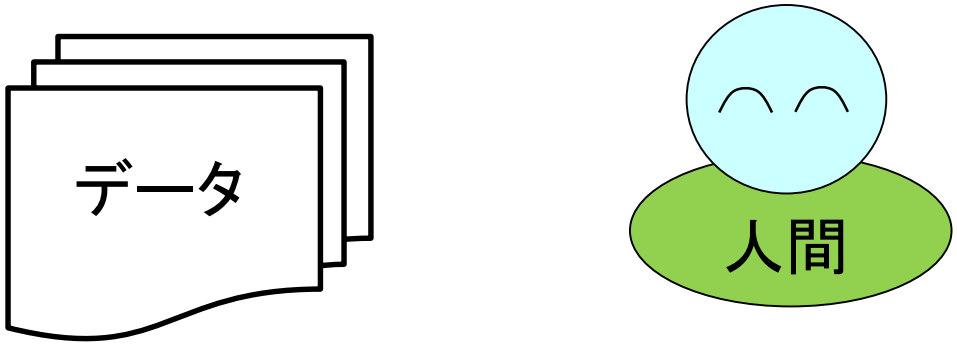
データ解析

第１回：統計モデルは真ではない。最後の敵は「実世界」。

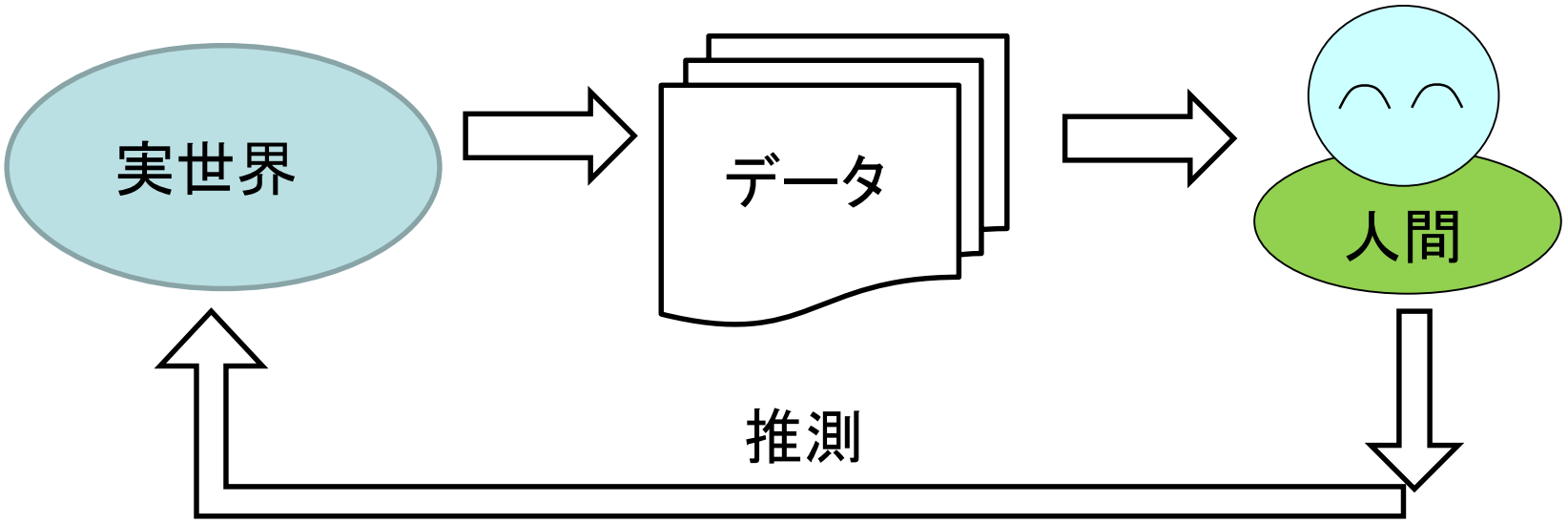
