的表現と概念を結びつけるため「**教師あり学習**」が最後に必要。
- 自動特徴抽出によってこれまでの学習手法の限界を克服した
 - 内在的な特徴量による構造的な理解
- 人間の「思考の枠組み」を超えた正解の低次
 - 「**アルファGo**」が定石にない手で碁の名人に勝つ



医療ビッグデータ時代における 医療人工知能の役割

Deep Learningなどの
非記号的AI・機械学習への
注目

医療ビッグデータ時代の到来

- ゲノムなどの網羅的分子情報やmHealth分子画像など, 診療情報の急激な増大
- 医学知識, 論文情報の急速な増大



従来の医学知識をモデル化する
human-mimicなアプローチよりも
ビッグデータから新しい知識を抽出する
データ駆動型AIへの期待

AI医療のパラダイム転換

医療ビッグデータ時代の到来

- (1) 次世代シーケンサなどによる「ゲノム/オミックス医療」による網羅的分子情報蓄積
- (2) モバイルヘルス(mHealth) によるWearable センサ情報の継続的蓄積 (unobstructed monitoring)
- (3) Biobankによるゲノム・コホート情報

大量データの急激な
コストレス化かつ高精度化



ゲノム : 13年→1日(1/5000) 3500億→10万円(1/350万)

個別化医療・予測医療
健康・医療の適確性の飛躍的な増大



医療の「ビッグデータ革命」

～何が新しいのか～

1) 臨床診療情報

- 従来型の医療情報
 - 臨床検査、医用画像、処方、レセプトなど

2) 社会医学情報

- 従来型の社会医学情報
 - 疫学情報・集団単位での疾患罹患情報

3) 新しい種類の医療ビッグデータ

- 網羅的分子情報・個別化医療
 - ゲノム・オミックス医療
 - システム分子医学・Precision Medicine
- 生涯型モバイル健康管理 (mHealth)
 - ウェアラブル・生体センシング

旧来のタイプの
医療データの
大容量化

新しいタイプの
医療ビッグデータ

医療の「ビッグデータ革命」

～ゲノム・オミックスデータの基軸的な特徴～

＜目的もデータ特性も従来型と違う＞

従来の医療情報の「ビッグデータ」

Big “Small Data” ($n \gg p$)

医療情報・疫学調査では属性数：10項目程度

— 目的：Population MedicineのBig Data

⇒個別を集めて「集合的法則」を見る

網羅的分子情報などのビッグデータ

Small “Big Data” ($p \gg n$)

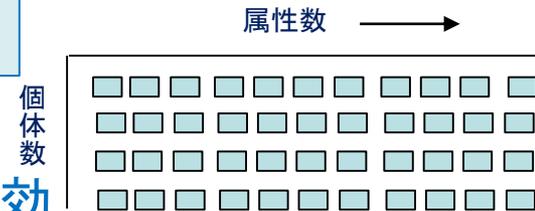
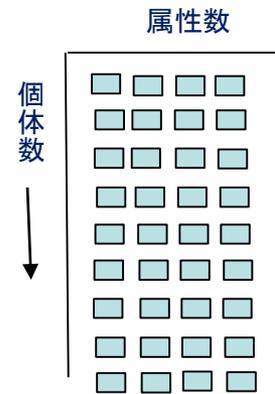
1個体に関するデータ属性種類数が膨大

属性に比べて個体数 少数:従来の統計学が無効

「新NP問題」：多変量解析:GWASで単変量解析の羅列

— 目的：例えば医療の場合Personalized Medicine

⇒大量データを集めて「個別化パターン」の多様性を抽出



新しいデータ科学の必要性

医療の「ビッグデータ」革命は どんな既存のパラダイムに挑戦しているか

- Population medicineのパラダイム転換
 - <One size fits for all>のPopulation医療はもはや成り立たない
 - 個別化医療 “Personalized (Precision) medicine”
 - 個別化医療を実現するために<個別化・層別化パターン>を網羅的に調べる：どこまでの粒度で個別化・層別化すればよいか
- Clinical research（臨床研究）のパラダイム転換
 - 臨床研究を科学にする従来の範型RCTは、個別化概念に破綻した
 - <statistical evidence based>呪縛からの解放
 - 「標本」統計・「推測」統計学に限定されない臨床研究
 - Real World Data: ビッグデータ知識生成（BD2K）
- 創薬の戦略パラダイムの転換
 - ビッグデータ創薬の可能性
 - 網羅的分子データからの計算機創薬・システム創薬
 - Transdisease Omics, Drug networkのDual Network Topologyによる創薬

次世代シーケンサのインパクト

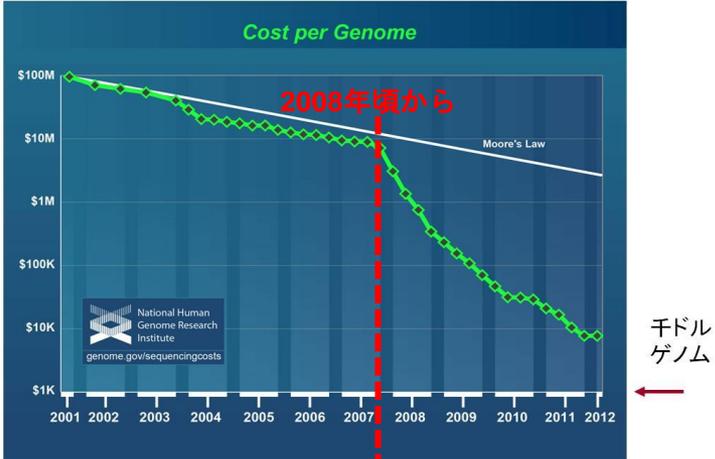
次世代シーケンサを始めとするhigh-throughput分子情報収集の急激な発展

急速な高速化と廉価化 ヒトゲノム解読計画13年,3500億円⇒1日,10万円

2005~ NGS 454 (LS,Roche)
2007/8~454, Solexa (Illumina),
SOLiD (LT,TF)
シーケンス革命



	HiSeq2500	Ion Proton
本体価格	約1億円	約3500万円
モード / チップ	ハイアウトプット ラピッドラン	Ion Proton I
解析時間	11日	27時間
リード長 (bp)	2 x 100	2 x 150
データ産出量 (Gb)	約600	約120
試薬コスト (ヒト1人全ゲノム)	数十万円	不可 エクソームのみ



HiSeq X システム 10台構成 (経費1/5)

