クレジット：
UTokyo Online Education 学術俯瞰講義 2019　田邉 國士

質量分析と機械学習を融合したがん診断支援装置の開発：学際的研究から産学連携へ

機械学習と統計科学：
Epagogics(帰納学)の慫慂

於 東京大学・駒場

2019. 6. 18

たなべ　くにお

田邉 國士

早稲田大学/理化学研究所/統計数理研究所

私の立ち位置

所属歴

- 統計数理研究所(1967-2005）
- 早稲田大学 理工学部（2005-現在）
- 総合研究大学院大学、山梨大学、理研、
  North Carolina State Univ, Brookhaven Nat. Lab., Univ. of Baselなどを併任

専門: 数理科学

- 数値解析（数値代数とアルゴリズム）
- オペレーションズ・リサーチ（最適化理論とアルゴリズム）
- 統計科学（ベイズモデル、逆問題）
- 蓋然的推論機構（帰納的推論機械(=機械学習)）

所属学会歴

- 数学会、AMS
- 応用数理学会, SIAM
- オペレーションズ・リサーチ学会
- 統計学会、 応用統計学会, ISI
- 情報処理学会
- 計測自動制御学会, IEEE
- 科学基礎論学会

近年、囲碁の名人に打ち勝つことで広く世間にも知られるようになったAIですが、その頭脳部をなす学習機械がどのような方法論的概念に基づいているかを知る人は多くありません。今日 Deep Neural Netwok と呼ばれる学習機械がもてはやされていますが、便利な道具としてこの学習機械のソフトウェアを操る技術者も、それが対象をどのように捉えるものであるかについて方法論的な意味を自覚しているとはいえません。

本日は、DNN, dPLRM を含む学習機械がもたらす科学方法論およびエンジニアリングへのインパクトについて「がん診断」などを例にとり述べ、学習機械一般の数学的本質を、短時間のためテクニカルな内容は避けて俯瞰したいと思います。

トピックス:

- **科学推論**
- **機械学習の数学と本質**
- **統計学の推論　……「うそ、まっかなうそ、統計学」**
- **説明・解釈とはなにか？……　社会的承認**

キーワード

科学方法論、実在、存在、表象、事象、分節化、名辞、記号、言語、数学

仮説演繹法、ニュートン・デカルト・パラダイム

機械学習、ディープ・ニューラル・ネット、罰金付きロジスチック回帰、SVM

確率、母集団、仮説検定、ベイズ統計学、頻度論統計学、回帰と共起分布

解釈・説明と社会的承認

著作権等の都合により、
ここに挿入されていた画像を削除しました

新聞記事
朝日新聞　2019年5月6日
見出し：説明できるAIへ　進む研究
https://www.asahi.com/articles/
DA3S14003210.html

AIや機械学習について、しばしば求められる質問・要求

- 機序はどうなっているのですか。説明して下さい。

  eXplainable AI = XAI を唱える人々がおりますが、私は不可能と思います。
  Evidence-based Medicine の衝突は避けられないかもしれない。

- どの位の数の訓練用データがあればどの位の正確な判断ができるのかを統計学的に示してください。

  この要求にも答えることは原理的に出来ません。統計学的推論が導く帰結は、すべて特定の想定の下での条件付き命題であるからです。統計学に過大な期待をしてはいけません。

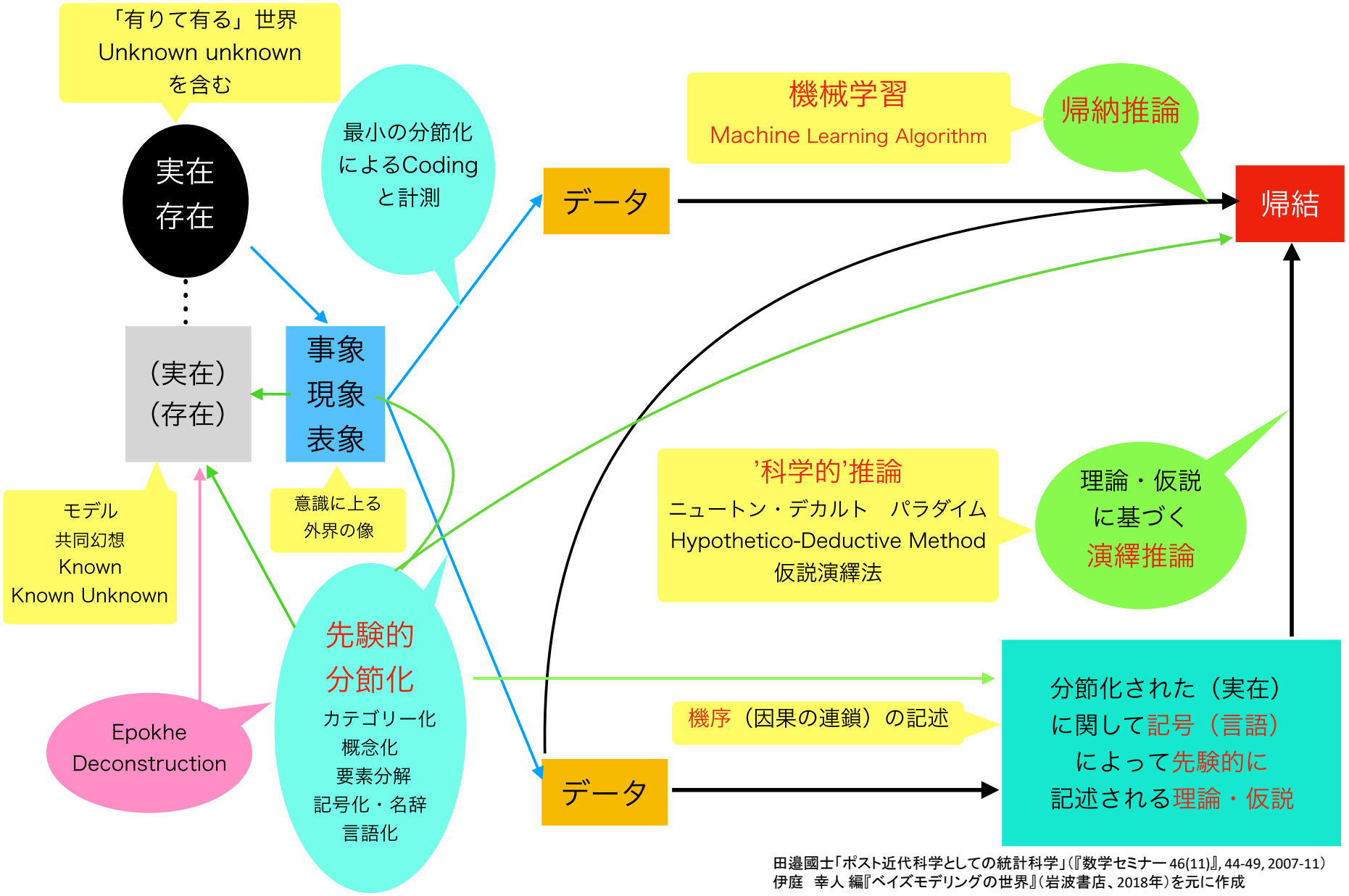
  その理由をこの講義で示したいと思います。

ニュートン-デカルト・パラダイム ＝ 仮説演繹法

従来の科学やエンジニアリングはニュートンとデカルトによる方法論に基づいて発展してきました。すなわち事象の素因となる原始要素を見つけ出し(先験的分節化）、それを精密に同定・測定し、論理的な推論によって導かれる帰結を予測し、事象の観測データと照合して、事象を理解し制御する方法論です。

実際、科学研究の現場においては原始要素の精密な観測・測定が要求されますし、エンジニアリングにおいても精密な観測を担保する機器の製作に意を用います。因果関係における関与因子の数や相互作用の単純性を秘かに仮構しているこのパラダイムは無意識の裏に現代人に刷り込まれており、科学研究は疎か社会的諸制度が要求する規制に対する適合証明（例えば薬品の有効性の証明）は、このパラダイムに則り行わねばなりません。これに基づかないものは非科学的であるとして退けられてます。

推論　Inference



田邉國士「ポスト近代科学としての統計科学」(『数学セミナー 46(11)』, 44-49, 2007-11)
伊庭　幸人 編『ベイズモデリングの世界』(岩波書店、2018年)を元に作成

「がん」とは何か？　定義は？

「がん」 は

1. 細胞単体に付与される概念か？

2. 細胞群に付与される概念か？

3. 細胞群とそれを囲むミクロな環境に付与される概念か？

医師の回答：

染色された細胞群の顕微鏡下での形態学的観察に（主に）基づいて、病理医が「がん」と診断したものが「がん」である

著作権等の都合により、
ここに挿入されていた画像を削除しました

ResearchGateの記事
2013年11月11日　ReseachGate
Is there any attempt to model
"Physics of Cancer"?