渡辺澄夫

データと人間



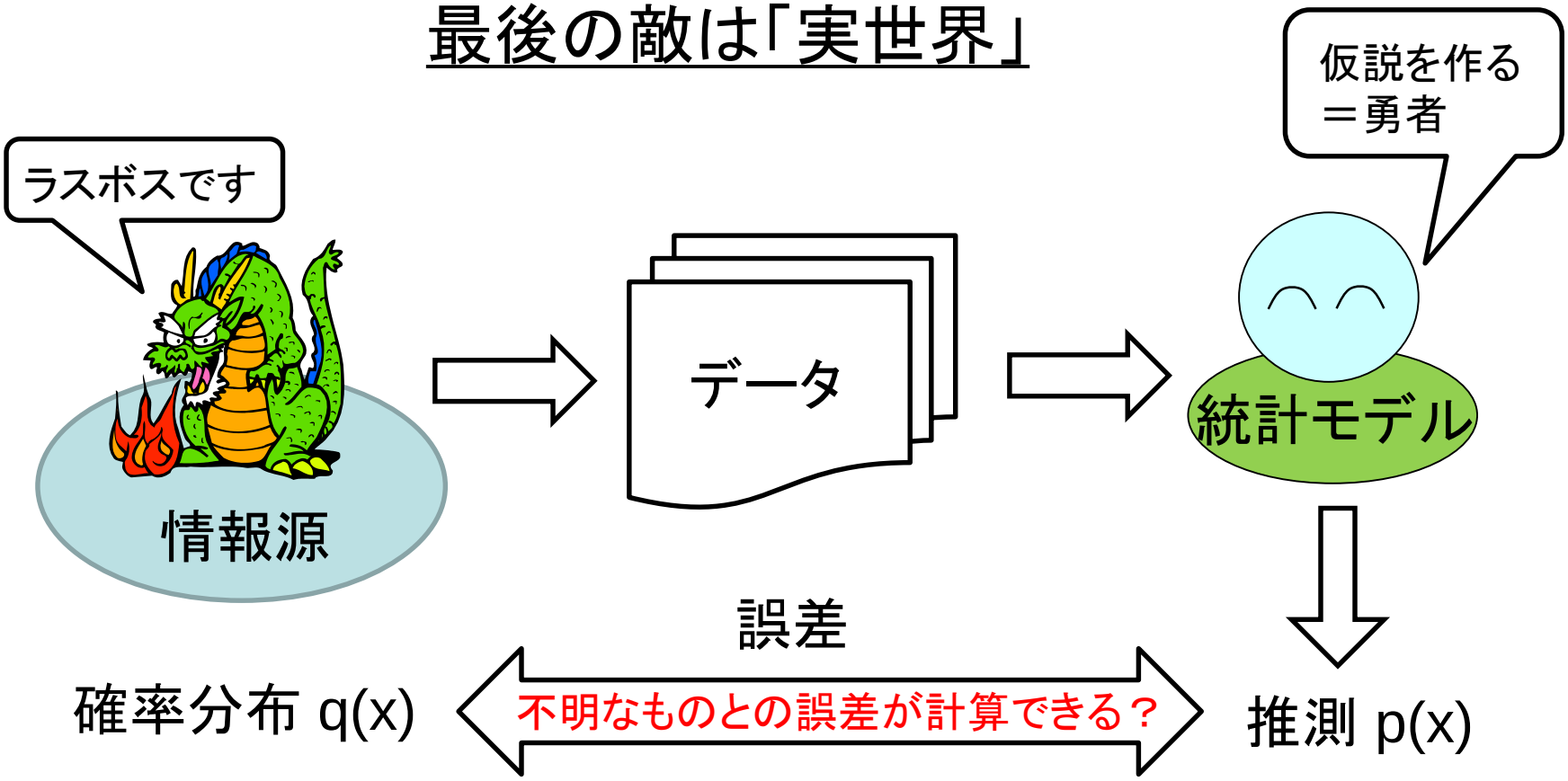
データと人間がある。コンピュータやネットワークの発展のおかげで
膨大なデータが得られるようになった。