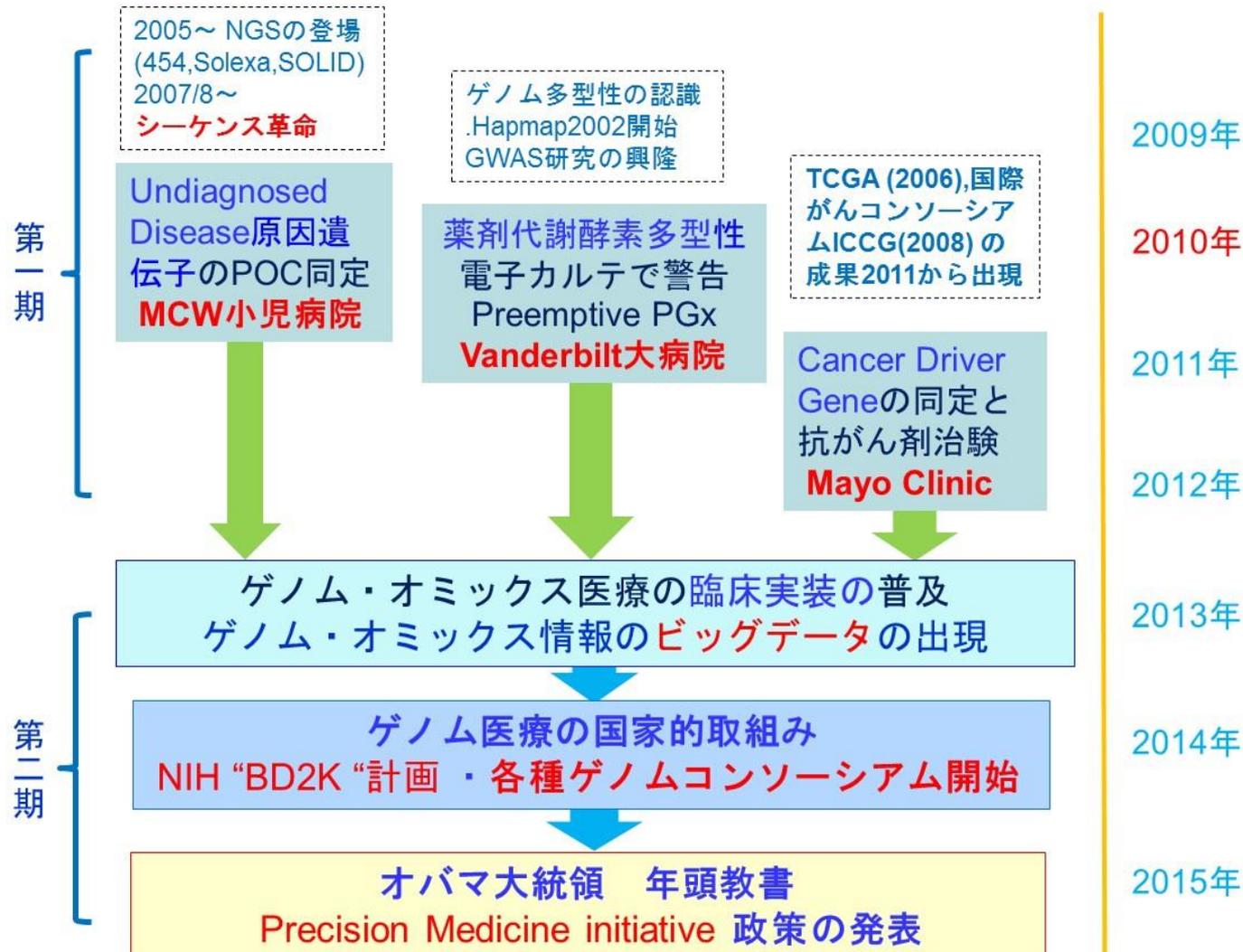


DNA Sequencing Cost: the National Human Genome Research Institute

シーケンス革命 2007/8

ゲノム(配列決定)機器の進歩は、計算機のムーアの法則を越えている!

ゲノム・オミックス医療の流れ



医療ビッグデータ時代とAIの必要性

ゲノム医療の実践

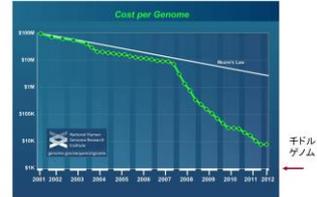
第1段階 ゲノム医療の発展

次世代シーケンシングの臨床普及 (2010~)

全ゲノム (X30 : 100Gb)・エキソーム解析 (X100 : 6Gb)

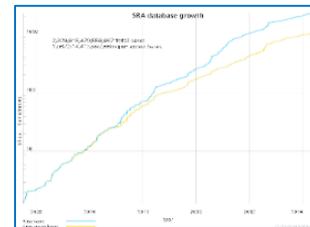
米国では数十の著名病院で実施

ゲノム・オミックス情報の蓄積



DNA Sequencing Cost: the National Human Genome Research Institute

2000兆塩基 (2 Pb)
が登録(NCBI:SRA)



医療ビッグデータ

第2段階 医療ビッグデータ時代

医療情報との統合

電子カルテからの
臨床フェノタイプ

医療ビッグデータ

学習アルゴリズム

ゲノム医療知識

人工知能AI



MayoClinicでは
10万人患者WGS

臨床表現型 eMERGEプロジェクト

electronic Medical Record + Genome (NIH grand)

電子カルテからphenotyping

- **phase I (2007-2011) 臨床表現型情報のタイピング**
 - 電子カルテを通して臨床phenotypingするときの形式
 - EMR : 臨床phenotypingとbiorepositoryに基づくGWASが可能か (EMR-based GWAS)。ELSI側面も検討
 - eMERGE-I: Mayo Clinic, Vanderbilt大学, Northwestern大学など 5 施設

- **phase II (2011-2015) 臨床実装**
 - 電子カルテと遺伝情報の統合
 - 電子カルテへのゲノム情報の統合
 - PGxの臨床応用に関する試行プロジェクト
 - 結果回付 Return of Result (RoR)
 - 4施設がeMERGE-IIより加わる
 - いくつかの小児病院とMount Sinai/Gesinger

- **phase III : 2015より始まる**

- **CSER consortiumと連携**

- “Clinical Sequencing Exploratory Research” コンソーシアム
NHGRIにより予算化



国家戦略としての「医療ビッグデータ」

NIH「ビッグデータから知識へ」計画

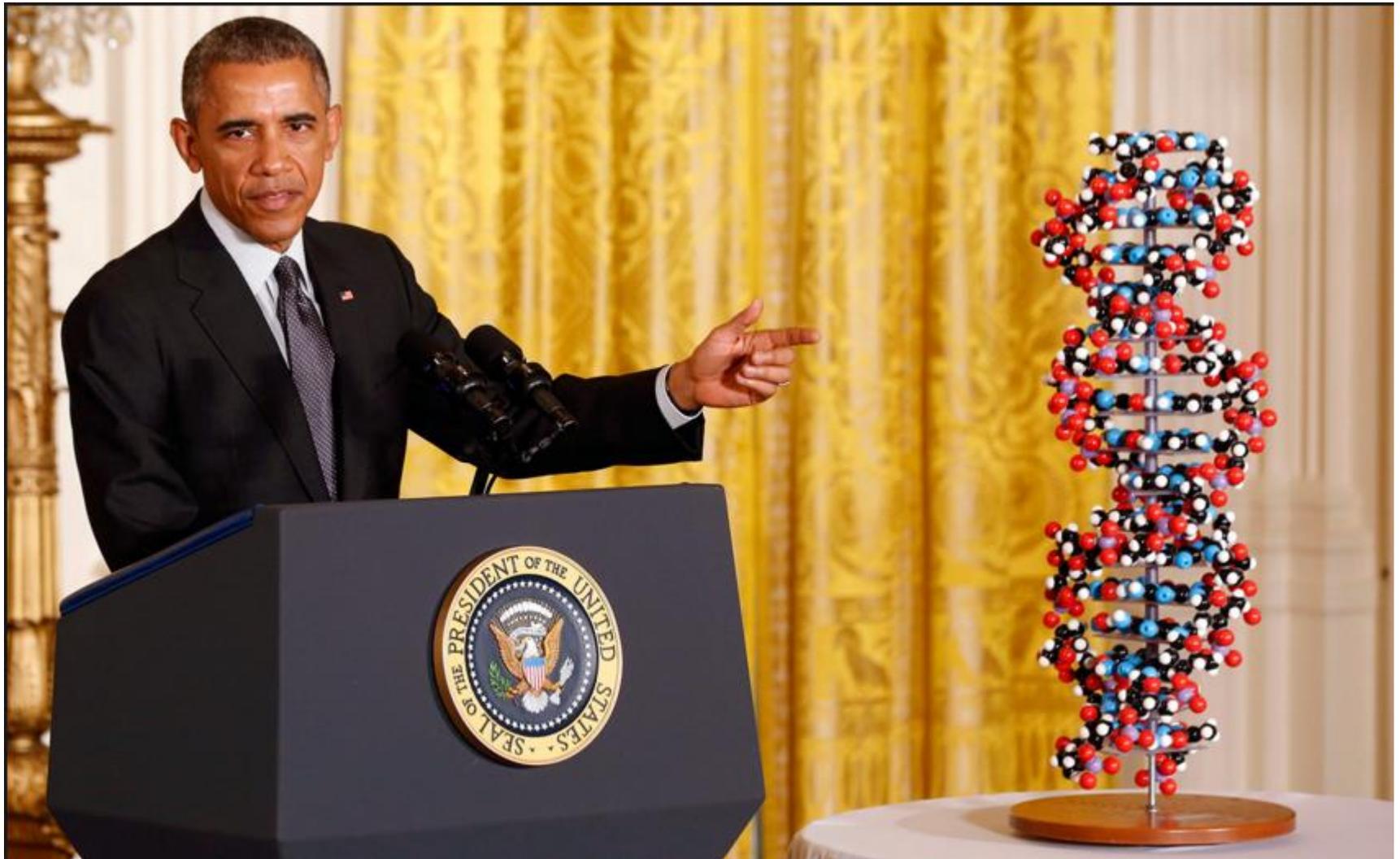
“Big Data to Knowledge” (BD2K) initiative

