

# ビッグデータ医療・創薬における 人工知能の展望

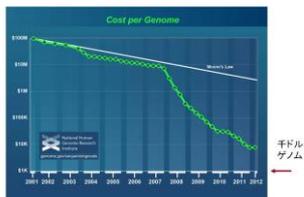
東京医科歯科大学 医療データ科学推進室  
東北大学 東北メディカル・メガバンク機構

田中 博



# 医療ビッグデータ時代の到来

- (1) 次世代シーケンサ (Clinical Sequencing)による「ゲノム/オミックス医療」における網羅的分子情報収集/蓄積
- (2) Biobank/ゲノムコホート普及による分子・環境情報の蓄積
- (3) モバイルヘルス(mHealth) によるWearable センサの連続計測による生理データの蓄積 (unobstructed monitoring)



DNA Sequencing Cost: the National Human Genome Research Institute

急激な大量データの出現  
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・医療の国民レベルの向上  
医療/ヘルスケアの適確性の飛躍的な増大



# ビッグデータ医療の2つの流れ

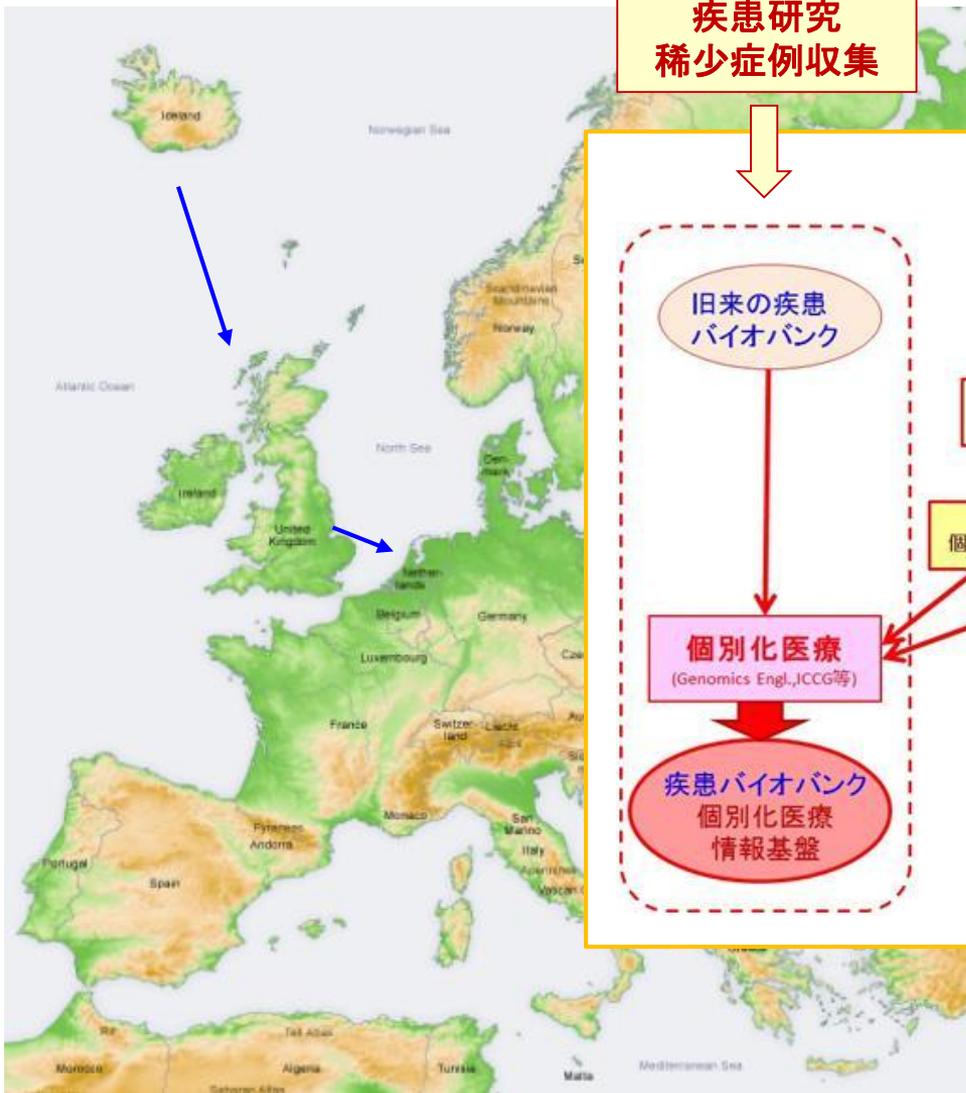
- 米国の流れ

- 次世代シーケンサの急激な発展による「シーケンス革命」からの怒濤の展開（2010から）
- 「治療医学」レベル質的向上のためにゲノム情報を取り入れた臨床実装の推進
  - 稀少疾患の原因遺伝子変異の同定
  - がんのドライバー遺伝子変異の同定と分子標的薬の選択
  - 薬剤代謝酵素の多型性の同定と個別化投与

- 欧州の流れ

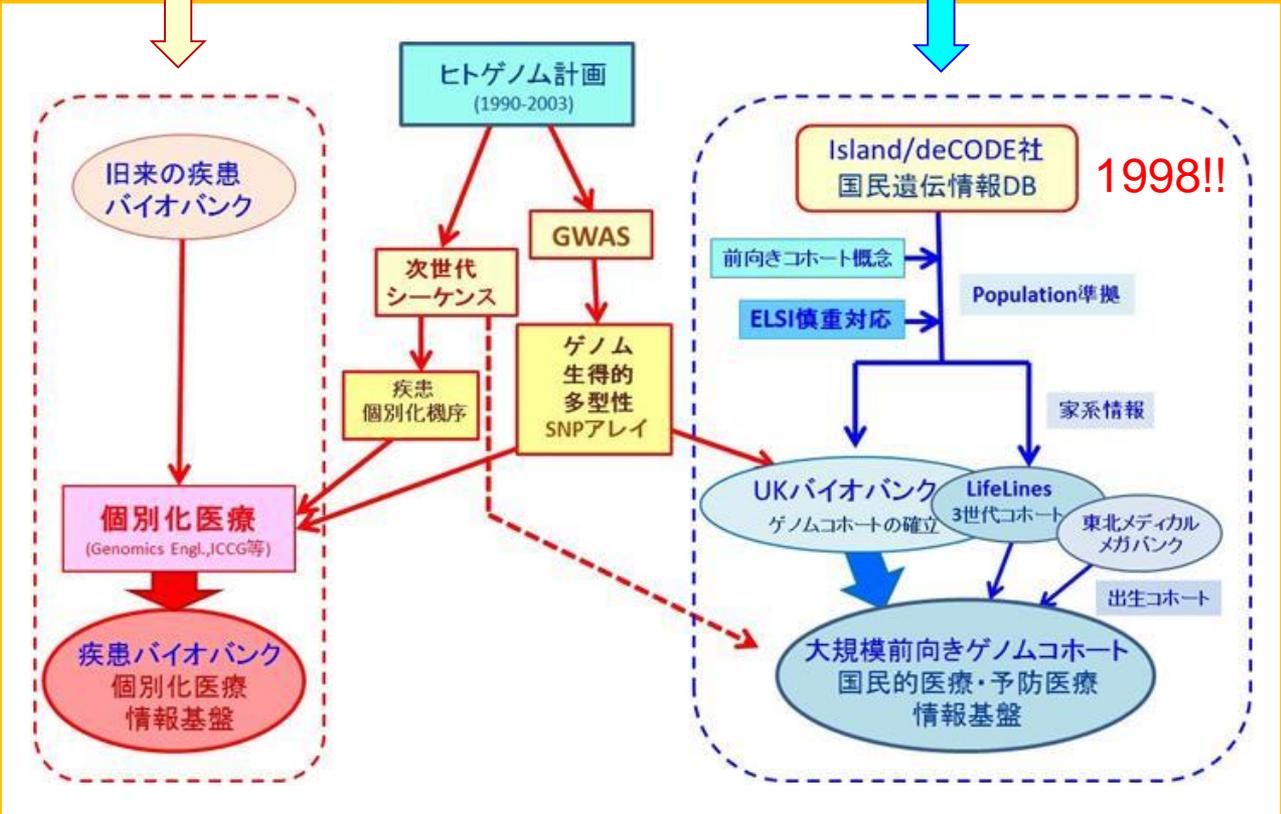
- 社会福祉国家の理念より国民医療（医療の国民レベル）の向上
- 「予防医学」レベル質的向上のためにゲノム情報を取り入れたバイオバンク推進
- 大規模前向きpopulation型バイオバンク/ゲノム・コホートの確立
  - 遺伝的素因だけでなく環境要因（生活習慣）との相互作用を解明し、「ありふれた疾患」発症を予測し、これに基づいて個別化予防する。
  - 疾患を発症前に対応して発症を防ぐ「先制医療(preemptive medicine)」や「予測医療(predictive medicine)の実現を目的