データ解析とは何ですか



データは実世界を観測して得られたものである。「実世界」とは人間から最も遠く離れた恐ろしい未知なる存在である。そこからデータが得られる。人間は生きるためにデータから実世界を推測する。データは有限であるから推測は正しくない。実世界は最後まで不明のままである。
これがデータ解析という物語の舞台である。そこには「演繹的に正しい」と言える結論は一切存在しない（非適切設定問題）。

最後の敵は「実世界」



実世界（＝情報源）が確率分布 q(x) である場合を考えよう。
人間は統計モデルを仮定しデータを用いて情報源を推測する。モデルは仮説であり実世界ではない。推測された結果 p(x) は間違っている。推測はどの程度に正しいのだろうか。弱い人間にできることはあるのだろうか。

統計学とは

※　このページに書いてある内容が理解できなくても気にする必要はありません。

確率論では、まず基礎となる確率を定義し、その上で様々な確率変数が従う
数学的法則を導出します。それは純粋に数学的なものです。美しく正しいです。
特別な場合を除いて人間は安心して暮らすことができます。

統計学では、基礎となる確率がわからないという状況を考えます。基礎となる確率が
わからないので、すべての方法すべての結論が正しくありません。それが統計学では
普通のことなんですが、それでは人間は安心して暮らすことができないです。

「すべての結論が正しくない」という状況にひとの心は耐え難いからです。そこで
そこから逃避するために昔の統計学者はしばしば「主義」を設定し、その「主義」の
もとで正しいかを論争してきました（例：最尤主義、ベイズ主義）。不毛な論争です。

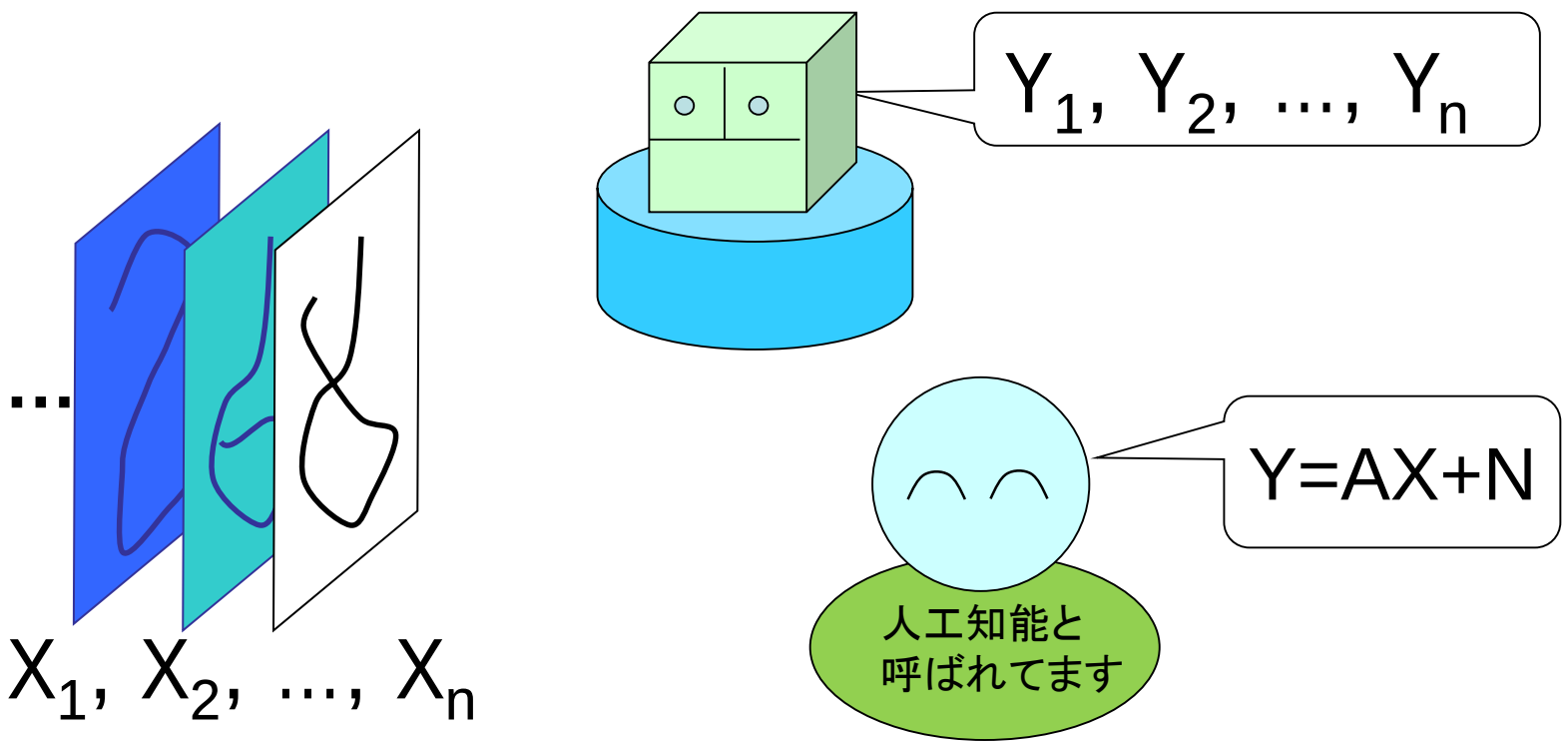
今日、統計学の本当の目標は「実世界」を推測することであり、その目標に対して
「主義の正しさを決めること」に意味はないことが知られています。学生のみなさんに
とって「安心して暮らせない学問」に出会うことは生まれて初めてかもしれませんが、
その状況から逃げず、がんばってください。生きるとはそういうことだろうと思います。