- **BD2K: “Big Data to Knowledge” Initiative 開始**
 - 次世代シーケンサによるゲノム・オミックス医療の普及
 - 臨床シーケンス情報蓄積の大量化蓄積に対応して政策立案
 - 研究費の配分**2013年**に提案。計画実施は2014年から
 - **データ科学のための副長官**（Associate Director of Data Sciences）を医療情報の世界から任命 **Bourne, PhD.**
- **Francis Collins長官談「NIH全規模での優先計画」**
 - 生命医療研究に喫緊の重要性を持つ、指数的に増大する生命医療データを活用する。
 - 「ビッグデータの時代は到来した」(Collins)
 - NIHがこの革命を作り上げる。**様々な異なったデータ種類に対するアクセスの統合・分析に主導的な役割を果たす。**

国家戦略としての「医療ビッグデータ」

- ゲノム・オミックス医療情報の全国的連携を目指して
 - 各先進病院で保持しているゲノム・オミックス医療情報の全米的な連携へ 臨床ゲノムオミックス医療DB
- NIH : BD2Kの2014年のGrandとしてのDDI（掘起し）
 - 医療におけるデータ科学の全米COE創設
 - Center of Excellence in Data Science
 - Univ. Pitts: Center for causal modeling and discovery of biomedical knowledge from big data
 - UCSC: Center for big data in translational genomics
 - Harvard: Patient-centered information commons
 - その他、コロンビア大学、イリノイ大学など11施設 32M\$
 - Data Scientist 人材養成への予算措置
 - データ発見索引 DDI (Data Discovery Index) Consortium
 - Data discovery index coordination consortium (DDICC)
 - データベースカタログの発展・Pub MEDのDB版
 - UCSD: BioCADDIE（バイオキャディ）を中心にDDI開発の準備を担当
 - BioCADDIE : Biomedical and healthCAre Data Discovery and Indexing Ecosystem
- 米国はすでに戦略的に対応している。わが国は？

オバマ大統領 Precision Medicine Initiativeを開始



2015年1月 大統領一般年頭教書演説

Precision Medicine

個人の遺伝素因・環境素因に合わせた (tailored) 医療
One size fits for all のPopulation 医療とは異なる

趣旨：基本は、個別化医療 Personalized Medicine の概念と変わらないが、目指していたのは診断/治療の個人化ではなく層別化であることを明確化

概念の拡張：Personalized Medicineが標榜された時から10数年経っている

医療ビッグデータ時代の到来による個別化医療の拡張

- (1) 遺伝素因 X 環境(生活習慣)要因のスキーマ重視
SNPや変異 (Genome)だけでなく環境・生活習慣要因(Exposome) の重視、疾患発症は2つの要因の相互作用を明快に強調。電子カルテの臨床表現型 (Clinical Phenome)も疾患発症後には不可欠。3つの成因の重視
- (2) 日常生理モニタリング情報の包摂
モバイルヘルス(mHealth)・wearable sensorによる大量継続情報収集の重視
- (3) ゲノムコホート・Biobankの重視
Precision Medicineを実現する基礎として、ゲノムコホート/Biobankが必要であることを認識。Real world dataの重視

医療ビッグデータ時代における 人工知能の役割の変化

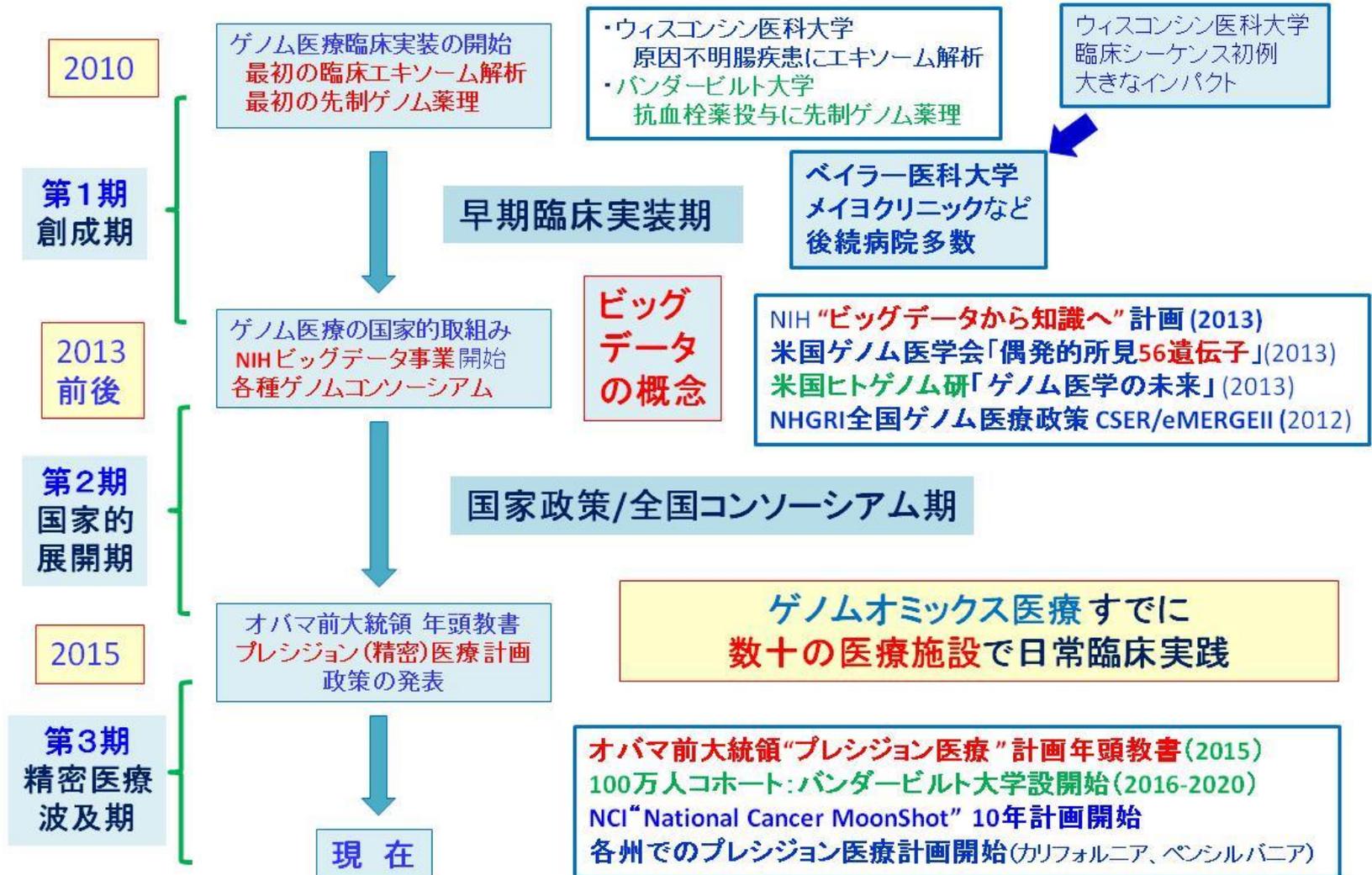
- 網羅的分子情報の急速な蓄積
 - 治療医学における次世代シーケンスの普及による（米国、わが国ではまだ）
- 大規模生命医学知識データベース
 - 疾患に関係する分子情報、パスウェイ情報・薬理情報などそれぞれの知識DBが大規模化
- 大量の論文情報
 - 新規の分子疾患関連情報報告（がん遺伝子など）



膨大なビッグデータを患者の個別化病態に合わせて縮約して特徴表現量を求めたり、適切な機械発見を行い、解釈を提示する役割

ゲノム医療からビッグデータ医療へ

2005～次世代シーケンサ登場
2007～シーケンス革命



「ビッグデータ」のData 原理

問題点 属性値数(p) ≫ サンプル数(n)

p: 数億になる場合あり n: 多くても数万、通常数千



これら膨大な属性変数がすべて独立ならばビッグデータの構造解析は不可能。単変量解析の羅列 (GWASのManhattan Plot) しか可能でない



ビッグデータ・スパース仮説

ビッグデータは、多数であるが属性値数より少ない独立成分が基底となって、相互にModificationして構成されている。
(独立成分の推定は、サンプル数とともに増加する)

データ次元縮約の原理 (**principle of compositionality**)