# 第2の流れ 欧州のバイオバンクの普及



疾患研究  
稀少症例収集

「集合的遺伝情報」による  
国民レベルでの医療向上

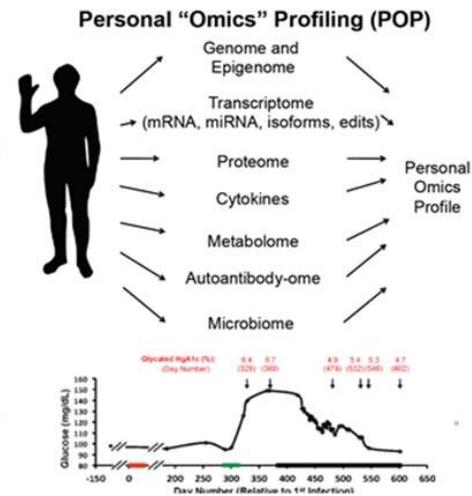
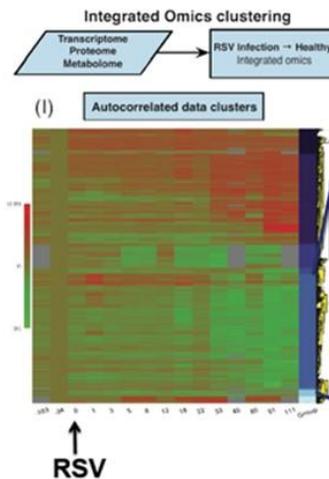


# もう一つの医療ビッグデータ第3の流れ モバイルヘルス

- Quantified Self
  - 米国での運動、Wearable Computerと生体センシングを結合して自己の健康・行動をモニターする。サンフランシスコから広がる
  - 東北大学 - 東芝COI
    - 「さりげないセンシングと日常人間ドックで実現する理想自己」
- Dr. John Halamka
  - 埋め込み式マルチセンサー
- Dr.Snyder
  - Integrated personal omics profile (iPOP)



ECG; EEG; Skin Conductivity; EVG



# 医療におけるビッグデータ

## I 次世代シーケンサによるゲノム情報

— 網羅的分子情報の急速な蓄積

## II 大規模バイオバンクによる情報蓄積

— ゲノム情報・環境生活関連情報

## III モバイルヘルスによる生理変量

— 連続的生理モニターによる情報蓄積

新しい  
タイプの  
医療ビッグ  
データ

---

## IV 表現型医療情報の蓄積

— 電子化の普及による医療情報の蓄積

旧来のタイプの  
医療データの  
大容量化

# 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(IV)の「ビッグデータ」

$N$ -Big Data ( $n \gg p$ )

医療情報・疫学調査では属性数：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(I~III)などビッグデータ

$P$ -Big Data ( $p \gg n$ )

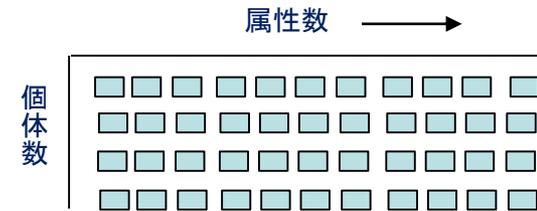
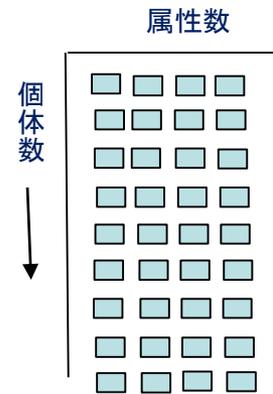
1個体に関するデータ属性種類数が膨大

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

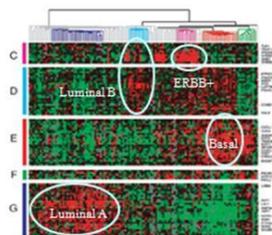
# ビッグデータは 医療のパラダイムを変革する

- 医療は近年大きくパラダイム変換しつつある。
- 2000年頃から、「ビッグデータ医療」の概念出現の前に、パラダイム変換の概念として、次の2つの概念が提示されてきた

## 個別化医療

Personalized Medicine

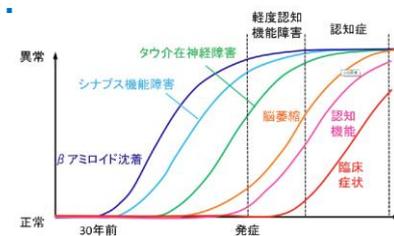
- ・ 従来のpopulation医学<One size fits for all>はもはや成り立たない
- ・ 同一の病名で括られているが、内在的亜型が多数存在
- ・ 医療の隅々に浸透するポピュレーション医学の桎梏克服



## 先制医療

Preemptive Medicine

“By making use of precise molecular knowledge (分子情報を使用) to detect disease before symptoms are manifest, and intervening before disease can strike.”  
(発症前に検出・治療)



# ビッグデータ医療はパラダイム 変換を実現可能にする

ビッグデータの医療・創薬への効果(20~30年)

I ゲノム・オミックス情報

疾患成立機序の解明

II 大規模バイオバンク情報

長期生涯疾患過程の解明

III モバイルヘルス情報

短期生涯疾患過程の解明

Disease Big Data

個別化医療・先制医療

生涯医療・先制医療

Life-long Big Data

# 21世紀医療の長期世代交代

- 第1世代(1930~1970)

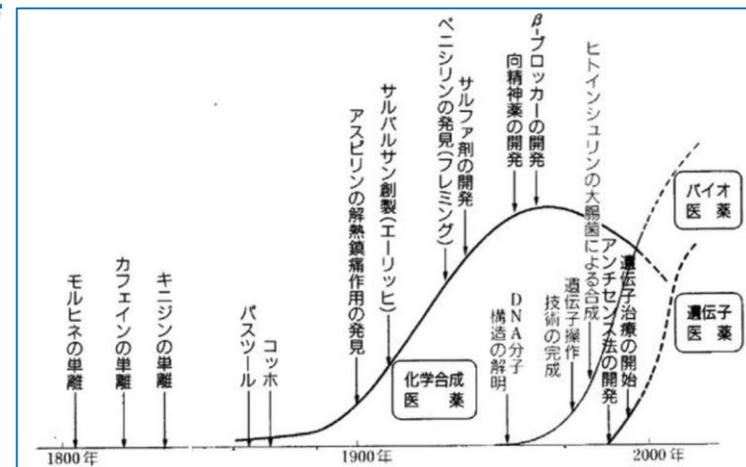
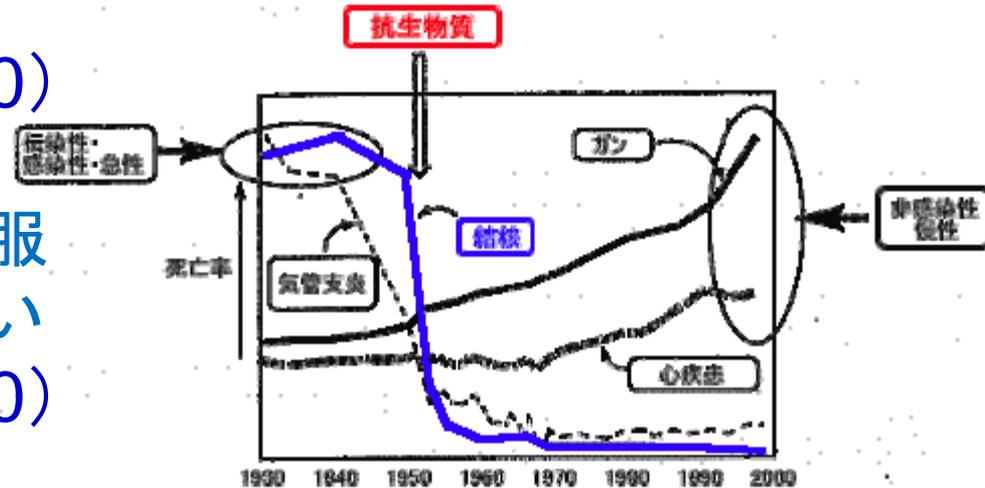
- 抗生物質の登場
- により細菌感染症の克服
- 疾病の病原菌との闘い