もっとも確実な演繹推論：数式モデルによる推論

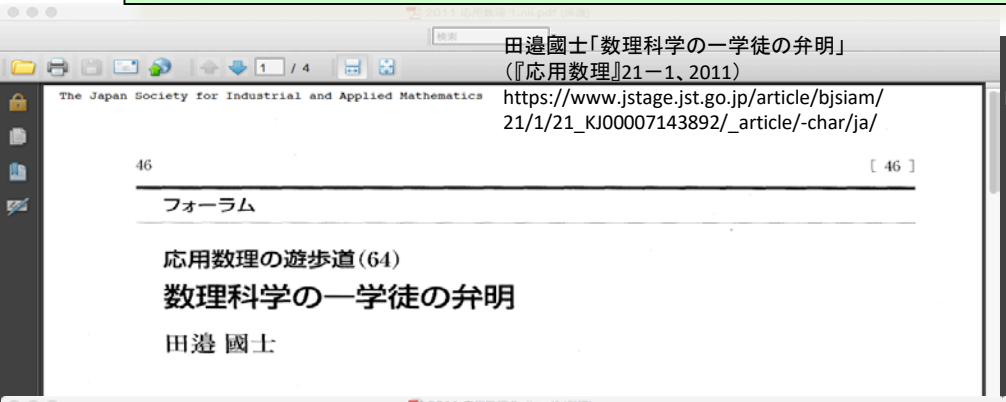
著作権等の都合により、
ここに挿入されていた画像を削除しました

演繹推論のイメージ図

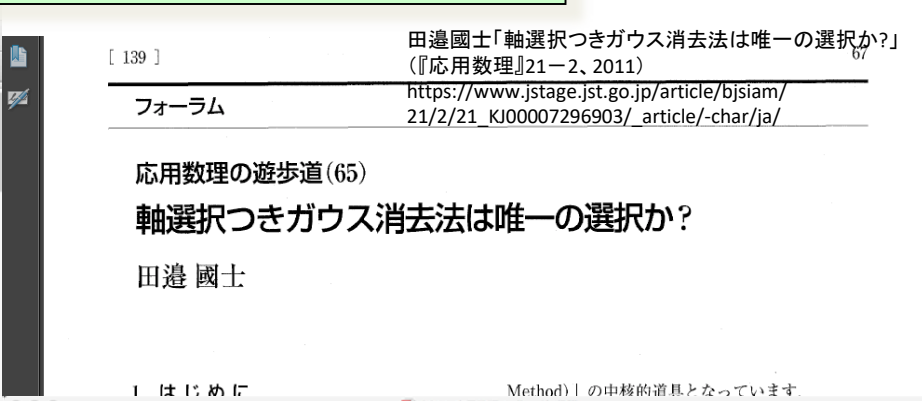
著作権等の都合により、
ここに挿入されていた画像を削除しました

書籍の表紙
Nassim Nicholas Taleb
The Black Swan : The Impact of the Highly Improbable
Penguin、2008

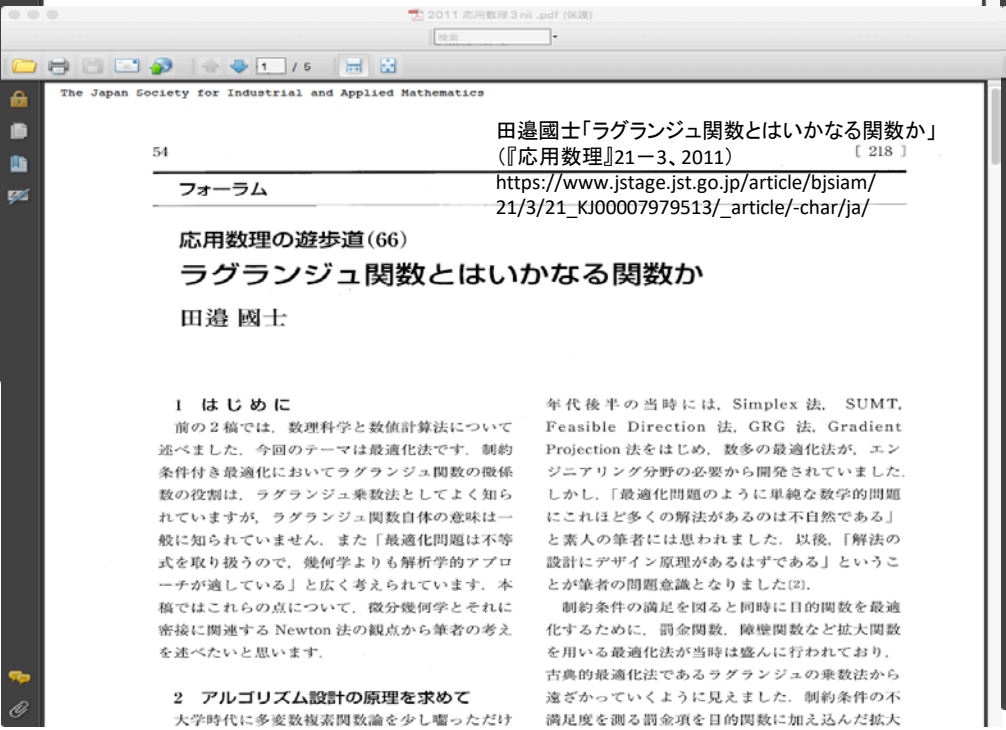
帰納という原罪



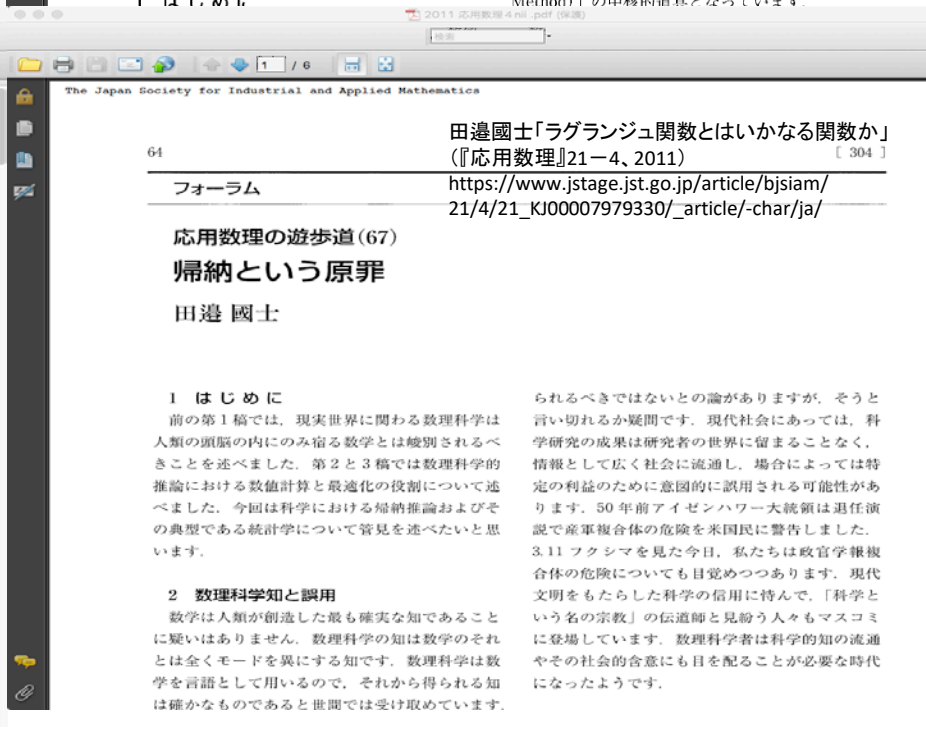
田邉國士「数理科学の一学徒の弁明」
(『応用数理』21－1、2011)
https://www.jstage.jst.go.jp/article/bjsiam/21/1/21_KJ00007143892/_article/-char/ja/



田邉國士「軸選択つきガウス消去法は唯一の選択か?」
(『応用数理』21－2、2011)
https://www.jstage.jst.go.jp/article/bjsiam/21/2/21_KJ00007296903/_article/-char/ja/



田邉國士「ラグランジュ関数とはいかなる関数か」
(『応用数理』21－3、2011)
https://www.jstage.jst.go.jp/article/bjsiam/21/3/21_KJ00007979513/_article/-char/ja/



田邉國士「ラグランジュ関数とはいかなる関数か」
(『応用数理』21－4、2011)
https://www.jstage.jst.go.jp/article/bjsiam/21/4/21_KJ00007979330/_article/-char/ja/

応用数理 Vol.21, No.1~4, 2011

Hypothetico-Deductive Paradigmの限界

- 確実な知識は仮説演繹法によってのみ獲得されるべきであるという信念は、現代の科学者に広く受け入れられており、社会学や心理学分野にさえこれを奉じる者が多い。

- 仮説演繹法はもともと天体力学などの物理の領域から発展してきたもので、比較的単純な因果関係が措定できる事象を対象としたものである．しかし、演繹主導のこの方法には、適用上の限界がある。対象が「物」であった近代科学とは異なり、現代社会が取り扱うべき対象は、異なる時間スケールでダイナミックに変化する多様な要素が複雑かつ階層的にフィードバック結合した自由度を持つシステムであり、実体論的な仮説を容易に想定できるようなものは少ない。

- 脳の認知活動の研究やゲノムデータの解析の場合のように、多数の要素が輻輳して多様な関係で結ばれているような事象の解析においては、仮説を構成するために先験的に分節化されるべき原始要素（力、質量、加速度などのような）を見出すことは容易ではない．