統計的モデリングは実世界に対する人間の挑戦であり、それは人工知能ではでき
ません（今のところ）。だからこそ職業もあって仕事もあるのかもしれないですね。

学ぶこと

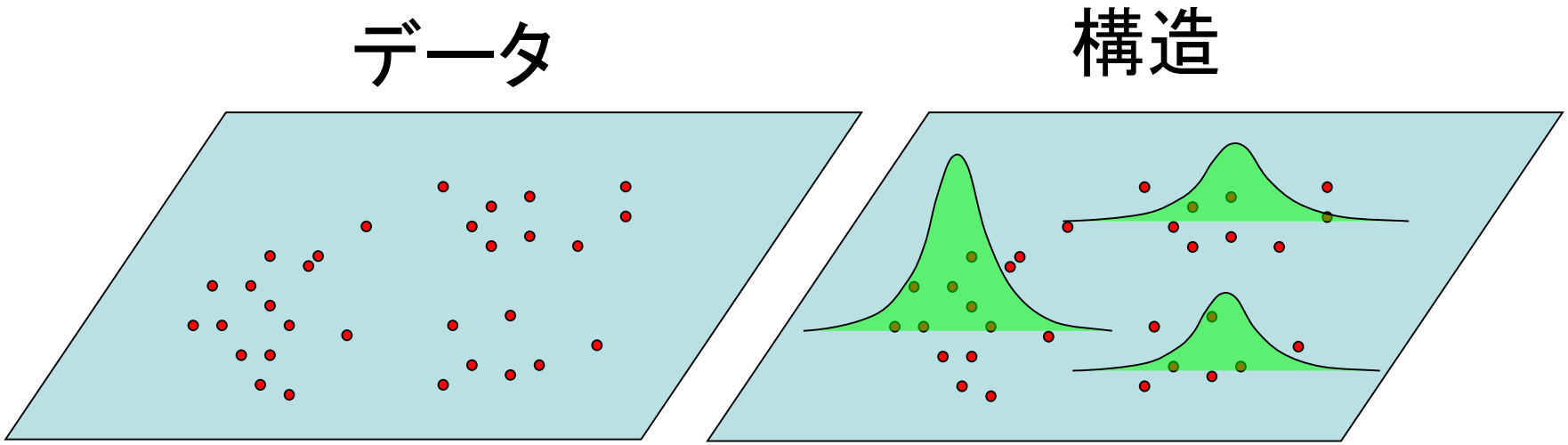
1. 回帰・判別分析
2. 因子・主成分・クラスタ分析
3. 時系列予測
4. ベイズ法・階層ベイズ法
5. 統計的検定
6. 評価法
7. 実世界