- 第2世代(1970~2010)

- 分子生物学の発展
- 分子的機序による疾患との闘い
- 分子標的薬・抗体医薬の登場

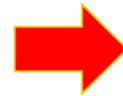
- 第3世代 (2010~2040)

- 網羅的分子情報・
- モバイルヘルスの発展
- ビッグデータ・AIの登場
- データ駆動型医療の登場



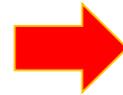
# 医療の長期的パラダイム変換

Population医療



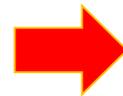
個別化医療

Reactive 医療



Proactive 医療

Occasional 医療



Life-long 医療

# 米国のゲノム・オミックス医療の 3つの流れ

2008年

2009年

2010年

2011年

2012年

2013年

2005~ NGSの登場  
(454, Solexa, SOLID)  
2007/8~  
**シーケンス革命**

ゲノム多型性の認識  
.Hapmap2002開始  
GWAS研究の興隆

TCGA (2006), 国際  
がんコンソーシアム  
ICCG(2008) の  
成果2011から出現

Undiagnosed  
Disease原因遺  
伝子のPOC同定  
**MCW小児病院**

薬剤代謝酵素多型性  
電子カルテで警告

**裸のヒトゲノムの異常から発症する  
疾患しか対象としていない!!**



抗がん剤治療  
**Dana Faber**

**ゲノム・オミックス医療  
臨床実装(clinical implementation)**

# ゲノム・オミックス医療の進展とビッグ・データ

2005~ NGS登場 (454 Life sci)  
2007~ シーケンス革命



2010

ゲノム医療臨床実装の開始  
臨床WESの最初 (MCW)  
先制PGxの最初 (VU)

- MCW Nic君原因不明腸疾患 WES XIAPの変異同定・骨髄移植
- Vanderbilt preemptive PG (PREDICT計画) 開始

Wisconsin医科大学  
臨床シーケンス初例  
大きなインパクト

第1世代

Early adopter  
時期

Baylor医科大学  
Mayo Clinicなど  
後続病院多数

2013  
前後

ゲノム医療の国家的取組み  
NIH "BD2K" initiative 開始  
各種ゲノムコンソーシアム

ビッグ  
データの  
概念

- NIH "Big Data to Knowledge" 計画 (2012/13)
- ACGM incidental finding list 56 genes (2013)
- NACHGR report "Future is here" (2013)
- CPIC guideline, EGAPP guideline 2013.14

第2世代

国規模の計画/全国Consortium  
時期

2015

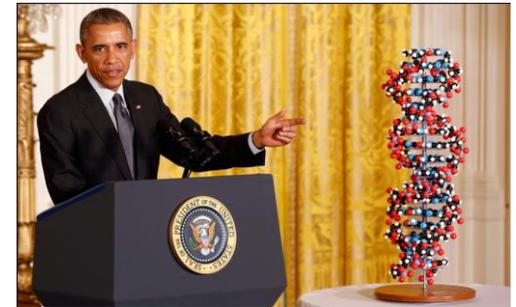
オバマ大統領 年頭教書  
Precision Medicine initiative  
政策の発表

ゲノムオミックス医療 すでに数十の医療  
施設でG/O医療が病院の日常臨床実践

- NIH "BD2K" COE in Data Science, DDI (2014)
- ASCO "CancerLinQ", Cancer Common
- "Precision Medicine (Obama)" 1 M genomic cohort

# Precision Medicine Initiative

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



2015

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

# ビッグデータ医療・創薬における 人工知能の期待

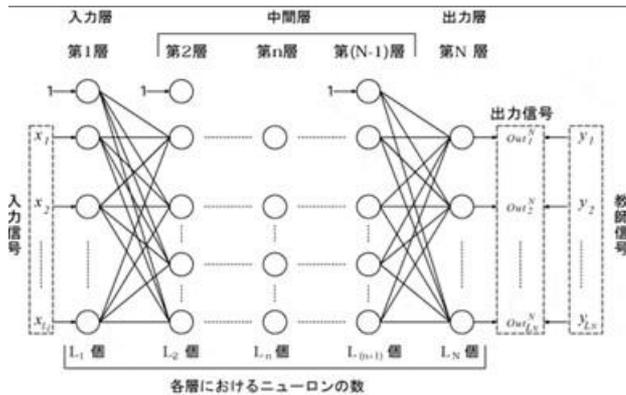
# 人工知能の最近の話題

- 「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
  - 「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。**人間の知識を投入していない。**
  - **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。
- 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



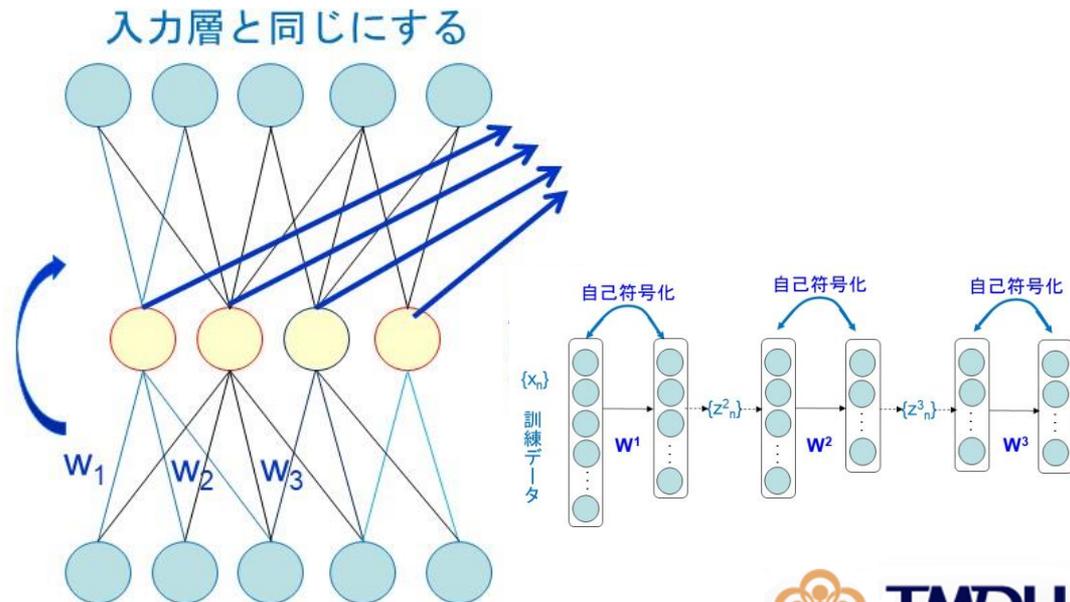
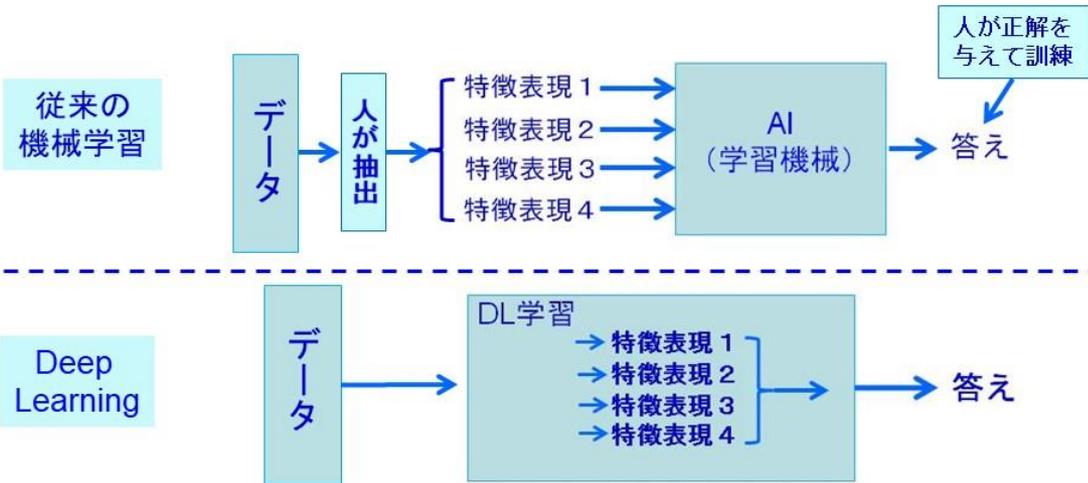
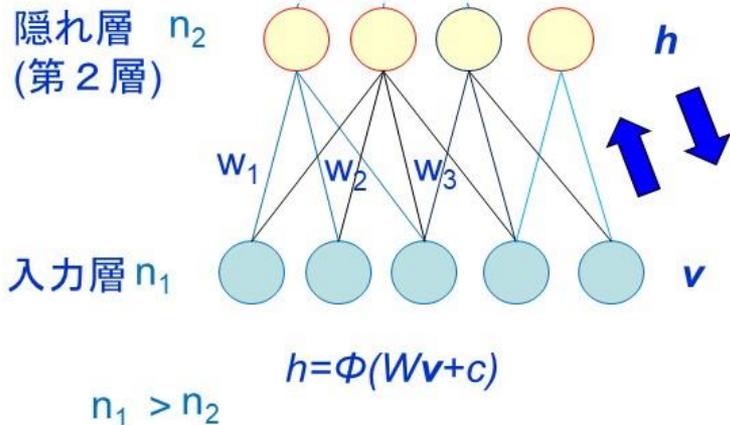
# Deep Learning 人工知能革命