IBM Watson

- **Learning systemの不可欠性: IBM Watson**
 - 自然言語処理、大量データベース探索、確信度付き解答: **Deep QAシステム** (jeopardy)
 - MITのSTARTと呼ばれるオンライン自然言語QAシステム: 質問をシンプルな質問に分解
 - CMUのOpen Advancement of Question-Answering Initiative (OAQA) システムが骨格
 - 質問解答に最も適切なテキスト資料を特定する知識源拡張アルゴリズム。テキストから知識を自動的に抽出
 - **大規模情報抽出、構文解析、知識推論により大量の情報資料をシステムの一般知識情報源に変換**
 - 自然言語理解に応用される統計学的学習理論 (例えば、カーネル法) が基礎
- **Memorial Sloan-Kettering Cancer Center (MSKCC)**
 - The Oncology Expert Adviser software (OEA)
 - IBMワトソンの計算能力および自然言語処理技術と、MSKCCが持っている**臨床知見** (分子・ゲノムデータ、がん病歴の膨大なリポジトリなど) を組み合わせ、個々の患者にとって最高の治療方針を決定するのに役立つ、最新の研究に基づいた詳細な診断情報や治療の選択肢を見出す
- **New York Genome Center**
 - がん専門医ががん患者に対してより良い個別ケアを提供できるよう支援するツールとしてゲノム研究専用デザインされたWatsonの試作システム
 - 最初の対象: 脳腫瘍のglioblastoma (グリア芽細胞腫)、ゲノム配列と医療情報、医学文献から**個別化治療**を提案
- **東京大学医科学研究所**
 - Watson for Genomics (WfG) を66歳女性「**急性骨髄性白血病**」と診断されて医科研病院に入院
 - 2種類の抗がん剤治療を半年続けたが回復が遅く敗血症などの危険も出た。
 - がんに関係する女性の**1500の遺伝子変異情報**をWatsonに入力。STAG2遺伝子 (cohesin複合体のサブユニット: 染色分体欠損) の変異原因と発見
 - WfGは急性骨髄性白血病の「**二次性白血病**」タイプと診断。
WfGは抗がん剤を別のものに変えるよう提案。
 - 女性は**数カ月で回復して退院し、現在は通院治療**を続けている
 - Watsonが治療法を助言した同様な例、医科研で41例
 - がん研究に関連する約2000万論文、1500万件以上の薬剤関連情報を学習

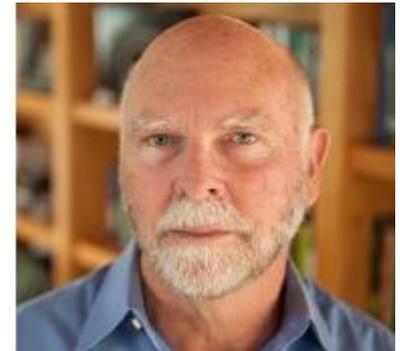


東京大学医科学研究所の Watson for Genomics

- Watson for Genomics (WfG) と東大医科研と共同研究
 - がん研究に関連する約2000万論文、1500万件以上の薬剤関連情報を学習
- 66歳女性「急性骨髄性白血病」と診断されて医科研病院に入院
 - 2種類の抗がん剤治療を半年続けたが回復が遅く敗血症などの危険もあった。
- がんに関係する女性の**1500の遺伝子変異情報**をWatsonに入力。
- STAG2遺伝子（cohesin複合体のサブユニット：染色分体欠損）の変異原因と発見
- WfGは急性骨髄性白血病の「**二次性白血病**」タイプと診断。
WfGは抗がん剤を別のものに変えるよう提案。
- 女性は**数カ月**で回復して退院し、**現在は通院治療**を続けている
- Watsonが治療法を助言した同様な例、医科研で41例

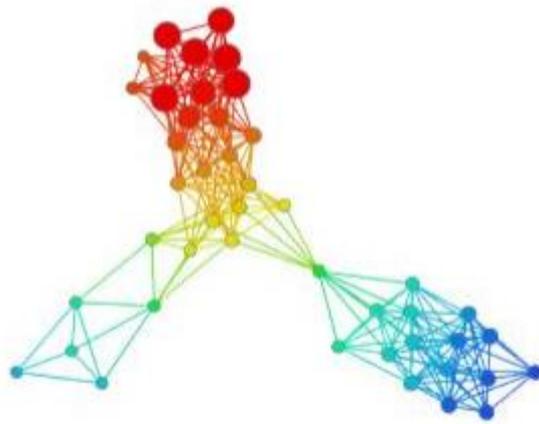
そのほかの機械学習

- **The ASCO (米国臨床癌学) CancerLinQ initiative**
 - 診療の現場(EHR)から大量の診療データを集め分析
 - 17万人のがん症例データベースを構築。新しい臨床治験へのガイドライン作
 - 各がん1～2万人の症例を集める
 - 学習システムを構築し治療知識を統計学習、ニューロネットを駆使して学習。
BigDataにおけるLearning systemの不可欠性
 - 2013年に、CancerLinQのプロトタイプを完成、10万人以上の乳がんを蓄積、完全規模へ継続構築中
- **Cancer Commons initiative**
 - Rapid learningのインフラ整備
 - 目的：患者の個別症例と最新の知識を更新
 - 個々の患者の”Donate Your Data”(DYD)登録
- **Craig Venter “Human Longevity Inc.”**
 - 健康寿命伸長のための**ゲノム科学、幹細胞治療**
 - 初期資本7000億円・医療費削減、HiseqX 5sets
 - 一年**40000ゲノム**（幼児から老人まで、患者・健常者も）収集し
最大のゲノムDBを作る、臨床情報も収集、腸内細菌も含む 一日5人のヒト全ゲノム
 - がん（Mores Cancer Centerと提携）、糖尿病、認知症などの成人疾患に
- **Google X project, “Baseline”**
 - 健康に関する尺度発見、Conrad AのもとにDuke大学やStanford大学が協力
 - 現在175名、先制医療的なバイオマーカ探し、今後拡大

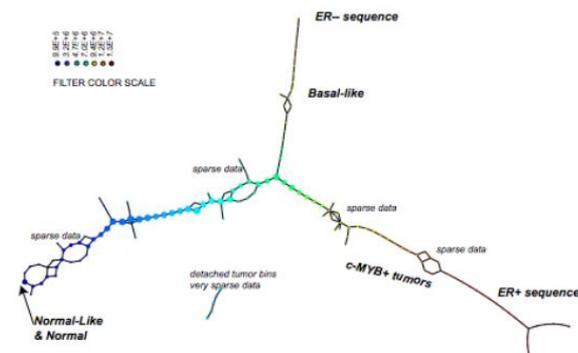


疾患のTopological Data Analysis

- 患者の網羅的分子情報や病態ビッグデータ
- Topological Data Analysis (TDA) を用いた機械知能以“病気の形”を描き臨床家に理解を容易にさせ、治療方針を決定させる



糖尿病のTDA病型



乳がんのTDA病型

DL型NNの医療応用への期待

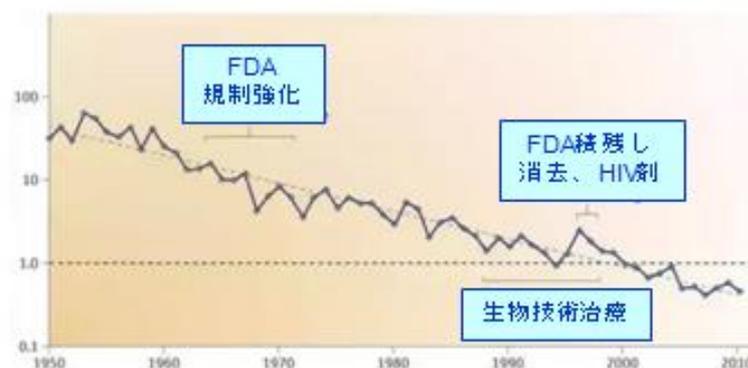
- DLの医療の応用は開始段階で応用成功例は少ない
- 本質的に「教師なし学習」:人間が思いつかない解を提示
- 画像分類・解釈と文章理解が優れているので、遺伝子発現プロファイル解析や病態推移の理解への応用が期待される
- いくつかのDeep Learningを用いた医療応用
 - ヒトmicrobiomeの分類・階層的表現を得た
 - 6つのがんで遺伝子発現をmiRNAとともに分類した。
 - 異なったMicroarrayを含むがん発現を分類の特徴表現を導き分類した。
 - Convolution ネットワークを使用して画像としての遺伝子発現を分類した。
 - 遺伝子発現プロファイルの自動アノテーション