回帰・判別分析



確率変数 (X,Y) のデータが得られたとき、XからYへの条件つき確率を推測する方法が回帰分析・判別分析である。認識・予測など応用が広い。簡単な問題なら線形回帰で解けることも多い。近年、脚光を浴びている深層学習やサポートベクタマシンは、この道の奥にある。

因子・主成分・クラスタ分析

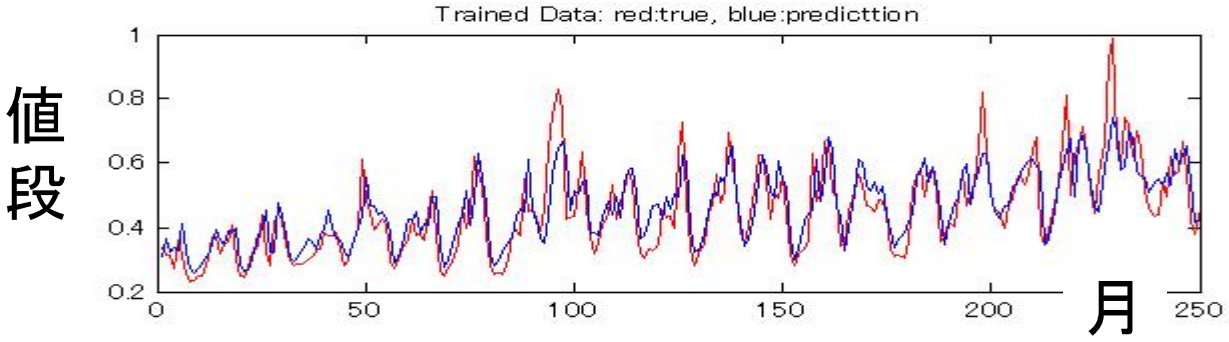


確率変数 X のデータが得られたとき、(Xに対する情報Yが与えられていなくても)、X の構造や潜在変数を取り出す方法が因子分析・主成分分析・クラスタ分析である。

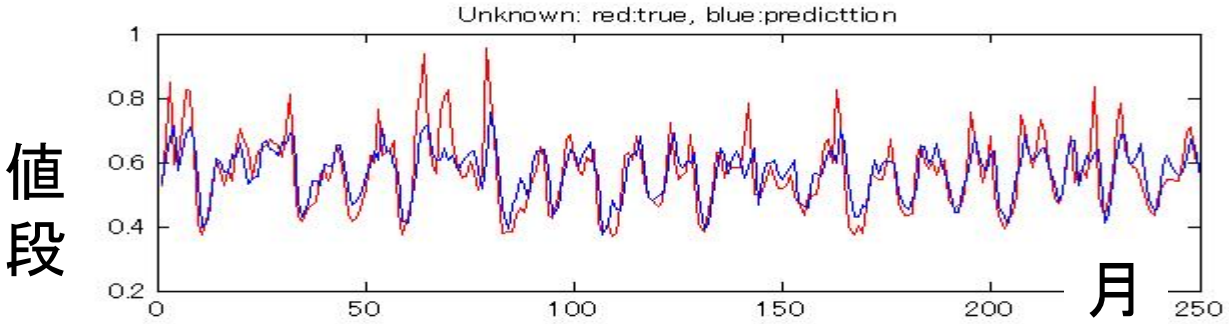
構造や潜在変数を取り出す数理的な技法だけでなく、得られた構造や潜在変数が何を意味するかを考察する力量が必要になる。データサイエンティストへの第一歩。またトピックモデルなどもこの領域に属する。

時系列予測

例：1970年1月から2013年12月までの白菜の値段
「政府統計の総合窓口」のデータを使用しています。
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do