それまでのニューラルネットワーク



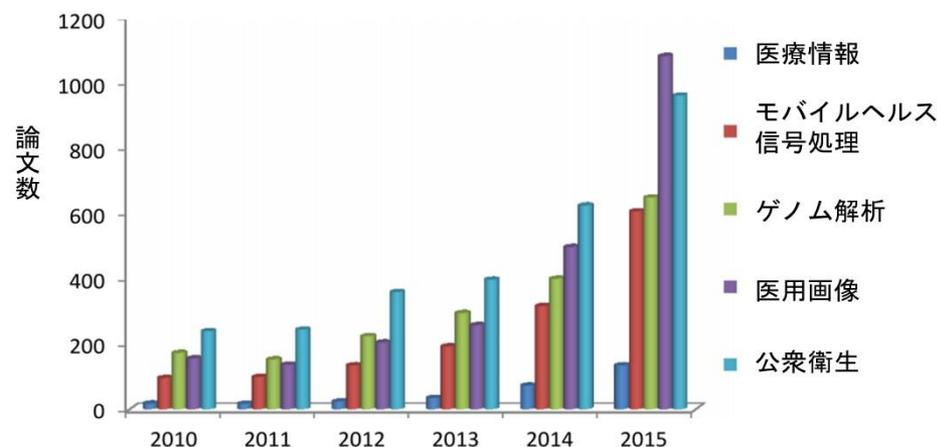
各層におけるニューロンの数

誤差を逆伝搬



# 医療への深層学習のこれまでの応用

- **ゲノム・オミックス医療**
  - 遺伝子多様性、タンパク質相互作用、がんの分類
- **医用画像処理**
  - 病態部位の自動認識、セグメンテーション、異種モダリティの統合
- **モバイルヘルス・生理信号処理**
  - 運動・カロリー推測
  - 行動支援・生体信号解析
- **医療情報**
  - 電子カルテ処理
  - 病態解釈
- **公衆衛生**
  - 流行病予測
  - 疾病への社会行動予測



深層学習を臨床意思決定支援の使うのは本来の使用ではない

# Deep learning: 創薬からの注目

- **Kaggle** (データサイエンス競技会)に**Merck社**が出題  
**Molecular Activity Challenge (2012)**.

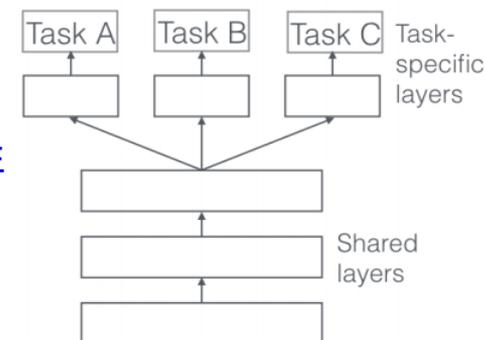
- 15種類の標的分子に対する化合物データ  
セットの有効性結果から、**構造活性相関のデータ**  
を学習して、未知の化合物の構造から分子の生物学的活性  
を予測するモデルの開発コンテスト  
勝利したモデルはdeep learning を用いた  
**1人の医薬品科学者もいなかった。**

- **Unterthiner**の大規模な構造活性相関 (QSAR)研究

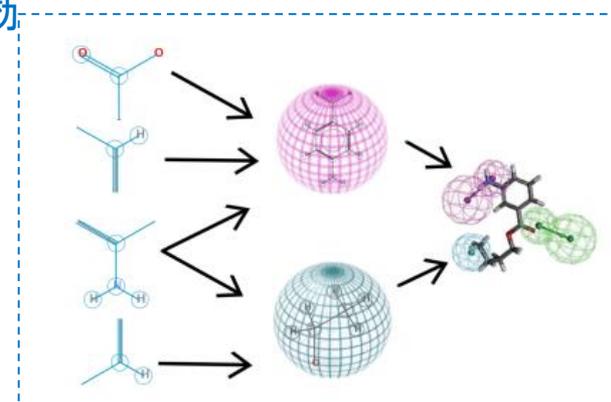
- ChEMBLに対するdeep learning
- 13 M 化合物特徴量 (ECFP12), 1.3M 化合物, 5k 薬剤標的
- Ligand-based 標的予測, 7種の予測法とAUC比較
- Deep learningがSVM, k-最近隣法, logistic回帰より有効
- 特徴量の抽出、薬理機序への理解

- **Google in collaboration with Stanford (2015)**

- Stanford 大学の Pande 研究室と共同研究  
バーチャルドラッグスクリーニングに対する  
deep learningによるツール開発  
"Massively Multitask Networks for Drug  
Discovery"

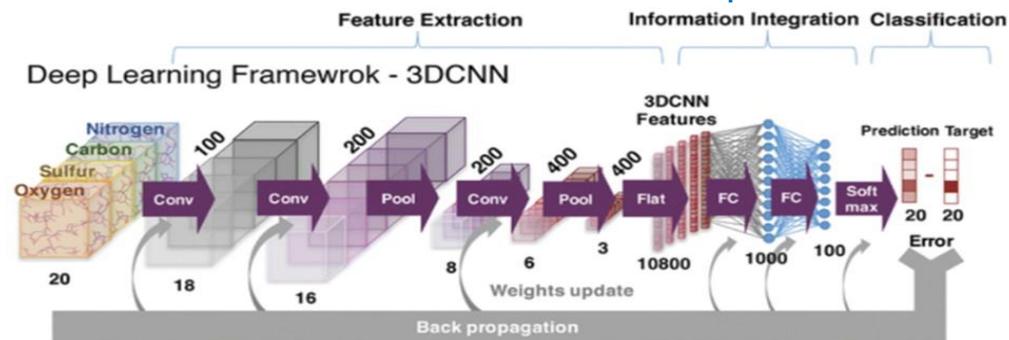


Multitask型DNN



# AI創薬の方法

- **Virtual Screeningへの人工知能・機械学習の応用**
  - **Ligand-based** AIバーチャルスクリーニング
    - 化合物のfingerprint や記述子から学習
  - **Structure-based** AIバーチャルスクリーニング AtomNet
    - 標的・化合物の3D座標⇒ Convolutional Deep Neural Network (CNN)