- たとえ関係する要素群を見出したとしても可能な仮説の数は組み合わせ論的に爆発的に増大し、各仮説の妥当性を検証することは計算論的に現実性がないばかりでなく、照合すべきデータ量が不足し、有意な実験的検証を行うことが困難であることが多い。

機械学習による人工知能の登場

因果の関係が錯綜する複雑な事象に関わるデータの解析および複雑事象の制御に対する現代社会の要求は、前世紀末において科学的推論法の新しいパラダイムを生み出した。ニュートン以来の科学的推論の規準的方法である「仮説演繹法」、すなわち「仮説定立−演繹−実験」という手続きとは異なり、帰納的推論を機械的に実現する新しい科学的推論法として「機械学習」という概念が数理科学・情報科学の発達によって登場してきたのである。

人工知能 ⊃ 学習機械

人工知能(AI)の頭脳部は学習機械と呼ばれる計算アルゴリズムから成り立っています。学習機械に訓練用データセットを入力すると、人間の介入なしにデータに伏在する規則や構造を陰伏的に創発することができるので、因果関係の連鎖が複雑なためその関数関係をあらかじめ特定できないような現象の予測が可能となります。

近年、ニューラルネット、サポートベクターマシンを始め数多くの学習機械が開発され、その普及によって人工知能が実現し、多くの分野で実用化されています。dPLRM(dualPenalized Logistic Regression Machine, 2001)は、ニューラルネット、サポートベクターマシンなどと同じ学習機械の一つです。確率モデルに基づいて設計された統計的学習機械である点にその特徴があります。

Kunio Tanabe, Penalized logistic regression machines: New methods for statistical prediction 2, Proceedings of 2001 Workshopon Information-Based Induction Science, 71-76, 2001

ニュートン-デカルト・パラダイムからの離脱としての機械学習

機械学習の概念はニュートン-デカルト・ パラダイムを覆すものです。因果関係の連鎖の同定なしに、科学的推論を可能とするものが学習機械です。変数の間に措定されるべき機序の先験的な発見という従来人間が行うものとされてきた帰納の営為を要しません。帰納の過程を機械化することによって科学研究過程とエンジニアリングの方法論にもう一つパラダイムをもたらしているのです。

そのひとつの顕著な特徴として、学習機械における推論には精密な測定データは必ずしも必要ありません。個々の対象に付随するモードの異なる数多くの変数の観測データの組み合わせの情報から学習機械は帰納的推論を実行することが出来ます。その場合個々の変数の観測データの高い精密性は必ずしも必要がなく、多数の変数の組み合わせがもたらす情報の方が機械学習による推論過程に大きな比重を占めることがあります。このため機械学習は、単純な因果関係の措定が困難な生体現象のように、異なる時間スケールで相互作用しながら変化する多種な要素が結合した多自由度を持つ対象を探る上では格好の方法論です。

学習機械の応用の拡がり

画像認識
自動運転
ロボティクス
音声認識
話者認識
株価予測
高速株取引
顧客管理
良否識別
故障識別
……

医学データ解析
- ・血液塗抹検査画像からの、寄生生物の分類

医療診断支援
- ・ PESI/dPLRM
- ・「ホワイトジャック」

介入支援
- ・東大医科研・IBM：治療薬候補提示
- ・手術支援

介入予後予測　Intervention Response
健康管理
……

さまざまな学習機械がある

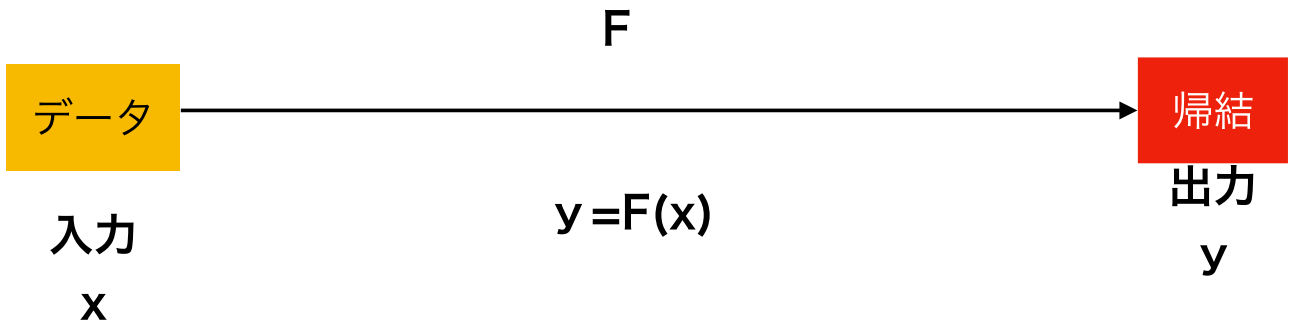
教師なし学習
- K-平均クラスタリング
- 階層クラスタリング
- ニューラルネットワーク
- 混合ガウス分布
- 自己組織化マップ
- 線形判別・2次判別

教師あり学習

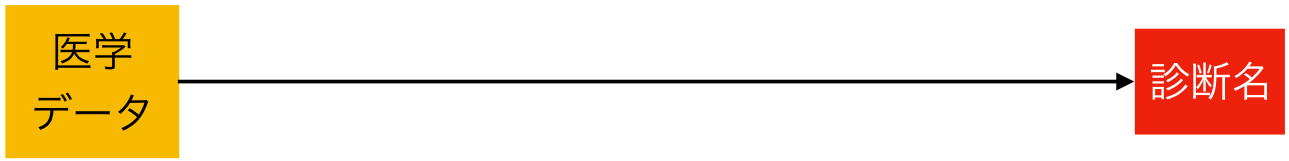
分類型
- K-最近傍識別
- 単純ベイズ分類
- 決定木、Random Forest
- アンサンブル学習
- ニューラルネットワーク（DNN)
- サポートベクターマシン、dPLRM

回帰型
- 時系列モデル（AR, ARMA model)
- 一般化線形回帰モデル
- 非線形回帰モデル
- 回帰木
- アンサンブル学習
- ニューラルネットワーク
- サポートベクター回帰
- ガウス過程回帰
- ベイジアン・ネット

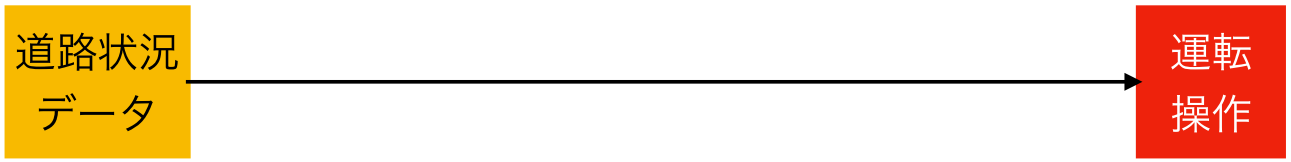
推論機構を関数と考える



診断行為を関数と考える



車の自動運転を関数と考える



関数　F を法則・理論・仮説を介さないで表現する

関数Fの柔らかいモデル（‘グニャグニャ’の関数モデル）を用意し、訓練用データ・セット $\{x_i, y_i\}_{i=1,2,\cdots N}$ を用いてそのモデルを特徴づけるパラメターの値を決める事により、関数 F を特定する。

ニューラルネット

カーネル・リグレッション関数

ロジスティック・リグレッション関数

ディープ・ニューラルネット

・・・・・・・・・・

天動説における周転円モデルのようなもの　--- 退行と見るか？

機械学習の概念的説明

機械学習においては、予測や診断などのターゲットとなるべき**Variable** $y = (y_1, y_2, \ldots, y_m)$ と観測可能な **Variable** $x = (x_1, x_2, \ldots, x_n)$ との相互関係をモデル化する。

単純化を厭わず述べると、y と x には多入力・多出力の関数関係

$$y \sim F(x) = (f_1(x), f_2(x), \ldots, f_m(x))$$

があると捉える。

従来の仮説演繹法パラダイムにおいては、**Variable x** から基本要素 x_e を分節化し、x_e と y の間に措定した因果関係などの機序に関わる仮説に基づいて F を規定し、(y, x) の観測データ・セットに基づいて F に含まれる自由パラメター群の値を推定することにより予測・診断の推論式 F を決定する。

機械学習のパラダイムにおいては、因果関係や機序などの‘真のモデル’の形式は全く分からないという前提の下に、極めて柔軟で可塑的な内部モデルを用意した上で、観測データを最も良く説明・予測する F に含まれる自由パラメター群の値を学習することにより予測・診断の推論式 F を決定する。

極めて柔軟で可塑的な内部モデルFとは

学習機械における F のモデルの要件と特徴:

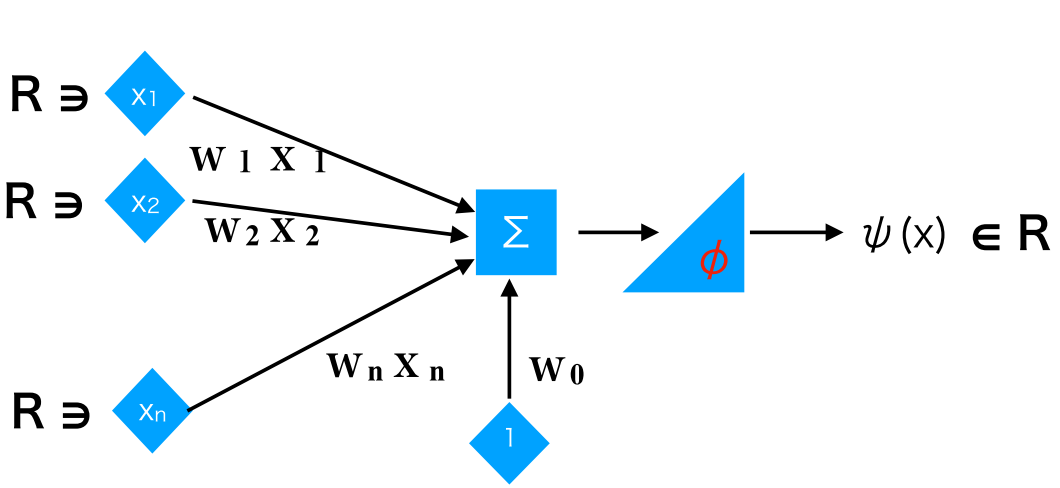
- **Variables** $y_1, y_2, \ldots, y_m, x_1, x_2, \ldots, x_n$ の間の複雑な絡まり合いの関係を豊かに表現できること
- Mode の異なる**Variables** $y_1, y_2, \ldots, y_m, x_1, x_2, \ldots, x_n$ の混合を許容すること
- 非線形関係を効果的に表現し、パラメター群を効果的に計算可能であること
- **Variable** $x = (x_1, x_2, \ldots, x_n)$ の先験的に分節化することを必ずしも要しない
- 先験的に特徴抽出を必ずしも要しない
- 機序・仕組みの知識を要しない
- ブラック・ボックス(機序は不明)である ☚ 留意 !!

Neural Network における内部モデルFの module

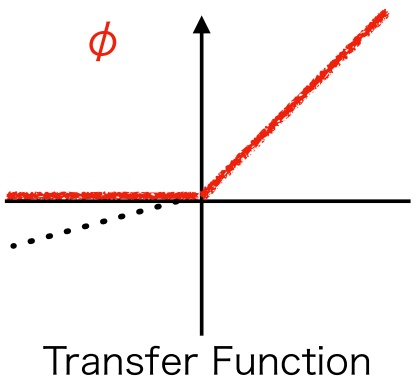
m(n+1) 個の未知パラメターを持つ区分的線形写像（函数）

$$g(x) \equiv (\psi_1(x), \psi_2(x), \cdots, \psi_m(x))^t \quad : R^n \rightarrow R^m$$

$$\psi(x) \equiv \phi\left(\sum_{i=1}^{n} w_i x_i + w_0\right) \quad : R^n \rightarrow R$$



ReLU function ϕ
(Rectified Linear Unit)



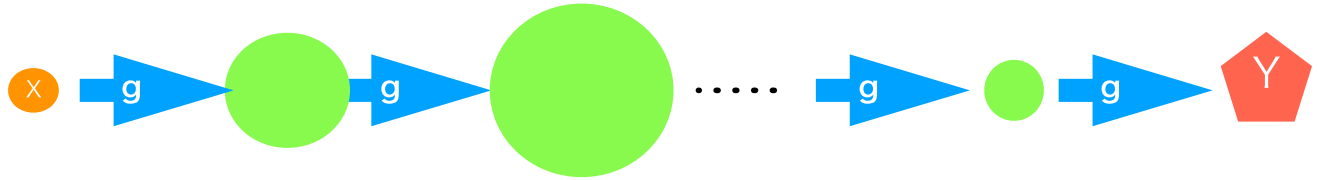
多層 Neural Network モデル (DNN)

d 層Neural Network: 合成関数

$$F(x) = g_d \circ g_{d-1} \circ g_{d-2} \circ g_{d-3} \circ \cdots \circ g_1(x)$$

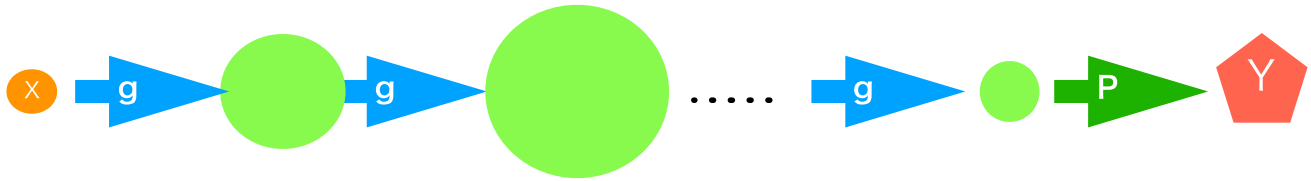
$$= g_d(g_{d-1}(g_{d-2}(g_{d-3}(\cdots g_1(x)\cdots))))$$

⇦ ReLUを採用したときは連続な区分的1次関数を表現

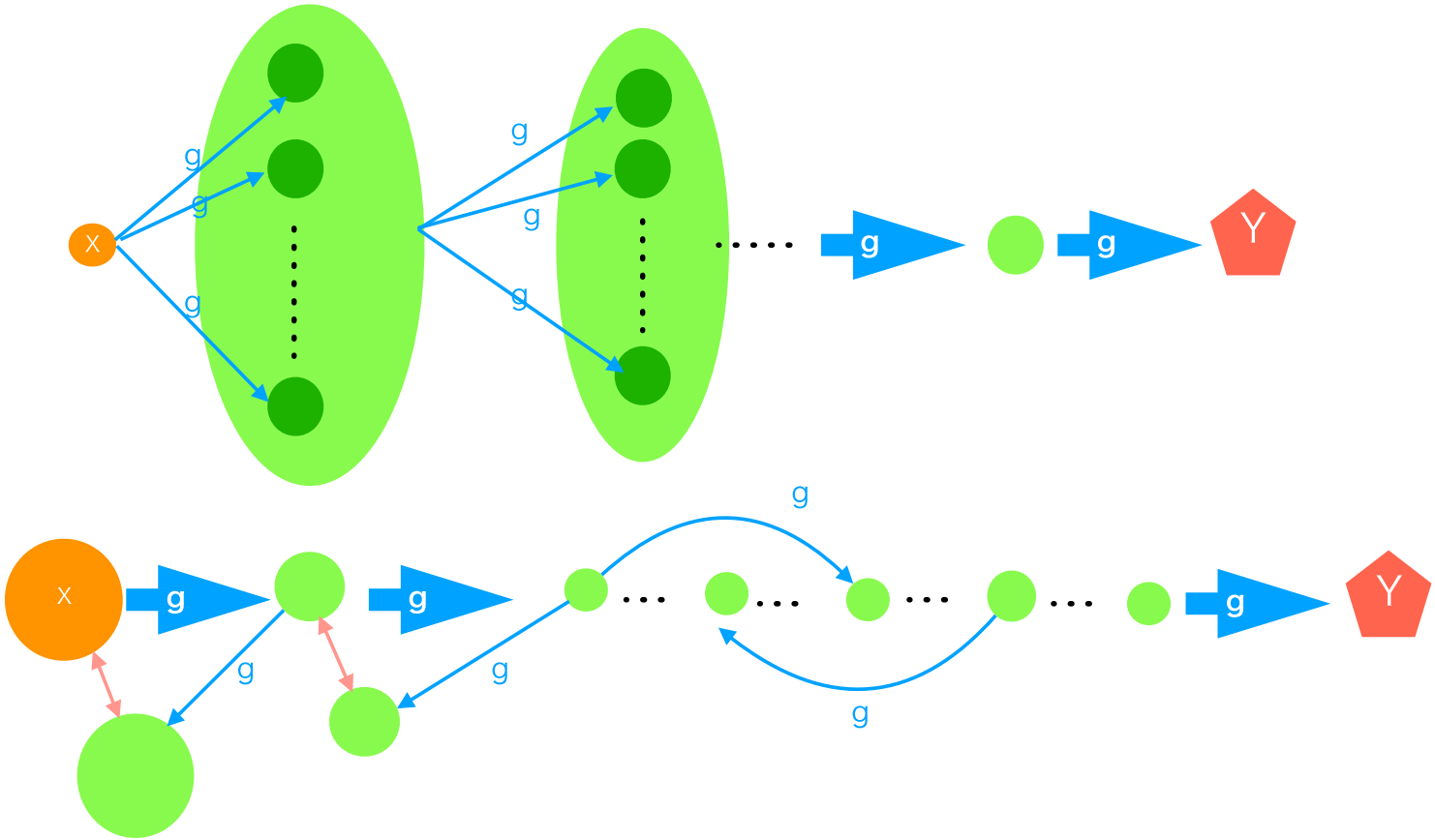


出力がカテゴリーの場合　　Multivariate Logistic Transform P(x)

$$P(x) \equiv (\sum_{i=1}^{n} \exp(x_i))^{-1}(\exp(x_i), \exp(x_i), \ldots, \exp(x_i))^t$$