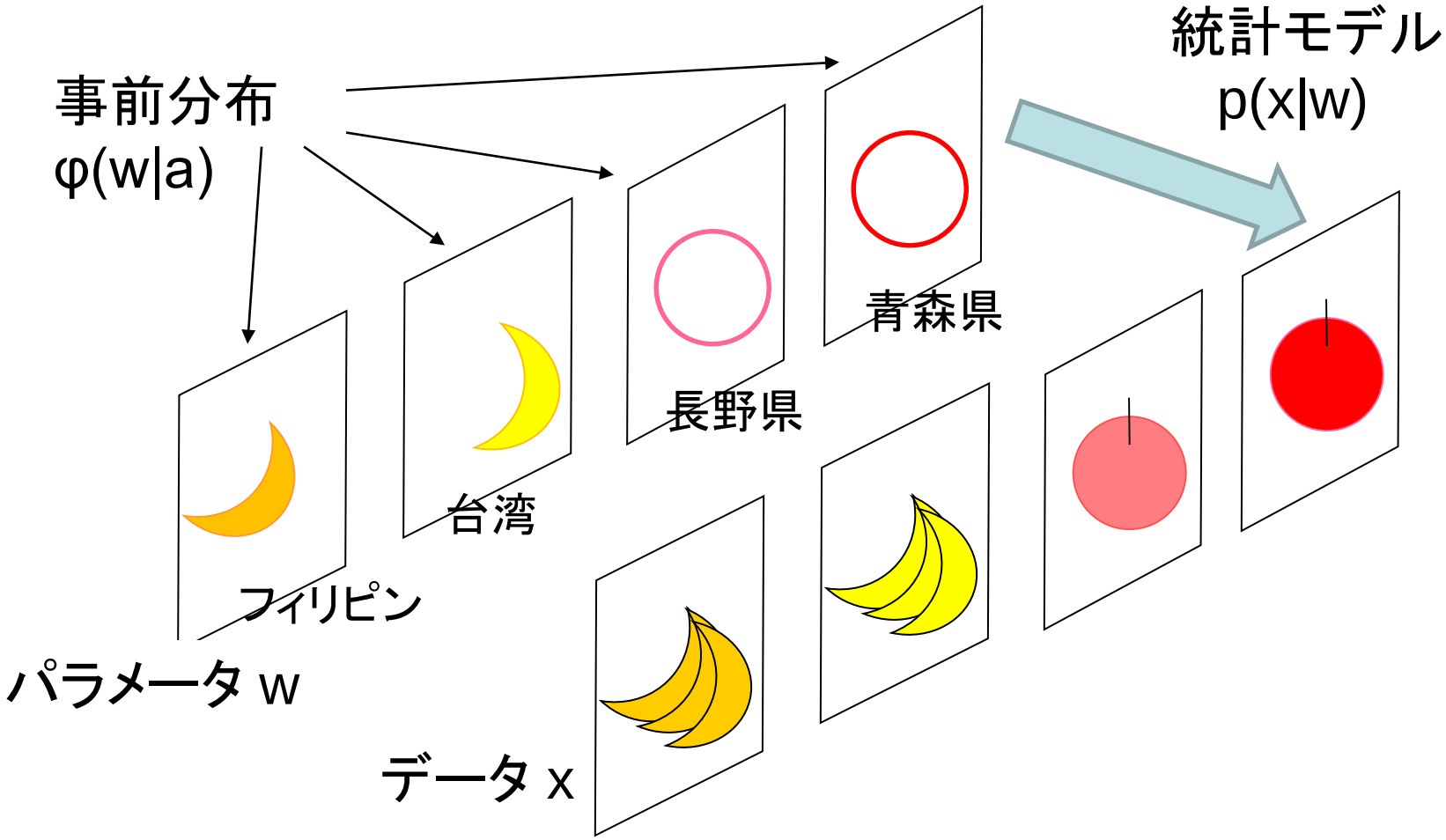
学習データ
赤：真
青：学習結果



テストデータ
赤：真
青：予測

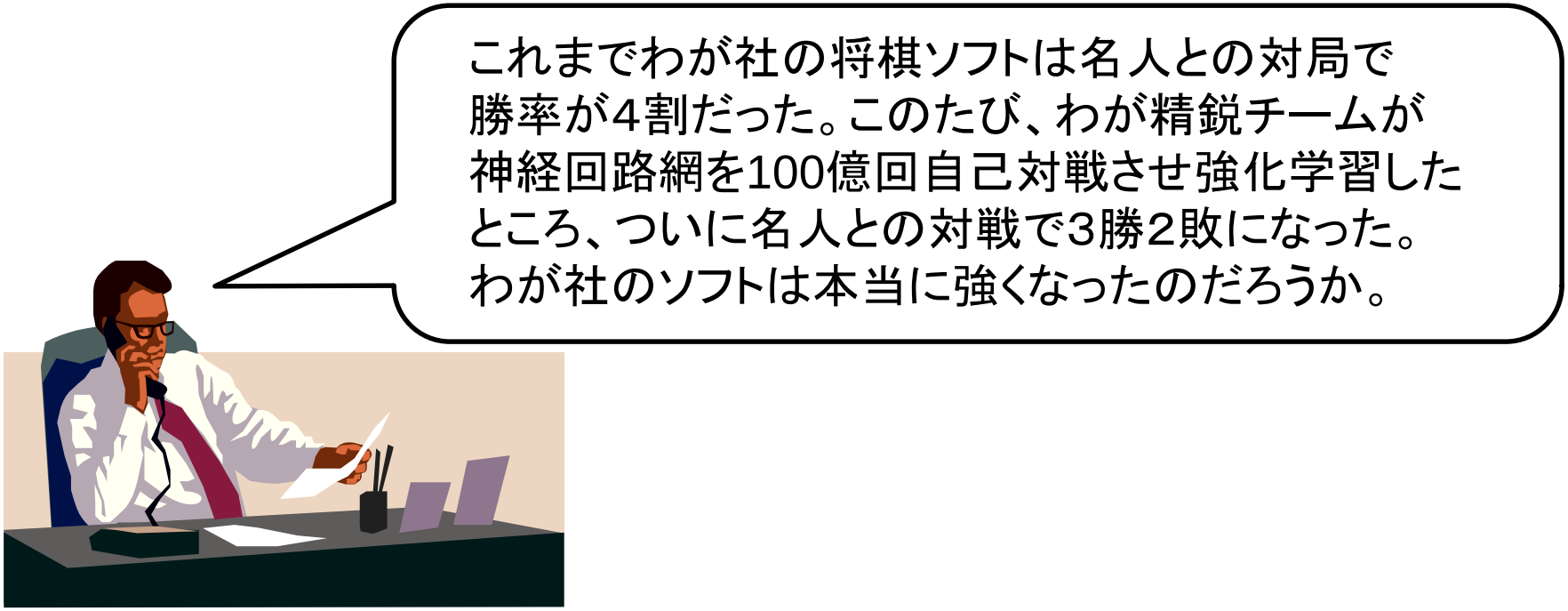
過去の時系列から未来を予測する。回帰分析とよく似ているが X が時間とともに変動する点が異なる。経済・金融などへの応用は多すぎるほどある。