- **標的分子探索に人工知能を用いた方法**
  - Hase, Tsuji, TanakaのNetwork Embeddingを用いた標的探索法
- **その他**
  - 化合物の人工知能を用いた自動設計
  - 合成経路設計
  - AI毒性学

# 薬剤標的分子探索への応用

対象疾患決定後、治癒に有効な生体側の標的分子を探索  
標的分子探索の範囲を限定

⇒ ヒト・タンパク質相互作用ネットワーク (PPIN)

## 学習的アプローチ

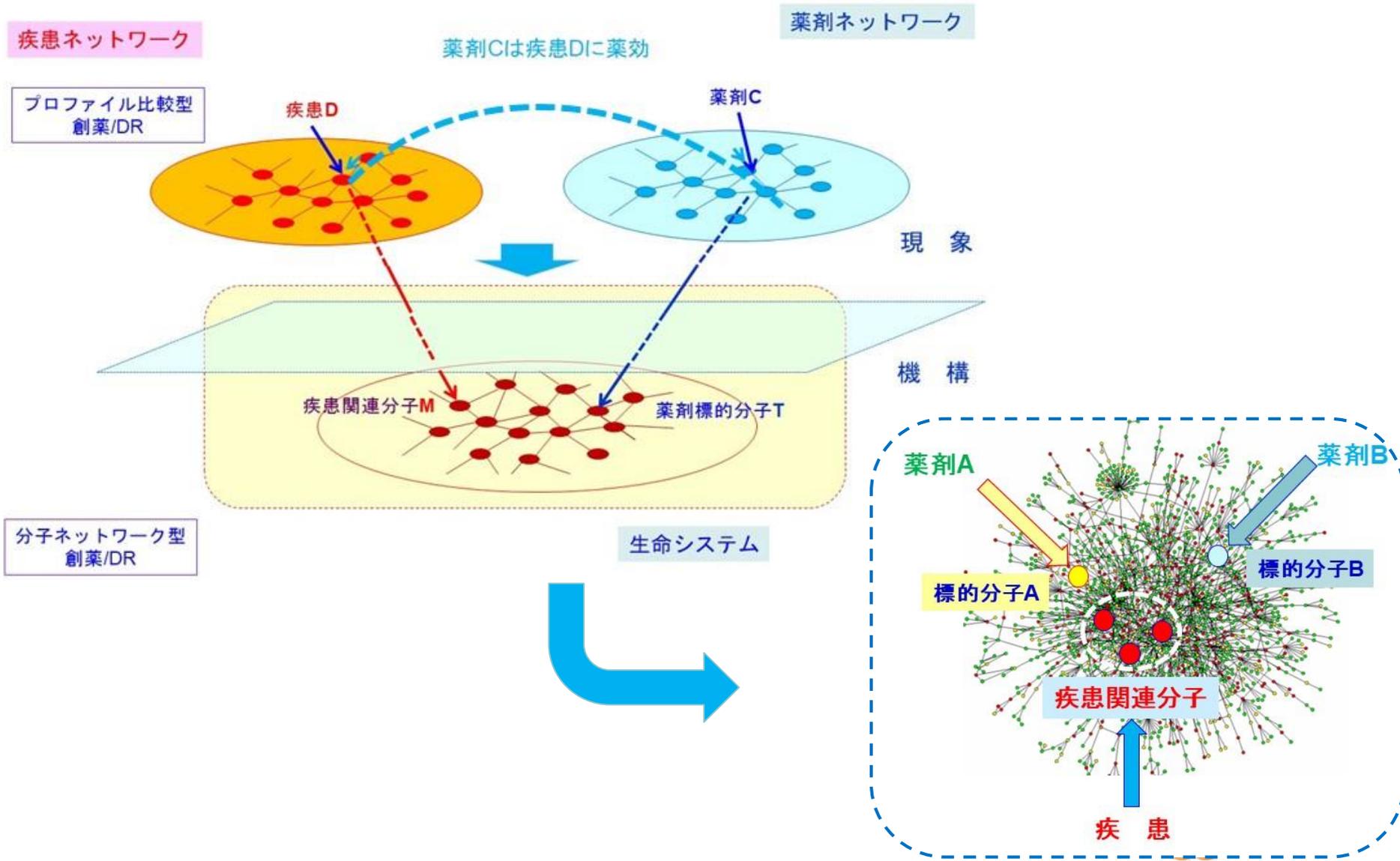
その疾患にこれまで有効な標的分子が既知  
既知の標的分子がPPIN上でどのような位置にあるのか  
帰納学習する

PPINはHPRDでは<1万タンパク質 × 1万タンパク質>の  
超多次元ネットワークで通常の機械学習では困難



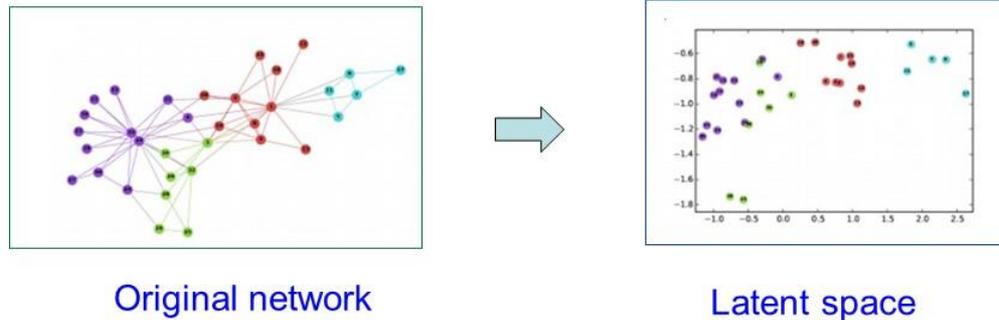
Deep learning による  
<ネットワーク埋め込み Network Embedding >

# 3層ネットワーク理論による創薬の枠組み



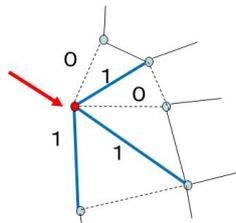
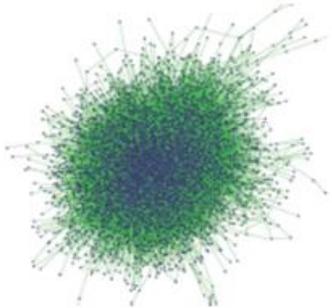
# Deep Learningによる Network Embedding

超多次元ネットワークをそれより遥かに低次元のLatent Spaceに写像



## Structural Deep Network Embedding

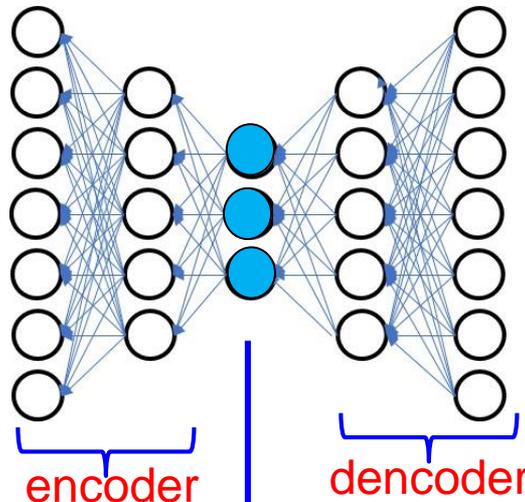
ネットワーク



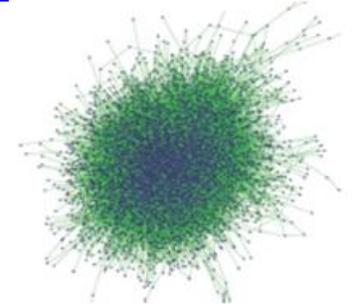
$$v_i = (0, 0, 0, 1, 0, 1, 0, \dots)$$

全節点の近接ベクトル

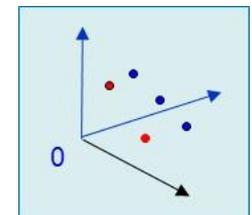
入力層



対称出力層



潜在空間  
Latent space



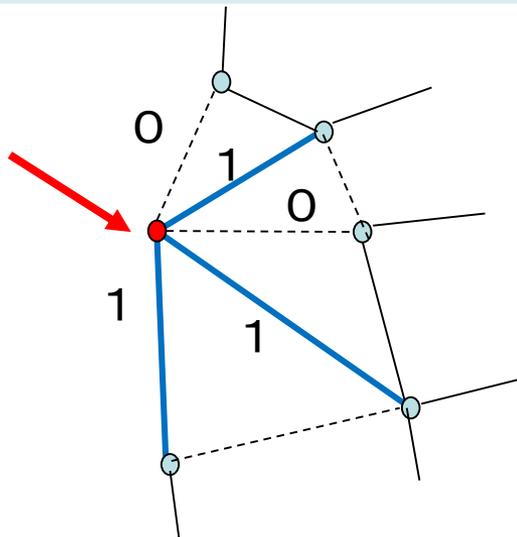
# Deep Learning と SVD (singular value decomposition)の精度の違い