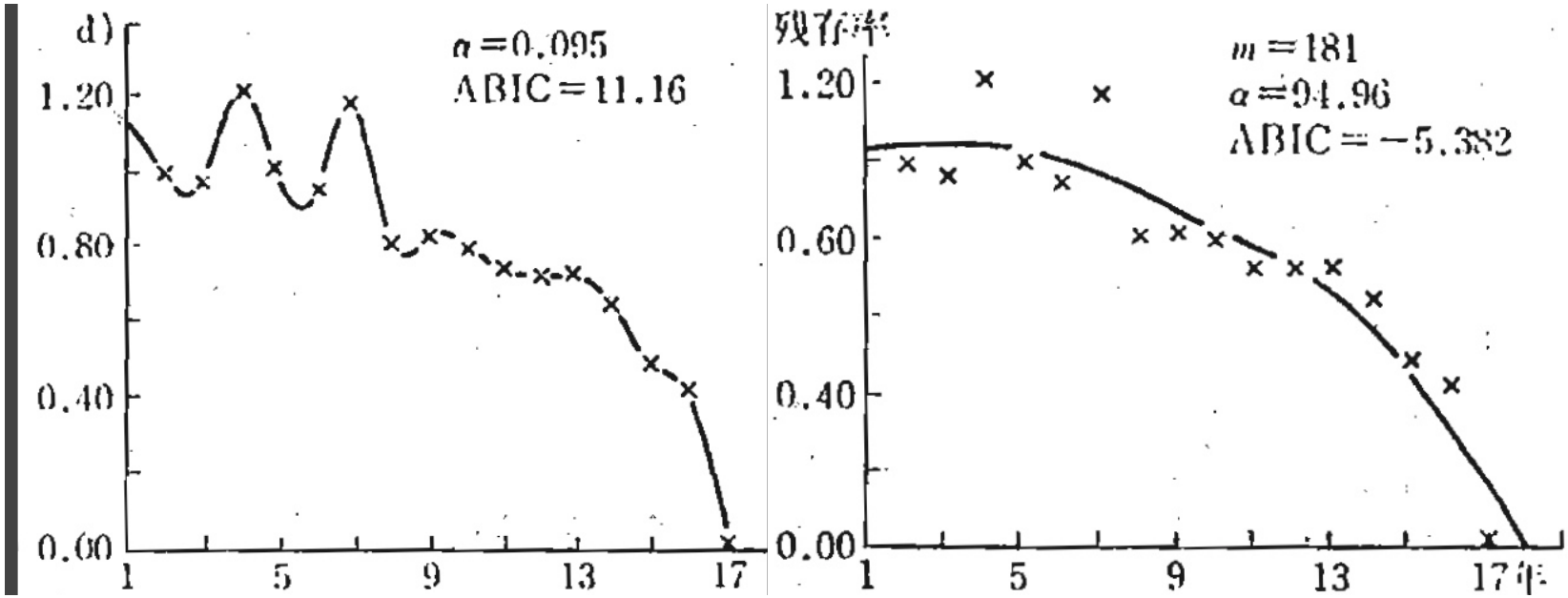


多層 Neural Network の変種



過学習: 可塑的なモデルでは Overfitting が起きる。

DNNではこれを防ぐために、訓練計算のアルゴリズムを途中で中止したり、その他さまざまな Tuning のため、人間が介入する必要がある。あるいはコンピュータの腕力を使って、やみくもに探索する。



圖：田邉國士、ベイズモデルとABIC, オペレーションズ・リサーチ, 1985.3

田邉國士「ベイズモデルとABIC」（オペレーションズ リサーチ 30, 178-183, 1985）

NLPにおいてDNNの一回の訓練に使われた消費電力

NLP: Natural Language Preocessing　言語処理分野

DNNの訓練は大企業に敵わない。アカデミズムは太刀打ち出来ない？

著作権等の都合により、
ここに挿入されていた画像を削除しました

Training a single AI model can emit as much carbon as five cars in their lifetimes
- MIT Technology Review 2019
The estimated costs of training a model
https://www.technologyreview.com/s/613630/training-a-single-ai-model-can-emit-as-much-carbon-as-five-cars-in-their-lifetimes/

Training a single AI model can emit as much carbon as five cars in their lifetimes
– MIT Technology Review 2019

SVMやdPLRMにおけるモデル

$$F_*(x) = w_1 h_1(x) + w_2 h_2(x) + \ldots + w_L h_L(x), \quad L \text{ は超巨大数}$$

(y, x) の観測データ・セットに基づいて F に含まれる自由パラメター群 $w_1, w_2, \ldots, w_L$ の値を推定することにより予測・診断の推論式 F を決定する.
☞ パラメターの数が多すぎて計算困難。

上記の推定問題の双対問題を考察すると、モデル

$$F_*(x) = \theta_1 K(x^1, x) + \theta_2 K(x^2, x) + \ldots + \theta_N K(x^N, x), \quad N \text{ はデータ数}$$

の F に含まれる自由パラメター群 $\theta_1, \theta_2, \ldots, \theta_L$ の値を推定することにより予測・診断の推論式 F を決定することと同値になる.
（カーネルトリック）

ただし $K(x^s, x^t) = h_1(x^s) h_1(x^t) + h_2(x^s) h_2(x^t) + \ldots + h_L(x^s) h_L(x^t)$

学習機械のモデルの統計モデル化

機械学習は、情報科学および認知科学の研究者によって、SVM, PAC学習などの計算論的学習理論として発展させられてきたが、近年では計算機科学や統計科学の研究者が参入し、HMM, CART, ベイジアンネット、AdaBoost, dPLRMなど様々な学習機械が登場し、データマイニング、ロボティクスなどの分野に幅広く適用されている。

これらの多くは F を確率統計モデルで表現する方向に進んで来た。

データの次元縮約は必要ない

高次元データの解析のために、前処理として次元縮約がよく行われますが、縮約によって情報が失われる事があります。

データ解析においてよく主成分分析などの多変量解析手法が用いられますが、変数間の関係が非線形でかつ正規分布などでは表現しきれない複雑な共起確率分布を持つ対象にはあまり役立ちません。

機械学習：共起関係を捉える

超多次元データの各要素変量の共起関係を捉える。

データの精密な測定は必ずしも必要がない。

Multimodal Data（異質混合データ）を取り扱うことが出来る。
　　カテゴリカルデータが混ざっても良い

アプリオリなデータの分節化は必ずしも必要ない

データの特徴量の抽出も必ずしも必要ない

アプリオリなデータの次元縮約も必ずしも必要ない

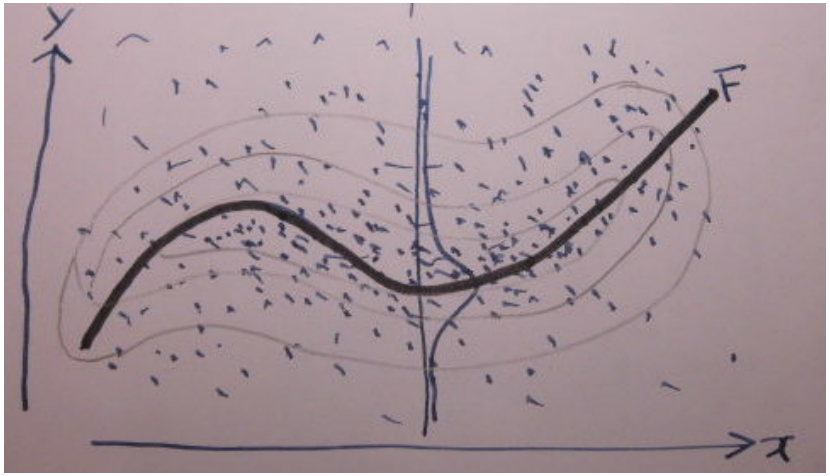
共起分布を捉える

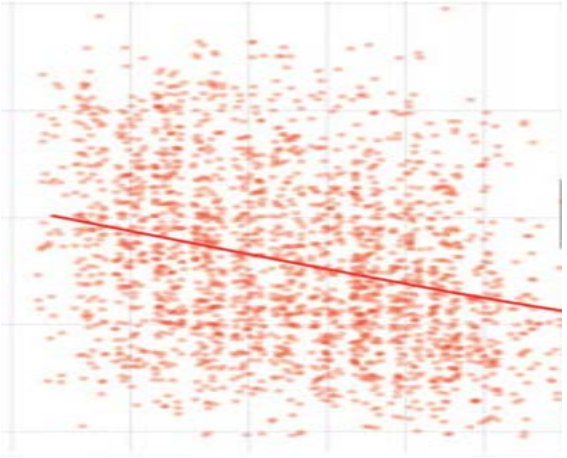
高次元データ {x, y} の各要素がとる値の共起確率分布 Simultaneous Distribution を捉える(近似する)事が本質

注)同時分布と訳されているが時間概念ははいらない。(誤訳！)