ベイズ法・階層ベイズ法



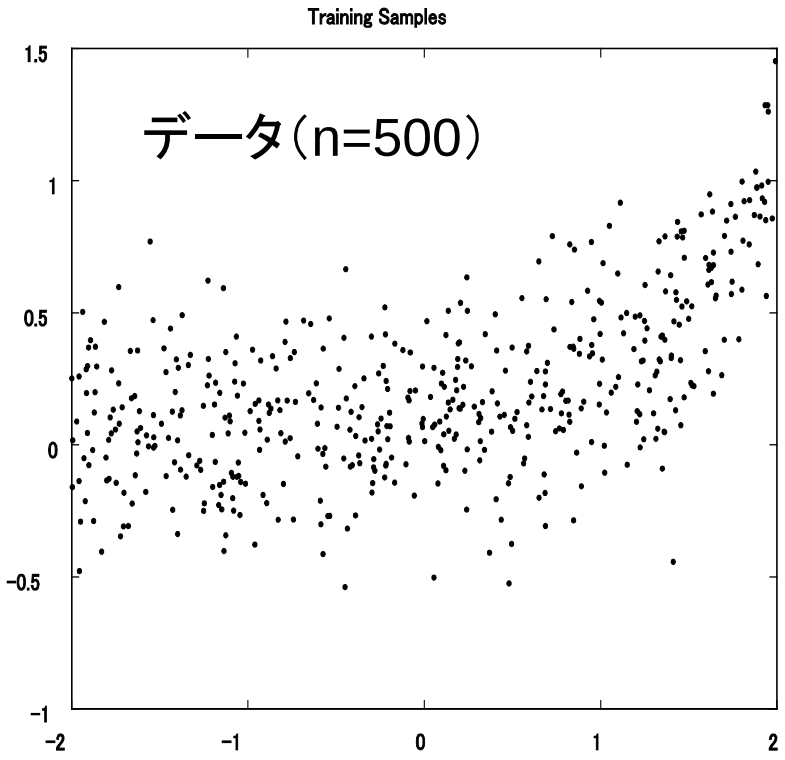
果物の特徴 x の確率分布を定めるパラメータ w は、果物が取れる地域により似ている点と異なる点がある。事前分布も含めてモデル化し推測しよう。現実のデータ解析で頻繁に現れる構造である。