あるタンパク質相互作用ネットワークのノードに注目する

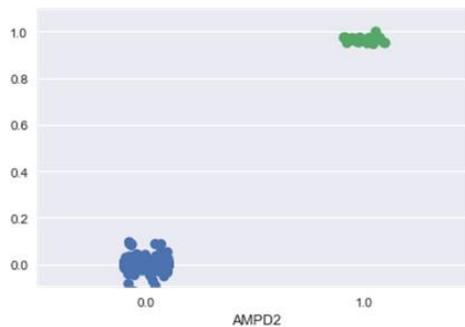
周りのノードで

結合しているノードは 1  
 結合していないノードは 0  
 とすると0, 1の近接ベクトルで結合を表現できる。

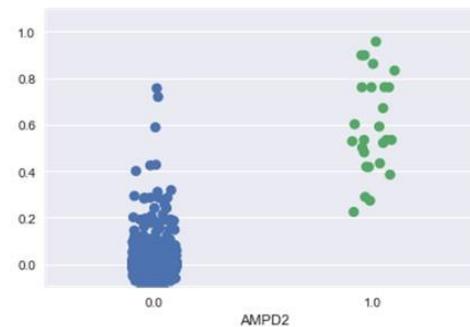
$$v_i = (0, 0, 0, 1, 0, 1, 0, \dots)$$



AMPD2 (adenosine monophosphate deaminase 2)  
degree=26



Autoencoder

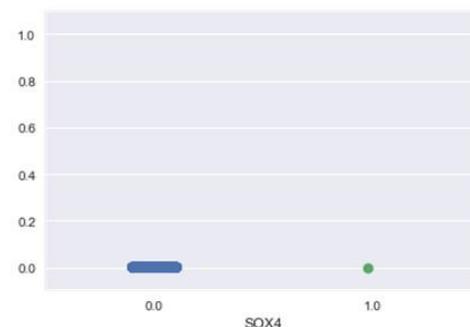


SVD

SOX4 (SRY-box 4)  
degree=1



Autoencoder



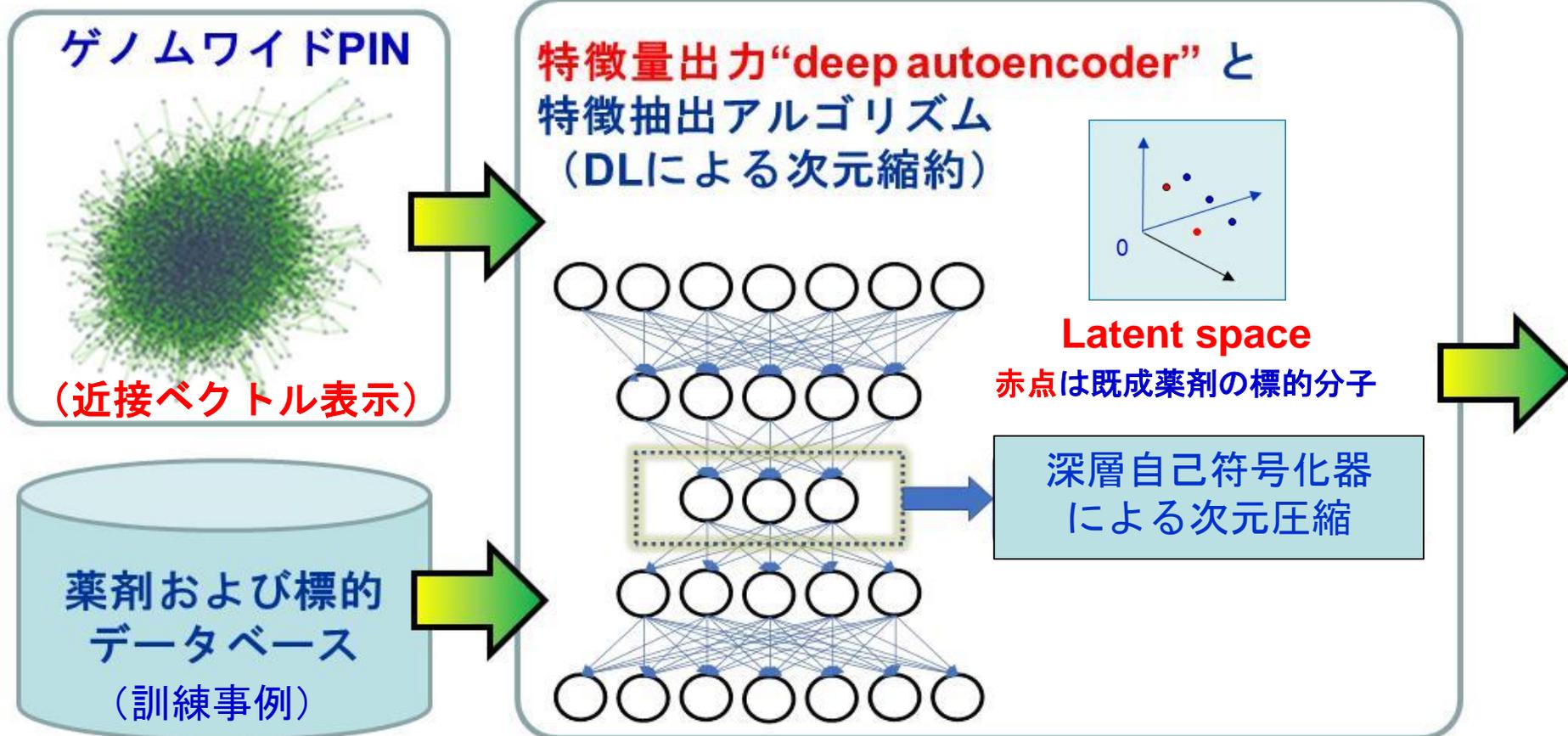
SVD

N=8,502

# Structural Deep NW Embedding(SDNE)による AI創薬 (Hase-Tanaka法) 1

入力

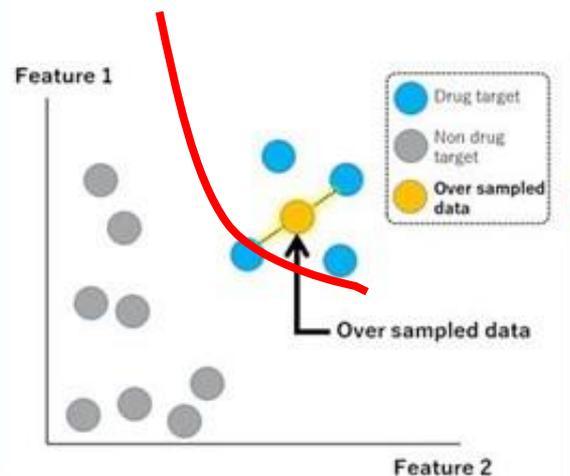
特徴量産出



# Structural Deep NW Embedding(SDNE)による AI創薬 (Hase-Tanaka法) 2

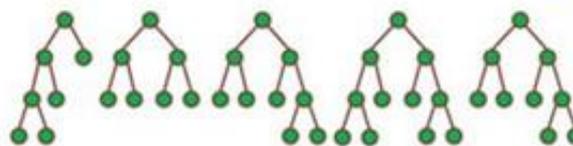
## 分類モデル

### 2群分類と標的判定 最新のアルゴリズム



Xgboost

標的性判定 algorithm to build  
a binary classifier



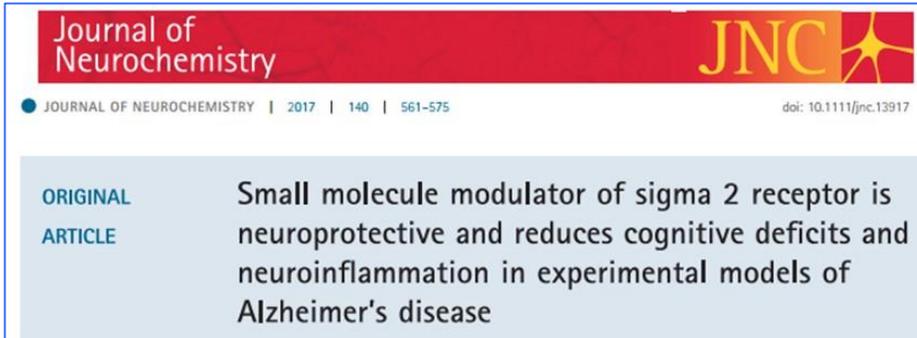
## 標的選定

### 標的性判定

遺伝子	標的確率
GRASP	0.982971
PGRMC1	0.982345
GPM6A	0.982345
NRP2	0.975194
PFKM	0.972128
DLGAP2	0.953659
CD81	0.941095
IQGAP1	0.926867
TROVE2	0.916886

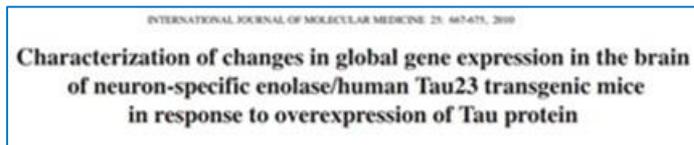
# 実験的研究との付合 1

## PGCM1 : progesterone receptor membrane 1

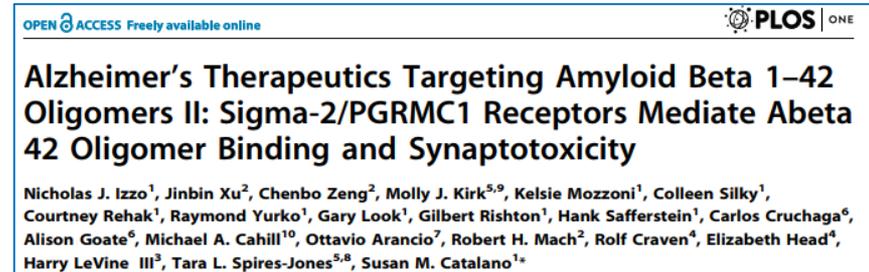
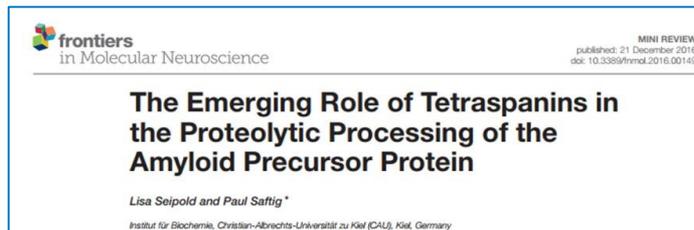


神経保護の効果 (neuroprotective) 認知不全・炎症に治療効果

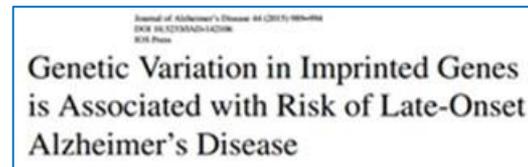
## GPM6A : Glycoprotein M6A



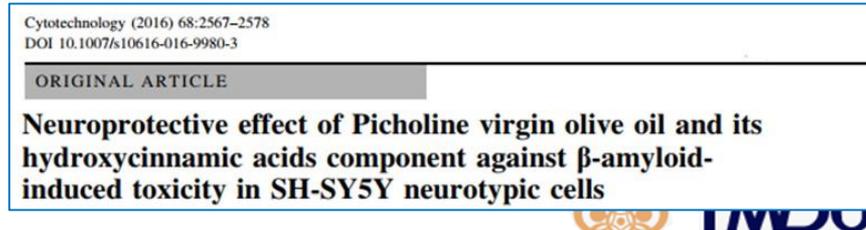