AI創薬

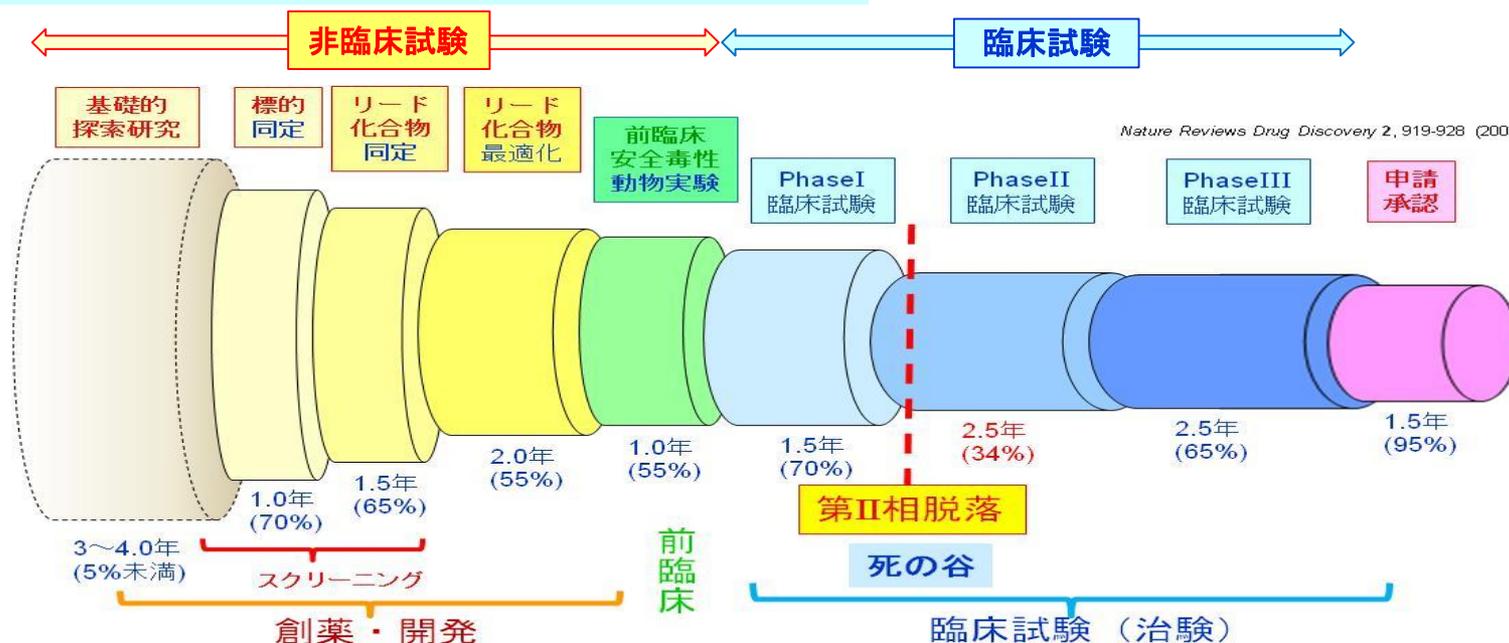
創薬をめぐる状況

- 医薬品の開発費の増大
 - 1 医薬品を上市するのに約1000億円以上
- 開発成功率の減少
 - 2万~3万分の1の成功率
 - とくに**非臨床試験**から**臨床試験**への間隙
 - **phase II attrition** (第2相脱落)
- 臨床的予測性
 - 医薬品開発過程の**できるだけ早い段階**での**有効性・毒性の予測**
- **臨床予測性の早期での実施**
 - 罹患者のiPS細胞を使う
 - ヒトの薬剤 - 生体関連のビッグデータを使う

10億ドル開発費で薬剤数



Nature Reviews Drug Discovery 11, 191-200 (2012)



Nature Reviews Drug Discovery 2, 919-928 (2003)

ドラッグ・リポジショニング

薬剤適応拡大

ヒトでの安全性と体内動態が十分に分かっている
既承認薬の標的分子や作用パスウェイなどを、体系的・論理的・網羅的に解析することにより**新しい薬理効果**を発見し、その薬を別の疾患治療薬として開発する創薬戦略

利 点

- (1) 既承認薬なので、ヒトでの安全性や体内動態などが既知で臨床試験で予想外の副作用や体内動態の問題により開発が失敗するリスクが少なく**開発の成功確率が高い**
- (2) 既にあるデータや技術（動物での安全性データや製剤のGMP製造技術など）を再利用することで、**開発にかかる時間とコストを大幅に削減できる**
- (3) **DR候補探索に疾患生命情報ビッグデータ知識DB**を使用できる。

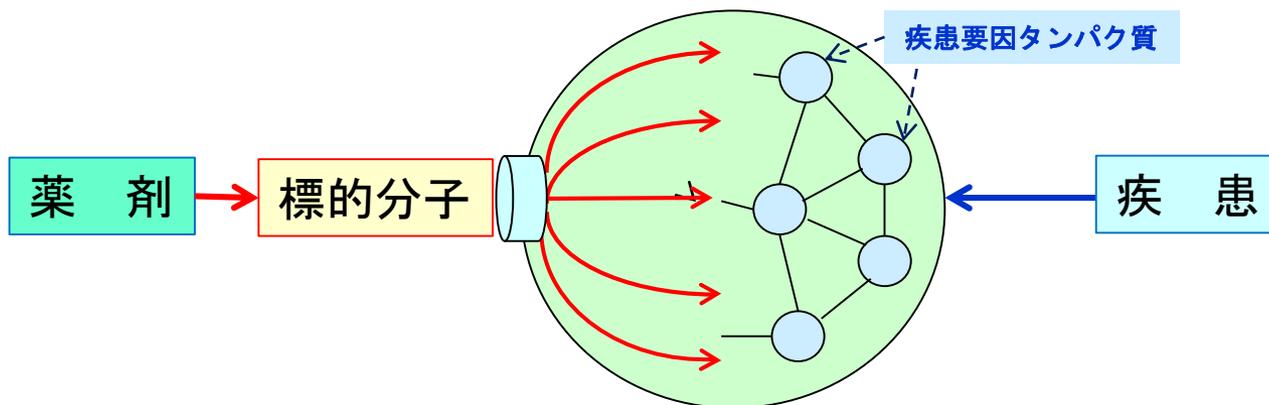
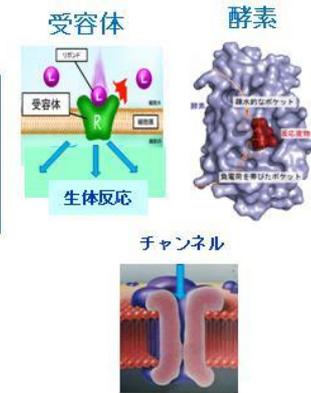
疾患・薬剤・標的の関係