共起確率分布の近似表現法には様々な方法がある。
その一つ

- 回帰関数(モデル) F + 確率変数





機械学習のエッセンス：共起確率分布を捉えて推論を行う。

標語的に表現するならば、機械学習は

因果関係

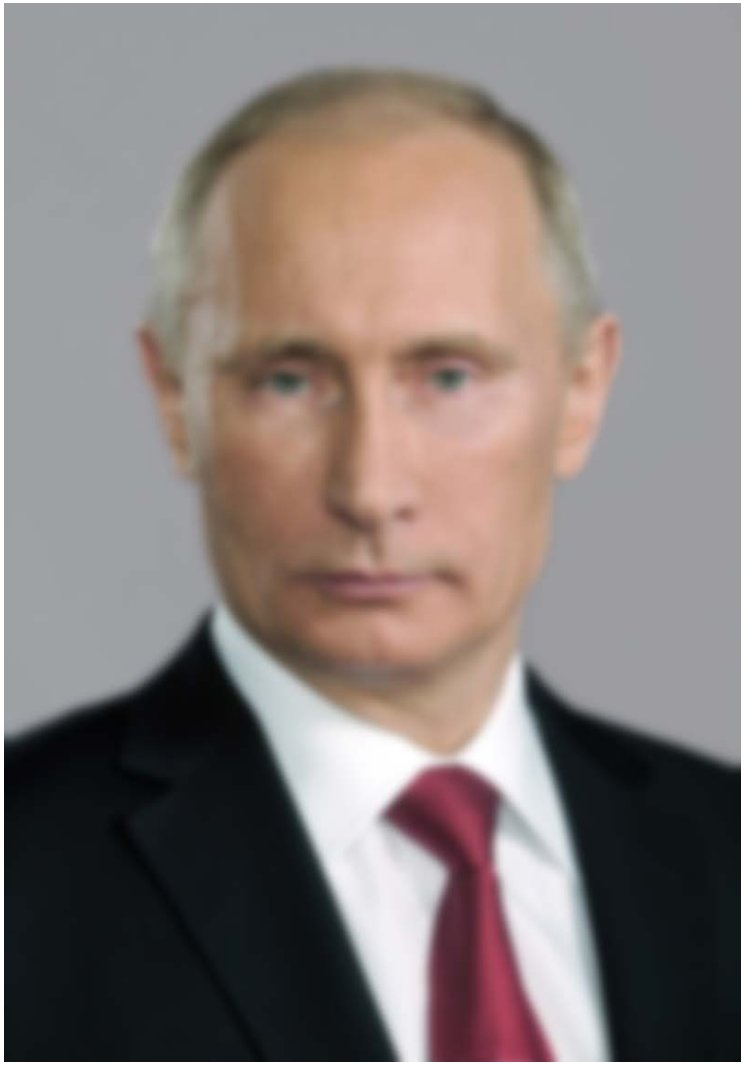
の観察ではなく

共起確率（分布）

の観察に基づく推論法である。

eXplainable AI = XAI が不可能と思う理由の一つです。

みなさん、これは誰でしょう

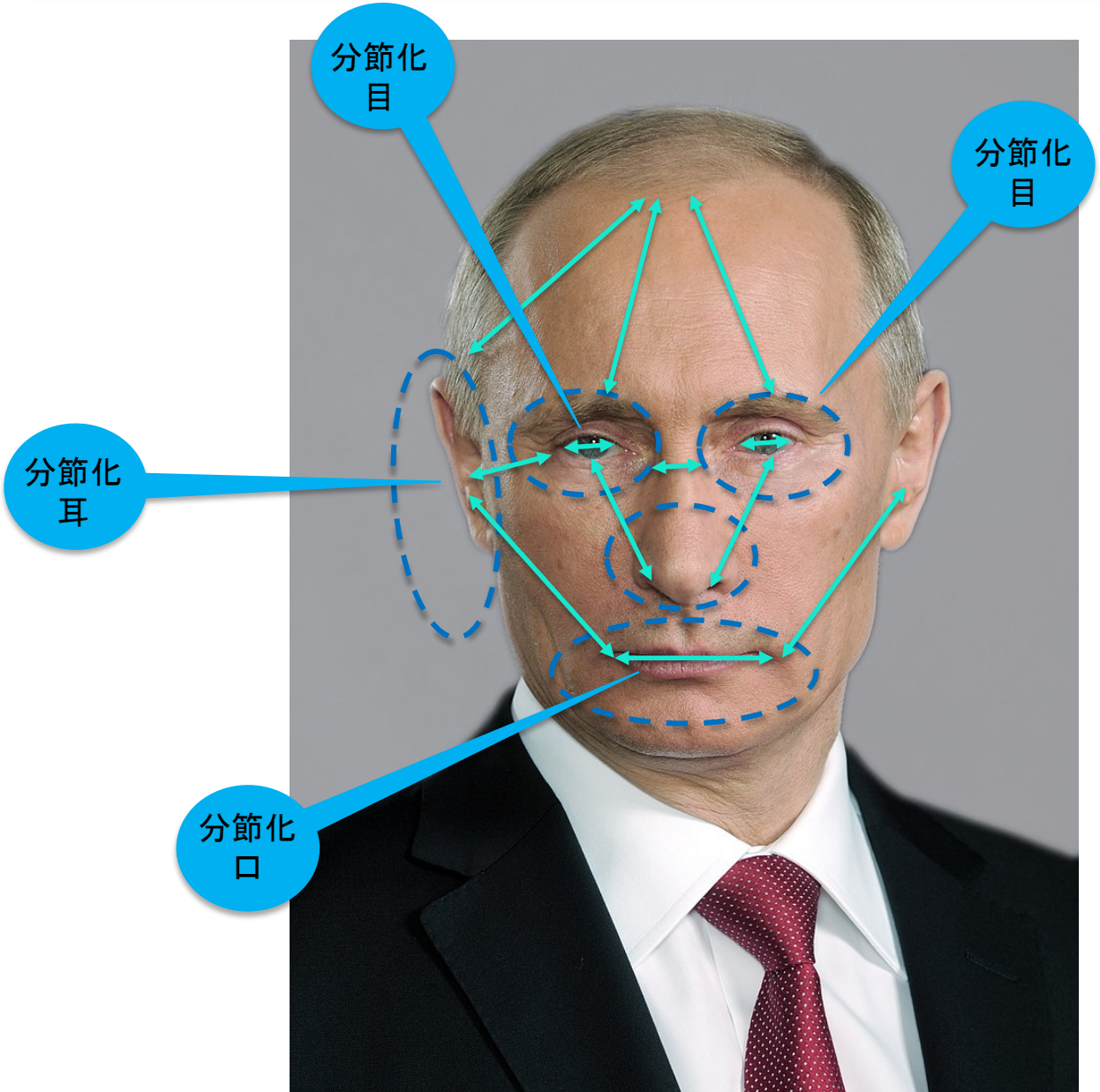


www.kremlin.ru
CC BY 3.0

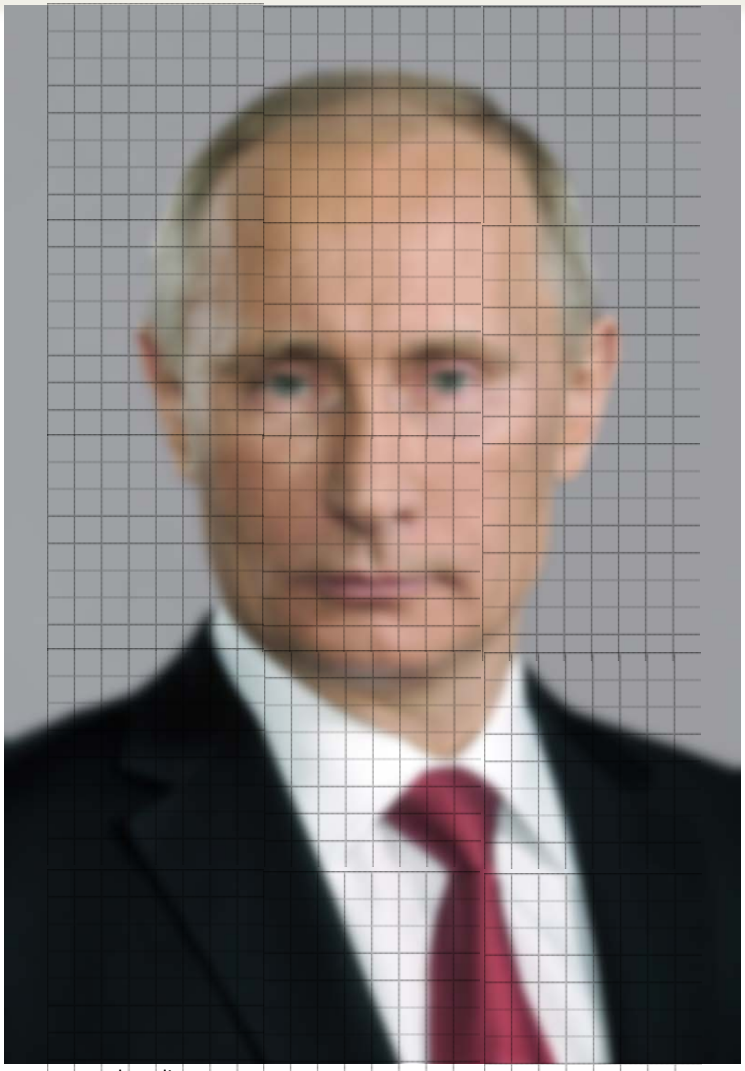
誰と識るのに、このような鮮明な写真が必要ですか？

www.kremlin.ru
CC BY 3.0

分節化も精密な測定も必要ではない。



各ピクセル毎の色の高次元データだけから推論



www.kremlin.ru
CC BY 3.0

Simple Repeat
https://simple-repeat.com

学習機械ではなく創発機械

私は「学習機械」の代わりに「帰納推論機械」という語を用いてきました．前者はLearning Machine の訳語であるが，訳語の語感には，あらかじめ‘正解’ があり，その正解を習う機械であるというニュアンスがある．NN，SVM，PLRM，dPLRMのいずれの数学的構造も，データに基づいて蓋然的な予測推論機序を˙内的に虚構˙する仕組みになっており正解があらかじめ想定されているわけではない．むしろ相矛盾するデータを含むデータセットからさえも，意味ある折り合いをつけて˙不変な何か˙を‘創発’する機械である．この点を強調するために，あえて「帰納推論機械」という語を用いてきた。