## CD81: Tetraspanins family



## DLGAP2 : DLG-Associated Protein 2

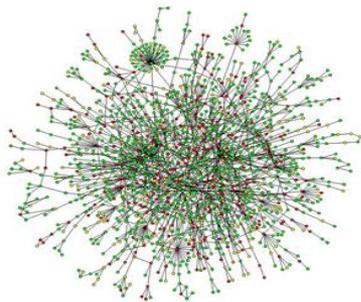


## PFKM: Phosphofruktokinase

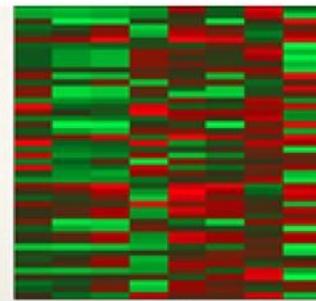


# 属性付加ネットワーク埋め込み法による PPIN・遺伝子発現の圧縮

タンパク質相互作用ネットワーク  
(PPIN)



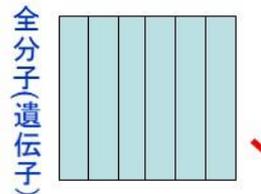
疾患罹患時  
遺伝子発現プロファイル



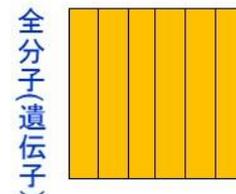
遺伝子発現  
相関ネットワーク



深層自己符号化  
による次元圧縮

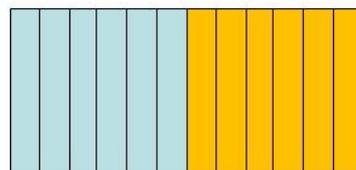


潜在変数(100次元)



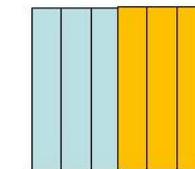
潜在変数(100次元)

融合



潜在変数(200次元)

深層自己符号化  
による次元圧縮



潜在変数(100次元)

Attributed Network  
Embedding

# 遺伝子発現プロファイル・PPINによる Network Embedding (Tsuji-Tanaka)

(a)PPIN単独での標的分子リスト

分子名	比率	アルツハイマー関連
PTGS2	0.500	多型がAlzheimer病と関連。
APH1B	0.500	$\beta$ アミロイドの前駆体生成に関与。
NCSTN	0.500	
MGA	0.333	
GUCY2F	0.250	
GUCY1B3	0.250	
GUCY1A3	0.250	
PHC3	0.200	
PIGW	0.143	

(b)PPINと遺伝子発現プロファイルでの  
標的分子リスト

分子名	比率	アルツハイマー病実験的研究との関連
HOXA1	0.0769231	発生制御遺伝子であるが、HOXAの領域の広範囲なDNAメチル化が、アルツハイマー病と関連[1]。
PDE5A	0.0666667	
NCSTN	0.0625000	$\beta$ アミロイドの前駆体生成に関与[2]
NR0B1	0.0400000	
F2R	0.0322581	新規のAlzheimer病創薬ターゲットとして、注目[3]
IGFBP3	0.0259740	アルツハイマー病の中心的な役割を演じる分子[4]
LDHAL6A	0.0243902	
TGIF1	0.0212766	Alzheimer病との関連の報告あり[5]
ITGB5	0.0185185	ネットワーク解析から、Alzheimer病との関連示唆[6]
HLA-A	0.0169492	この領域の多型とAlzheimer病との関係あり。炎症との関連で注目[7]。