病気の主要な要因

疾患要因タンパク質（複数）

薬：疾患要因タンパク質に影響を示す
標的タンパク質に作用し阻害する

薬剤の標的分子
受容体・酵素・チャンネルなど

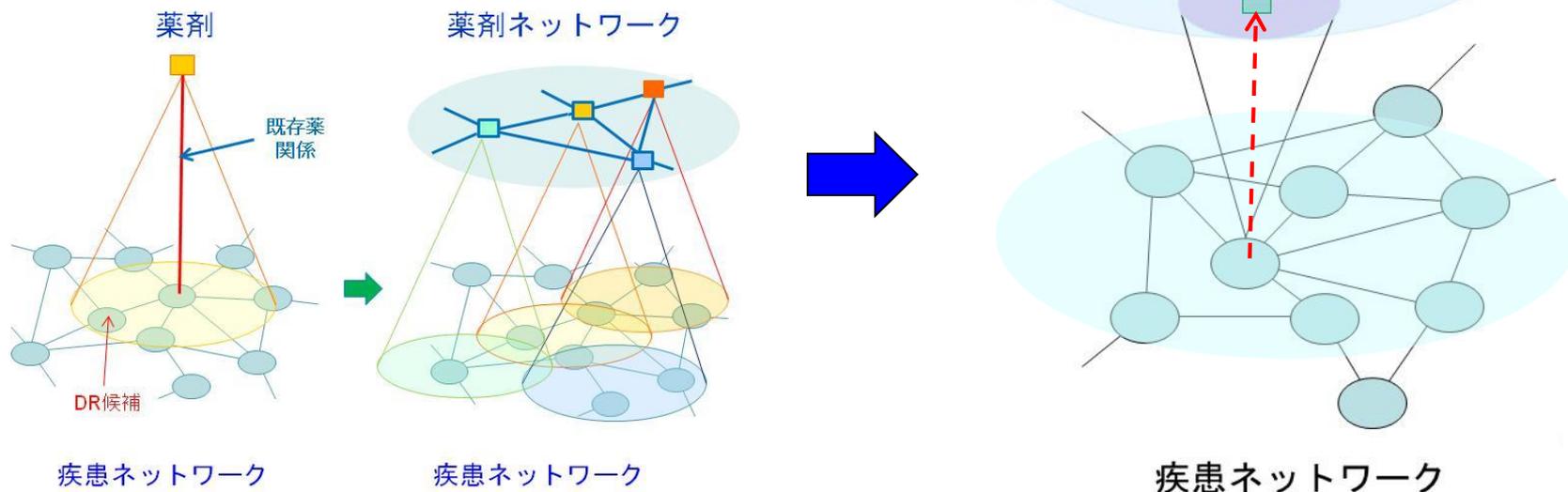


生体システム/ネットワーク

DRの方法論から創薬方法論へ

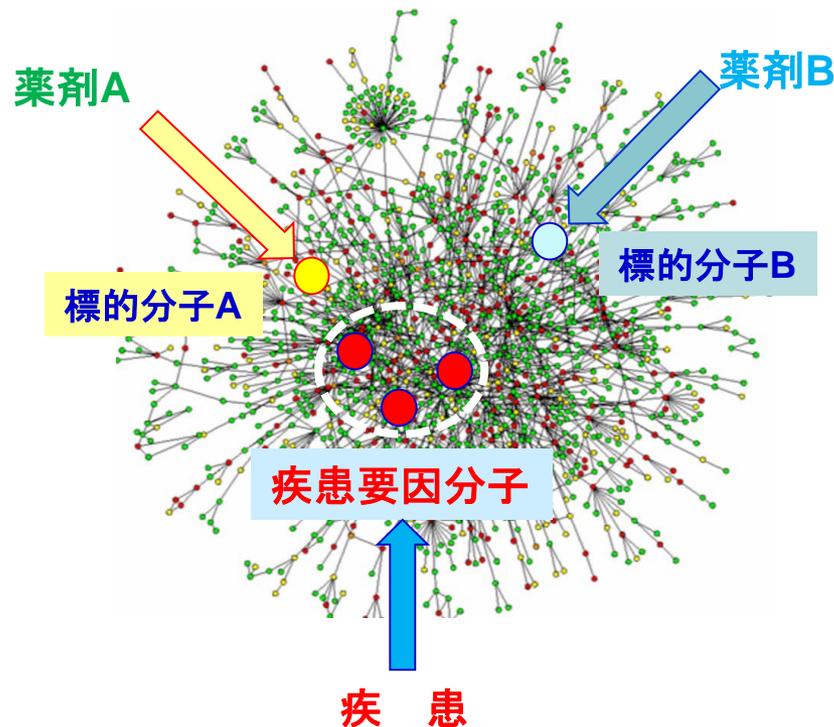
- **疾患ネットワークの十全な形成**
 - 内在的機序の近親性から疾患ネットワーク
 - 医薬品の有効性・毒性の近傍 Projection
⇒ DRにおける**有効性**はすでに確立
- **創薬への展開**
 - 薬剤階層のネットワークは既に確立
 - 投与時生体反応の近親性だけではなく
 - 化合物の構造的近親性(finnger print) からも作成
 - 疾患から逆投影。創薬の可能性探索
- **疾患ネットワークと薬剤ネットワーク間写像**
 - **双方向性・対等性**

疾患から薬剤ネットワークへの**逆投影**
Multi-Topology 双対写像 創薬方法論



標的分子や疾患要因分子の タンパク質相互作用ネットワーク (PPIN)

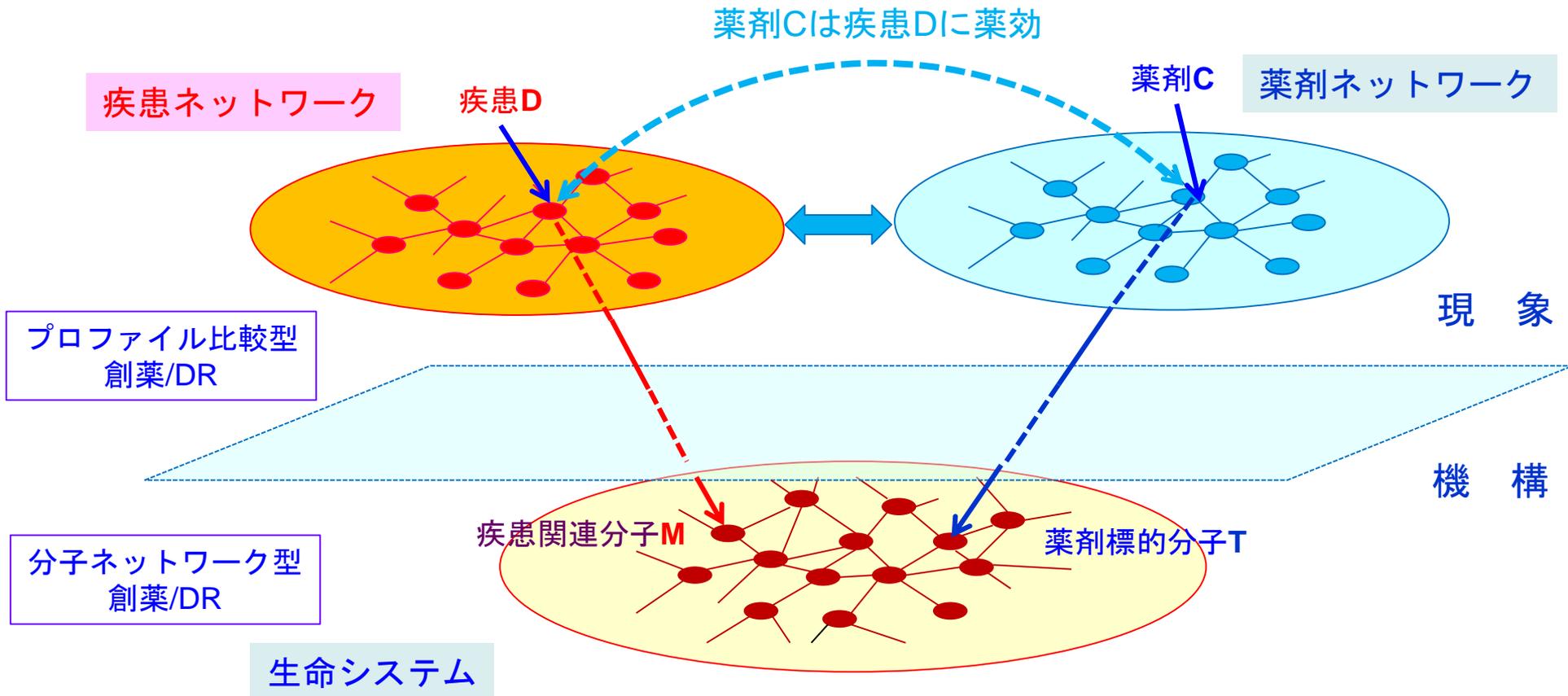
- 薬剤ネットワークと疾患ネットワークを媒介する第3の生体ネットワーク
- タンパク質相互作用ネットワーク (PPIN) での創薬/DR戦略
- PPIネットワーク場を基礎にして距離 (類似性) を検討
- **薬 剤** : 薬剤の**標的分子** (タンパク質) によって PPI場と繋がる
- **疾 患** : 疾患特異的発現遺伝子を**疾患要因分子** (タンパク質) へ翻訳、
- PPIN場内での**薬剤標的分子**と**疾患**の「**代理人(疾患遺伝子)**」の**距離・親近性**を基準に、**薬理作用のインパクト力**を評価



タンパク質相互作用
ネットワーク (PPIN)

プロフィール型計算創薬の原理

3層生体・薬剤ネットワークのFramework



Deep Learningによる 多次元ネットワーク縮約法

(Hase, Tanaka 2017)

- 医療・創薬ビッグデータへの応用性高い
- 超多次元ネットワーク情報構造の急増
 - ゲノム医療<網羅的分子情報–臨床表現型情報>
 - ゲノムコホートにおける<遺伝子情報–環境（生活様式）情報>
- Deep Learning-based Network Contraction
「DLネットワーク縮約法」