田邉國士、帰納推論機械 PLRMとdPLRM - 方法論、モデル、アルゴリズムおよび応用、システム/制御/情報、Vol.51, No2, pp.87-95, 2007

モードの異なるデータ(Multimodal Data)

患者データから肝がんの存在を予測するAIを開発

東大病院ら、正診率87.3%、AUROC値0.940と高精度

QLifePro 医療ニュース　2019年6月5日 (水)配信　消化器疾患　癌

東京大学医学部附属病院は5月30日、ディープラーニングを含むさまざまな手法から、収集された患者データから得られる予測能を最大化する学習アルゴリズムと学習パラメーターを自動抽出するフレームワークを作成し、患者データを用いた肝がんの有無の予測精度を検討したと発表した。この研究は、同院検査部の佐藤雅哉助教、矢冨裕教授、同院消化器内科の建石良介特任講師、小池和彦教授らと、島津製作所基盤技術研究所AIソリューションユニットの梶原茂樹主幹研究員らの研究グループによるもの。研究成果は「Scientific Reports」に掲載されている。

多要因が組み合わさり発症するさまざまながんに対し、単一腫瘍マーカーでの存在予測には限界がある。従来がんの有無の予測に使用される腫瘍マーカーは有効な手段だが、日常の診療においては腫瘍マーカー以外にもたくさんの情報が収集されるため、なるべく多くの情報を統合して診断を行うことが望ましいと考えられている。

研究グループは、線形回帰モデル、ニューラルネットワーク、決定木、サポートベクターマシン、ディープラーニングなど、さまざまな手法から収集された患者データから得られる予測能を最大化する学習アルゴリズムと学習パラメーターを自動抽出するフレームワークを作成し、日常の診療で得られる患者データを用いて、どの程度正確に肝がんの有無が予測できるかを検討した。なお、同フレームワークは目的とする予測対象（研究では肝がんの有無）と日常診療で得られる患者データセットを入力すると、フレームワーク内で最も予測精度の高い学習アルゴリズムと学習パラメーターが抽出され、最適な学習モデルが自動的に作られる仕組みになっている。

まず、1997年1月～2015年5月までに東京大学医学部附属病院を受診した肝疾患患者の中から、肝がんの予測モデルに入力する年齢、性別、身長、体重、HBs抗原とHCV抗体、アルブミン、ビリルビン、AST、ALT、ALP、γGTP、血小板値、AFP、AFP-L3分画、PIVKA-IIの16項目と肝がんの有無の情報が収集可能であった1,582人（肝がん患者539人、非肝癌患者1,043人）を用いて、肝がんの予測モデルの作成と精度の評価を行った。対象患者1,582人を、各アルゴリズムの訓練を行うための訓練データ（1,266人）、最適な学習パラメーターを抽出するための検証データ（158人）、作成された学習モデルの精度を検証するための評価データ（158人）の3つにわけて検討を行った。

PESI データを病理診断に利用するための装置およびソフトウェアの開発

それは2008年に山梨大・早大 医工融合研究プロジェクトから始まった

ガン診断における判断材料としてガンに特異的に検出されるバイオマーカーとよばれる生化学的分子の量の測定が広く行われています。多くの医学研究者や医療関係会社は各種ガンに対応するバイオマーカーの発見に多くの資源を投入しています。ニュートン–デカルト・パラダイムに従えばこのバイオマーカーの探索という研究開発の方針は当を得たものでしょう。しかし、個々のガンに対して１，２のバイオマーカーを探索するのは賢明なことでしょうか？　膨大な費用が掛かるるだけでなく、検証すべき仮説の数に比して必要な検証用データの収集は容易ではありません。

ガンは遺伝子上の突然変異によって引き起こされますが、それがもたらす細胞内外での代謝過程で生成される数多くの生化学的分子群の代謝経路は非常に異種多様です。同じ急性骨髄性白血病と診断されるものでも、その遺伝子の変異の仕方は多種異形です。ガンの発生に関与する代謝分子は多重に組み合わさってガンを発現しているのです。一般にガンの病名は臨床的所見にあるいはガン細胞の形態の観察による命名に過ぎず、実体論的に定義されるものではありません。ガンの物理モデルはないという論文もあります。従ってその代謝物質のひとつ二つをバイオマーカーとして特定することを試みるよりも、代謝分子群全体の中の組み合わせを観測して診断する方が遙かに理にかなっています。筆者らはこの点に着目して、肝臓ガン、腎臓ガン、大腸ガン、胃ガンの細胞から検出される極微量の液滴の質量分析データに学習機械を適用してガン診断支援装置の開発しました。しかも学習機械を適用するに当たって、精度の粗い質量分析データを用いて代謝分子群の個々の分子を同定することなく診断に成功しています。

がん診断支援装置PESI/dPLRMの開発

竹田ほか「質量分析法と統計的学習機械を組み合わせた新規がん診断支援装置の開発」『島津評論』69－3・4、2013年
https://www.shimadzu.co.jp/products/tec_news/srv69_34/report01.html

質量分析法と統計的学習機械を組み合わせた新規がん診断支援装置の開発

竹田　扇[1]　吉村健太郎[1]　出水秀明[2]　平岡賢三[3]　谷畑博司[4]
田邉國士[5]　中島宏樹[6]　堀　裕和[7]

要　旨

著者らは従前のイオン化法とは異なり前処理が不要で迅速に数 pL という極微量の試料をイオン化することが可能な探針エレクトロスプレー法（Probe ElectroSpray Ionization, PESI）と，得られたスペクトルを全て利用して経験的かつ高精度な予測が可能な学習機械（dual Penalized Logistic Regression Machine, dPLRM）を組み合わせることで，内視鏡室，検査室，手術室で即時診断可能な革新的がん診断支援装置の開発を進めている。本稿ではこの装置の基盤となっている上記二つの要素技術を簡単に紹介した後に，実際の具体的な応用例としてマウスを用いた解析，ヒトのがん組織を用いた判定例を紹介する。dPLRM を用いると一見違いがないように見えるがん組織と非がん組織のスペクトル群を高精度で識別することが可能であり，本装置の有用性が示された。

嘘、まっかな嘘、 統計学

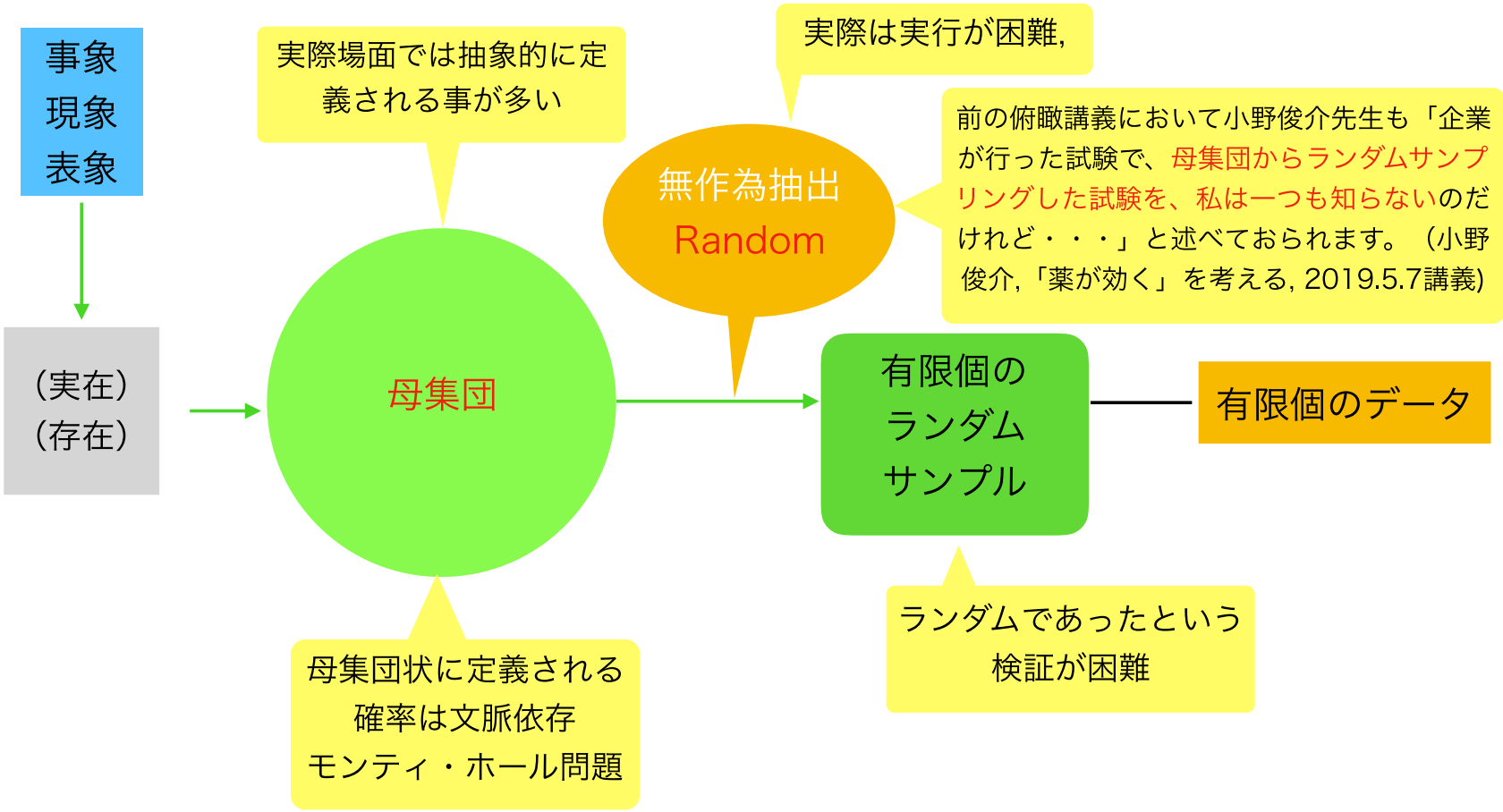
帰納推論の非論理性の問題は、古代ギリシャ人、中世の哲学者、近代の科学者を大いに悩ませた。アリストテレスの「単純枚挙」、ドウンス・スコトスの「一致法」、ウィリアム・オッカムの「差異法」など、彼らは帰納の手続きを定立することによって、帰納を演繹と同格に持ち上げようと努めたが、所詮は無理な試みであった。