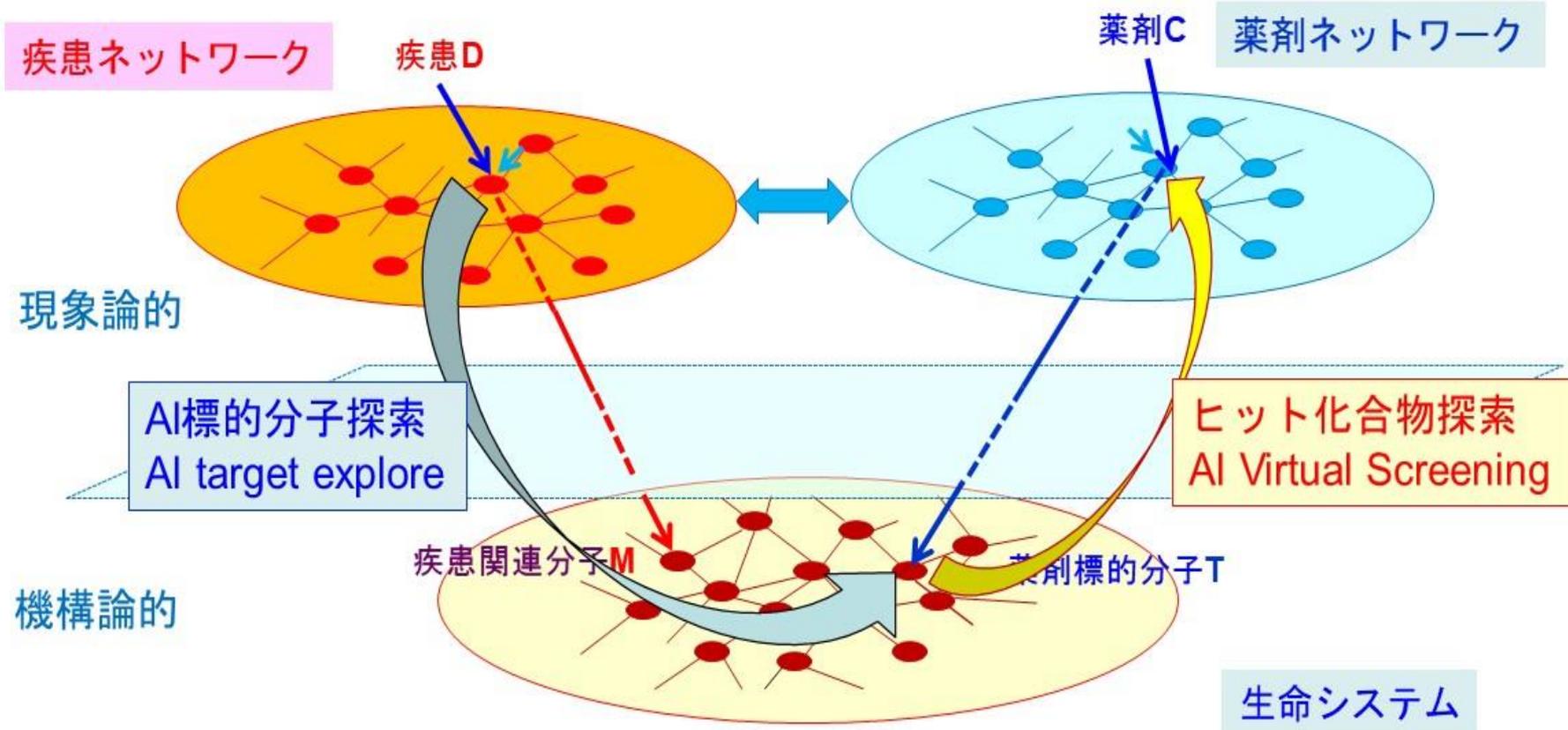
## 参照URL

- [1] <https://www.sciencedirect.com/science/article/pii/S1552526018300499>
- [2] <https://www.ncbi.nlm.nih.gov/gene/23385>
- [3] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4333706/>
- [4] <https://www.ncbi.nlm.nih.gov/pubmed/26637371>  
<https://repository.kulib.kyoto-u.ac.jp/dspace/bitstream/2433/215438/1/yigak04119.pdf>
- [5] <https://www.biorxiv.org/content/early/2018/03/22/286674>
- [6] <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0040498>
- [7] <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5696257/>  
<https://www.ncbi.nlm.nih.gov/pubmed/18936542>

# AI創薬の実現

3層生体・薬剤ネットワークによるAI創薬の過程

薬剤Cは疾患Dに薬効



# ゲノム医療の第2世代

一部の単一遺伝病を除き、大半の疾患  
(Common diseases)の発症は

疾患発症の相対リスク=

遺伝要因(G:genome) X 環境要因(E:exposome)

相互作用は加算的でもなく乗算的でもない  
＜(G,E) 組合せ特異的な効果＞である

GWASでSNPの相対リスクが低い  
(1.1~1.3)理由: GxE組合せ特異  
的效果を環境要因の全てに亘って  
平均しているからである



# GXEの記憶をとどめている 網羅的分子情報

- エピゲノム

- DOHaD仮説

- オランダ飢饉の時胎児→成人後肥満・心筋梗塞・糖尿病
    - 過度な低栄養：肝臓のPPAR $\alpha/\gamma$ （儉約遺伝子）  
メチル化低下・遺伝子発現がオン
    - Baker仮説：戦後の心筋梗塞増加：貧しい村

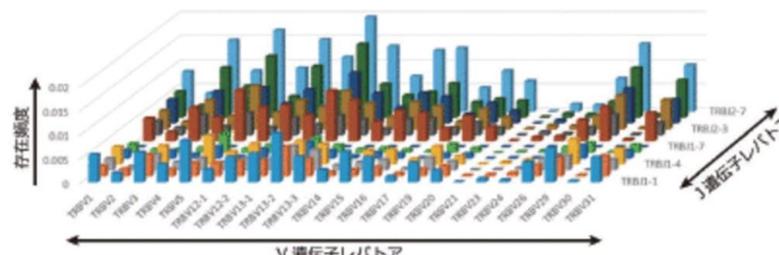
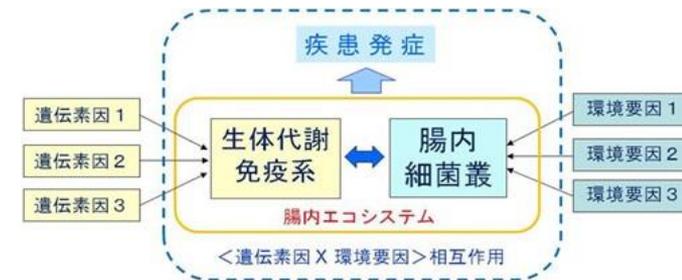


- ホロゲノム（メタゲノム＋宿主ゲノム）

- 腸内細菌層microbiomeが疾患（環境相互作用）の原因
  - 心筋梗塞：赤肉からのTMAO
  - 脳腸相関：自閉症の原因

- 免疫ゲノム（TCRレパトア）

- 環境病原体との関係：がんなど



# 第2世代のゲノム・オミックス医療

- 生涯的全体性においてその個人の疾患可能性の全体性を把握し、個別化予防、個別化治療に取り組む
- ゲノム・オミックス情報と医療・健康
  - **Clinical Sequencing**のインパクト
- **第1世代ゲノム医療**
  - ゲノムの変異・多型性の個別性に基づく
- **第2世代のゲノム医療**
  - 多因子疾患が対象、環境情報との相互作用
  - エピゲノム、メタゲノム・免疫ゲノムなど
  - 遺伝子X環境の相互作用を反映する**メタ・オミックスのバイオマーカ**が必要

**疾患メタ・オミックス修飾**

# 今後の医学〈知〉の展望

- ビッグデータ医療時代：次元縮約
- Deep Learningによる〈多次元ネットワーク情報構造〉の縮約
  - 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医療の「枠組み」実行方向は「見えてきた」

ヒトの仮説駆動的な〈知〉とAIのデータ駆動的な〈知〉との「共創的cocreativeな〈知〉」「ケンタウロスの知性」が医学のみならず人類の未来の進むべき途の探索を可能にする

ご清聴有難うございます