超多次元ネットワーク情報構造⇒
少数の特徴的ネットワーク基底の〈和〉
- 線形分解ではない。非線形分解で基底への射影

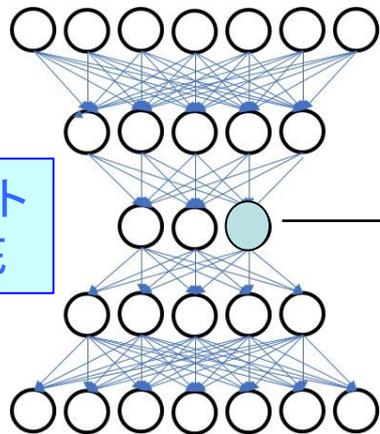
特徴的ネットワーク基底への分解

特徴的ネットワーク基底の和に縮約

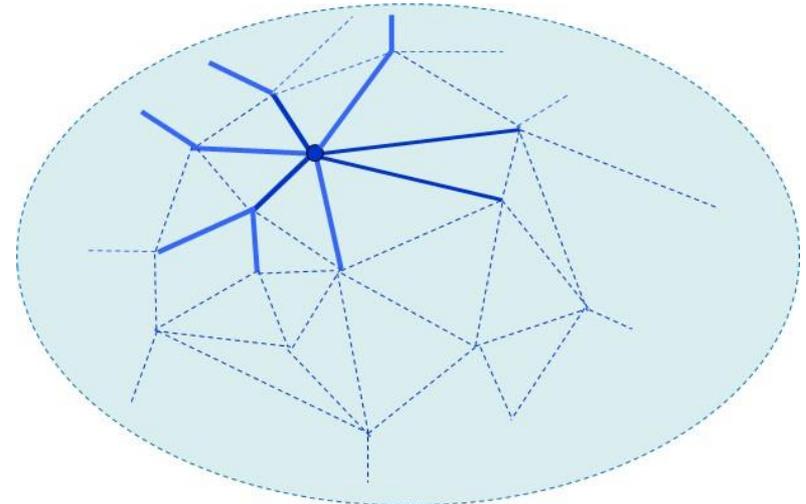
特定のノードを起点とした素NW（部分NW）の集合
全体NWを包摂する集合にDL反復自己学習

特徴的ネットワーク基底：トポロジーのみの構造/頻度構造

素NWの集合(全体包摂)



特徴的ネットワーク基底



Deep Learningによる創薬・DR

1) 生体ネットワーク (PPIN) 特徴量の抽出

- タンパク質相互作用ネットワーク(PPIN)のNW結合を学習し**特徴表現** (特徴NW基底) を出力。
- 学習集合を部分ネットワークの集合から決める
- ノードを起点とした素NWでPPIN全体を覆う集合

2) 多層Stacked Auto-encoderのDLで学習

- 特徴的NW基底の「教師無し」学習
- 次元縮約による特徴的NW基底の抽出

3) DL特徴NW基底空間における正例補完

- DrugBankからの正例とその増加 (SMOTE法)

4) DL特徴NW基底量を用いた機械学習分類

- Xgboot法などを用いたDL特徴量からの判別ネットワーク・タンパク質の標的性の判定

Deep Learningによる創薬・DR

分類部 DrugBankを利用した 当該分子を標的とする既製薬剤の探索

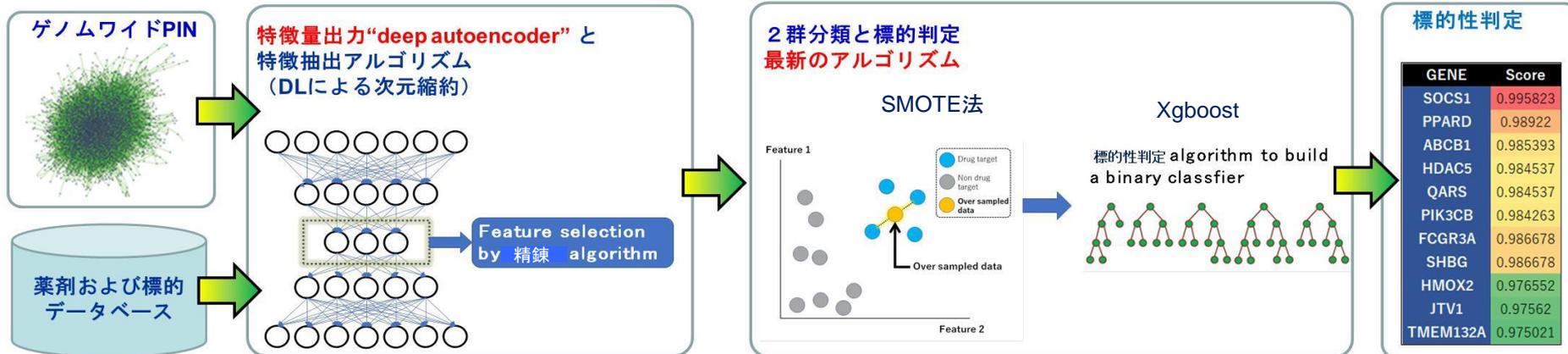
既製薬剤がない→新規薬剤探求（創薬）
既製薬剤がある→DRの検討

入力

特徴量産出

分類モデル

標的選定



従来の機械学習 (Random Forrest)と同じ成果は得られている

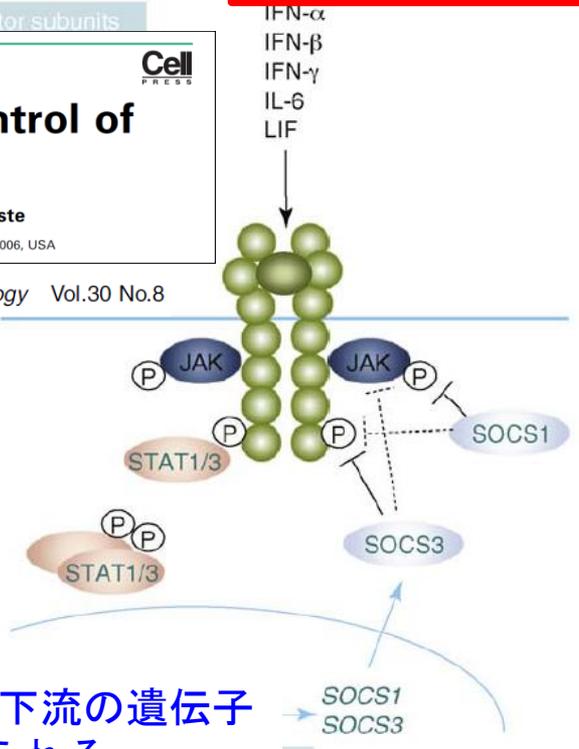
<疾患- 標的分子> 予測結果

Target disease	Top 25 genes for potential novel drug target for a target disease
Anti-Alzheimer's disease	SOCS1 ; BDNF; F2; PRKCS; YES; UBB; HSP110; HSP70; HSP90AA1; HSPG2; PARD6A; BLNK; PRKCI; YES1; PPP3CA; MMP11; SPTAN1; PTPRC; FBLN1; RPLP1; FAM107A; TRADD; NR4A1; LAT; NF2; PRKCE; KIT; NID1
Anti-Anxiety	S1PR1; CNR1; MTNR1A; CCL4; F2; TEC; IL8; CRHR1; AGTR2; OPRD1; IL8RA; RNF43; RHO ; SP6; RAB13; DRD4; IL8RB; MMP9; MMP2; OPRM1; IL1B; GNAS; S1PR3; KIT; GRM2
Anti-Rheumatoid	SLC22A5; GRASP; KIT; SLC22A4; CFH; COX2; HSP90AA1; UBB; DHRS3; SCTR; ADORA1; MUC1 ; C6orf47; NR4A1
Anti-Breast Cancer	SHC1; NFKB1; RELA; ID2; RAC1; SRC; JUNX1; HDAC2; I
Anti-Colorectal	SHC1; GHR; HRAS; PLCG1; NFKB1; MUC1; LYN
Anti-Pancreatic	SH2; DSCC1; TAF9; AIF1; IL2; NSMCE4A; ASPH;
Anti-Melanoma	RAP; HMGB2; FEN1; MUC1; LMNB1; GDE1;

SOCS1はJAK/STAT pathwayを介してサイトカインの応答を変動させ、中枢神経系の炎症を制御

Review
SOCS1 and SOCS3 in the control of CNS immunity
 Brandi J. Baker, Lisa Nowoslawski Akhtar and ETTY N. Benveniste
 Department of Cell Biology, The University of Alabama at Birmingham, Birmingham, AL 35294-0006, USA

Trends in Immunology Vol.30 No.8



しかし、SOCS1は上流の遺伝子なので、この下流の遺伝子を標的にした方が、長期投与には良いとも考えれる。

疾患-標的分子リンクの同定よりDRへ

機械学習で予測された、新規標的の情報(disease A と target の情報, 標的が disease A の新規標的分子、青いリンク)を、既知の drug-target-disease interaction network をマップし薬剤の新しい適用疾患 (赤リンク) を予測