帰納を救う試みとして、グロステストとロジャー・ベーコンは、帰納と演繹に加えて「試験（テスト）」という手続きを導入した。現代の統計学は、思想的には彼らに負っていると言える。しかし、試験（テスト）というものも有限の経験的事実であり、無限の事象について言及する一般命題（「カラスは黒い」）を、有限のデータから（たとえ確率的にせよ）導出することは論理的には誤りである。

統計学による推論をより客観的なものに変革しようとしたフィッシャー、ネイマンらの苦闘も、この「帰納」というものの本質に起因して、極めて限定的成功しか収められなかった。しかも彼らの理論の大前提、たとえばランダムネスの仮定、措定されたモデルが‘真のもの’を含んでいるという仮定、が成立しているかどうかは理論の外にある。大前提を理論外に放逐することによって、推論の客観性を獲得したことになるだろうか？

統計的データ解析において広く用いられている仮説検定法も、確率論を用いて一見論理的装いをこらしているが、潜在的に前提する想定が正しければ確率が計算できるのであって、想定の確かさに何ら保証がないことに留意する必要がある。

母集団という概念



確率と母集団

確率を想定するには、確率の対象となる母集団の概念が重要である。

モンティ・ホール問題，確率の文脈依存生

母集団は想像上の抽象的概念であり、検証することが難しい。

実際、現実の場面では母集団の定義が曖昧であることが多い。
「企業が行った試験で、母集団からランダムサンプリングした試験を、私は一つも知らないのだけれど・・・」（小野俊介,「薬が効く」を考える, 2019.5.7俯瞰講義)

「統計でウソをつく方法」というものがありますが、多くの場合母集団がははっきりしていないデータであることが多い。

統計学には2つのキャンプがあり、確率に対する考えおよび統計モデル構成の違いから論争があり決着がつかない。
ベイジアン　と　頻度主義統計学者

統計的仮説検定法

「有意性検定」においては、観測できる事象を構成する要素的事象に関する特定の確率分布（モデル）を措定する．このモデルは「帰無仮説」と呼ばれる、このモデルが真であると仮定した時に稀にしか起こらない事象群を確率の計算に基づいて定め、観測データがこの事象群に入ったならばこの仮説を「棄却」するという手続きをとる．措定された確率分布は人間が勝手に想定したものであるから、観測出来る生の事象の確率をこれから導き出すことは出来ないことは自明である．実際、棄却が論理的に意味することは「仮説の下で非常にまれな事象が起きたかあるいは仮説とされた確率分布が真ではないかのどちらかである」とフィッシャー自身も述べている。

ちなみに、この手続きには別の問題もある．一体どのくらい小さければ稀と判断すべきであろうか．非常に稀な事象を定める際に確率（有意水準と呼ぶ）としては5%や1%がよく使われるが、その根拠が判然としない

この検定の手続きは仮説を没にするための意志決定に関わるもので、仮説が棄却された時に、仮説の否定が支持されたとするのは（一見背理法のように見えるが）拡大解釈であり間違っている．世間ではこの点に関する誤解が少なくないようであるが、仮説検定論のこの論理的性格は十分に認識しておく必要がある 。

１９８３−８７年の間にドイツのダルムスタットのGIS研究所において、原子核の衝突実験から2回注目すべき新現象があることが見つかった．既知の物理から分かるバックグラウンドから標準偏差の約6倍の位置に離れて突出するものであり、‘信頼レベル’は99.9999%に相当するものであった．それが本物ならノーベル賞に値する新粒子の発見となる．この間、これに関しては百を上回る実験と理論の論文が公刊された。ところが10年を経てこの現象の存在を否定する報告がなされ、大論争となった．1996年のPhysical Review Letters誌には大御所２人による賛否両論が掲載された．翌年発行されたScience誌には、多数の物理学者によって‘新発見’は幻影であったと結論され、１０年の歳月と何百万ドルが無駄になったと報じている．同誌はこの件を‘an illustration of the questionable, perhaps delusive(偽りの), power of statistics’であるとしている．2015年にもスイスのCERNの Large Hadron Collider 実験データ解析でも同様なケースが報告されている。

The One That Got Away?, Science: News and Comment, 275, 10 January, 1997

統計学における人を誤らせる用語法

有意性検定　帰謬法（背理法）ではない！！

「統計学的に有意」であることは、現実的に有意であることを意味しない。「有意」であることと、例えば「臨床的に意味がある」こととはまったく別のこと。（「有意水準5%という社会的慣例ルールに従ったら「有意」と出ました。」ということに過ぎない）なぜ有意水準が5%なのか、0.01%ではダメなのか？　有意水準の値の採り方次第でコロコロ結論が変わる。

P値

近年やっとP値に対する批判・再検討がなされている。

Ronald L. Wasserstein & Nicole A. Lazar, The ASA's Statement on p-Values: Context, Process, and Purpose, http://amstat.tandfonline.com/loi/utas20

Andrew Gelman, Eric Loken, Data-dependent analysis—a "garden of forking paths"— explains why many statistically significant comparisons don't thold up., The Statistical Crisis in Science, American Scientist

信頼区間

真値が90%-信頼区間に入っている確率が0.95であることを意味しない。！！

社会的承認

これまで異端視されてきたベイズ統計学や機械学習も、近年やっと日米の規制当局であるPMDAやFDAが承認するようになっている。

実際、FDAも去年4月に初めて画像データに基づく機械学習による診断装置を承認した。

推論方法としての機械学習の根本的な弱点

機械学習の問題点が払拭された訳ではない。

・訓練用データセットがどのような母集団に由来するかが不明確である事が多い。

・したっがって、未知のデータ対して、訓練済みの学習機械を適用するする際に、訓練に使われたデータ・セットの母集団に未知データが属するか否かが不確かなので、学習き機械が出力する帰結がどのくらい信頼できるかが分からない。

・訓練不足の問題（数値的最適化が不十分の場合）

説明とは何か？　　解釈できるとは何か？

機械学習はブラック・ボックスなので、どのように推論したかを説明できない、分からないという批判があるが、機械学習の推論過程を自然言語で追尾して理解することは不可能であると私は思います。説明ができるような機械学習に限定・規制してしまうと、機械学習のもつ巨大な力を滅殺してしまうことになると考えています。

「説明には上手な説明と下手な説明があるだけ。正しい説明と間違った説明があるわけではない」（上野千鶴子・東大教授）

機械学習の適用領域と限界

学習機械は帰納的な推論機構なので、帰納という営為の持つ本質的限界を免れない。すなわち有限個の学習データセットから推論機構を学習するため、学習データセットから大きく外れた未知データに対する予測力はない。しかもそれが外れたデータであるか否かは予めわからない。この点は学習機械の適用に当たって十分留意する必要がある。また学習機械による推論にはその機序が人間には解釈不能であると言う問題もある。解釈して‘判る’という心の働きは解釈者が生得的および経験的に獲得した心象、概念、言語、数学などの既知の知識（これら自体もモデルの集合とみることが出来る）に照らして馴染みがあるか否かという主観的なものである。

現代の我々の頭脳は因果推論に深く染まっているので因果の連鎖をすべて提示されない限り、入出力の関数関係がブラックボックスでは‘判る’ということにはならない。しかし近い将来に学習機械のモデルが様々な分野で実用化される時代が来れば、事情は変わるものと思われる。解釈可能性の概念は歴史的・文化的文脈に強く依存するものである。「飛行機はなぜ飛ぶのかはわからない」という識者の発言を聴いたことがある。機械学習はエンジニアリングの分野には大きなインパクトをもって迎えられると思われるが、ここしばらくは科学界には受け入れられそうにない。しかし科学研究の探索的フェーズでは大いに活用されることを期待したい。