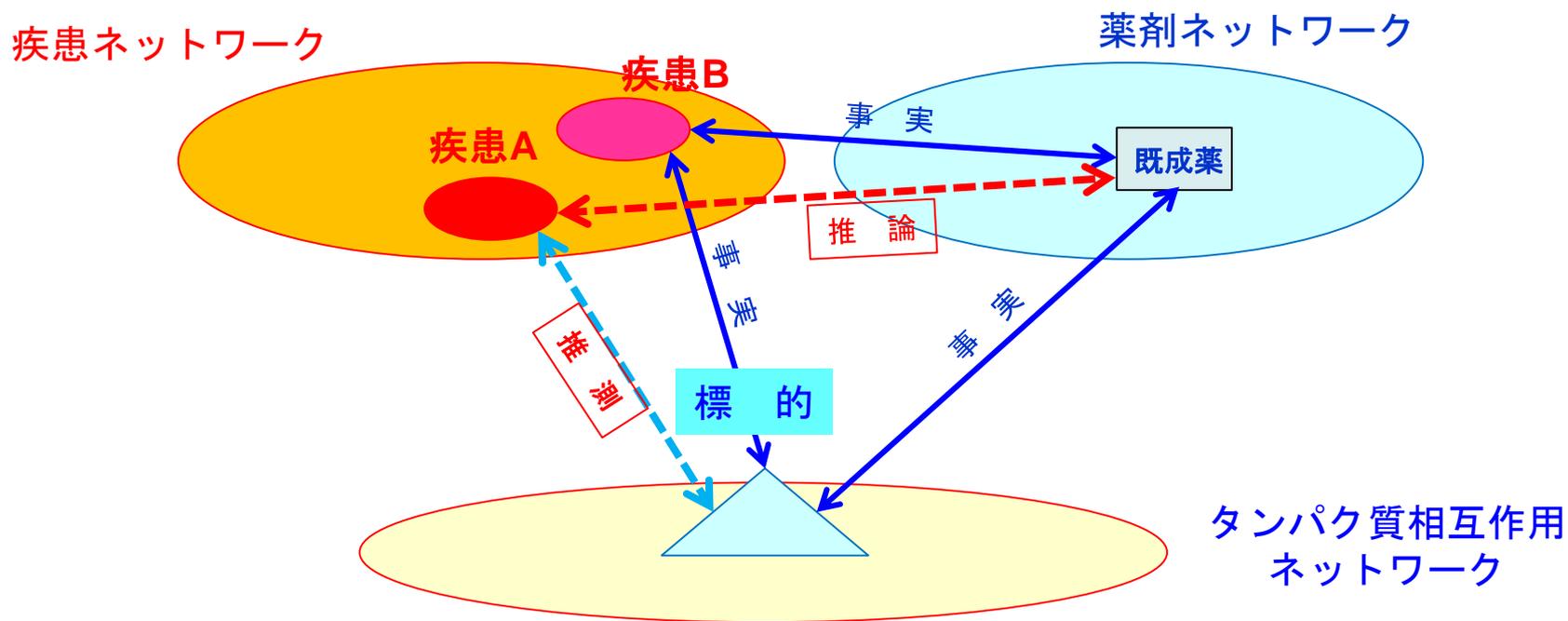


Table 3. Predicted repositionable drug candidates

Target disease	Candidate repositionable drug
Anti-Alzheimer's disease	Imatinib; Marimastat; Nilotinib ; Regorafenib; Sorafenib; Tamoxifen; Urokinase
Anti-Anxiety	3-Methylfentanyl; Agomelatine; Amitriptyline; Amoxapine; Antihemophilic Factor; Apomorphine; Aripiprazole; Bromocriptine; Buprenorphine; Butorphanol; Cabergoline; Canakinumab; Captopril; Chlorpromazine; Coagulation Factor IX; Codeine; Dextromethorphan; Dextropropoxyphene; Dopamine; Drotrecogin alfa; Ethylmorphine; Etorphine; Fentanyl; Halothane; Hirulog; Hydrocodone; Hydromorphone; Imatinib; Ketamine; Ketobemidone; L-DOPA; Lepirudin; Levallorphan; Levorphanol; Lisuride; Loperamide; Loperidine; Marimastat; Melatonin; Menadione; Methadone; Methadyl Acetate; Methotrimeprazine; Minocycline; Morphine; Naloxone; Naltrexone; Nilotinib; Olanzapine; Ondansetron; Oxycodone; Oxymorphone; Paliperidone; Pergolide; Pethidine; Pramipexole; Promazine; Propiomazine; Quetiapine; Regorafenib; Remifentanyl; Remoxipride; Risperidone; Ropinirole; Rotigotine; Sorafenib; Sufentanil; Suramin; Thiothylperazine; Ziprasidone
Anti-Rheumatoid	Acetylcholine; Adenosine; Amiloride; Aminohippurate; Aminophylline; Amphetamine; Ampicillin; Azidocillin; Benzylpenicillin; Cefalotin; Cefdinir; Cefixime; Cephalexin; Choline; Cimetidine; Clonidine; Cyclacillin; Desipramine; Diphenhydramine; Dopamine; Dyphylline; Enprofylline; Epinephrine; Furosemide; Grepafloxacin; Histamine Phosphate; Imatinib; Imipramine; Insulin, isophane; Ipratropium bromide; L-Arginine; L-Carnitine; Levofloxacin; Lidocaine; Liothyronine; Lomefloxacin; Mepyramine; Methamphetamine; Nicotin; Nicotine; Nilotinib; Norepinephrine; Norfloxacin; Ofloxacin; Oxtriphylline; Pentoxifylline; Probenecid; Procainamide; Quinidine; Quinine; Regorafenib; Rifabutin; Secretin; Sorafenib; Spermine; Testadenoson; Testosterone; Tetraethylammonium; Theophylline; Thiamine;

Neuroscience 304 (2015) 316–327

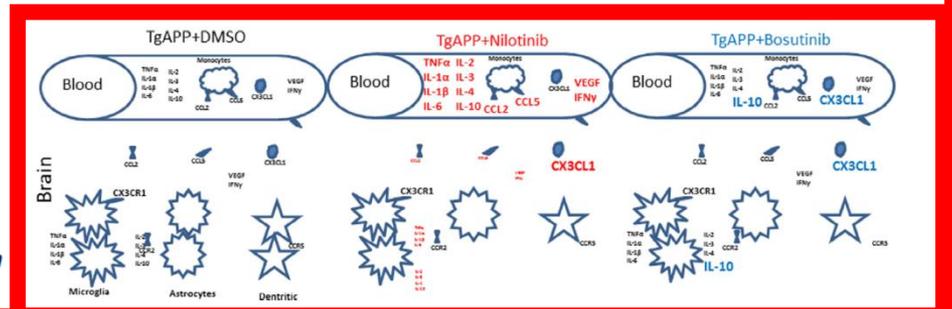
慢性白血病の抗がん剤であるニロチニブがアルツハイマーのDR薬剤として選定

NILOTINIB AND BOSUTINIB MODULATE PRE-PLAQUE ALTERATIONS OF BLOOD IMMUNE MARKERS AND NEURO-INFLAMMATION IN ALZHEIMER'S DISEASE MODELS

I. LONSKAYA,^a M. L. HEBRON,^a S. T. SELBY,^a
R. S. TURNER^b AND C. E.-H. MOUSSA^{a*}

^a Department of Neurology, Laboratory for Dementia and Parkinsonism, Georgetown University Medical Center, Washington D.C. 20007, USA

^b Department of Neurology, Memory Disorders Program, Georgetown University Medical Center, Washington D.C. 20007, USA



今後の戦略・方向

1. DLによる多次元ネットワーク情報構造の縮約はビッグデータ医療・創薬に適用可能
AI創薬の「枠組み」・実行方向は「見えてきた」
2. 本年中に、いよいよAI創薬の実装に着手しなければならない
3. 製薬企業、IT企業、医療機関を束ねた集中的プロジェクトを推進するために「ビッグデータ医療・AI創薬コンソーシアム」設立する

今後の戦略・方向

- DLによる〈多次元ネットワーク情報構造〉の縮約
 - 創薬だけでなく、ビッグデータ医療への適応可能
 - ゲノム医療の〈網羅的分子情報－臨床表現型〉の相関ネットワーク構造
 - バイオバンクの〈遺伝素因－環境要因〉と発症
- AI創薬の「枠組み」実行方向は「見えてきた」
- 本年中に、いよいよAI創薬の実装に着手しなければならない
 - 製薬企業、IT企業、医療機関を束ねた集中的プロジェクトを推進するために「ビッグデータ医療・AI創薬コンソーシアム」を設立する

ご清聴有難う御座いました