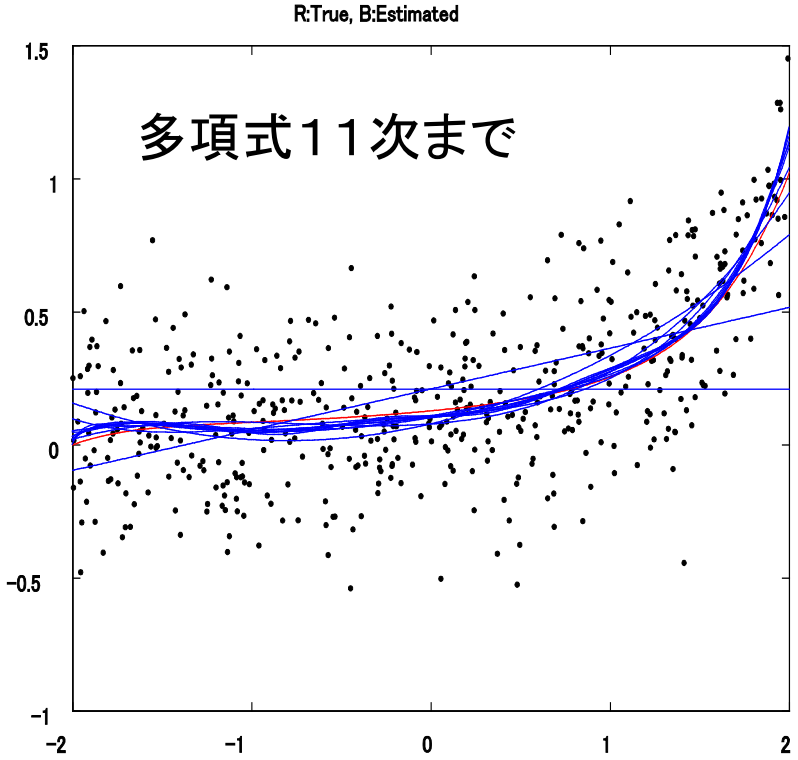
統計的検定



真の確率分布がわかることのない実世界から得られたデータに対して、弱い人間はデータの解析を行なって何らかの推測を行なうがその推測はどの程度に正しいのだろうか。統計的検定では、実世界に対して仮説を作りデータの説明可能性を数量的に調べる。

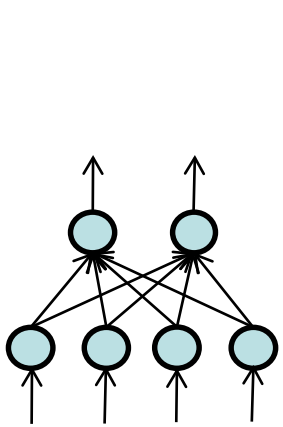
評価法



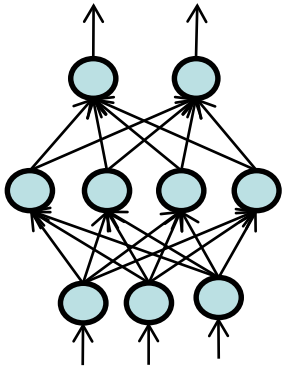


真の確率分布がわかることのない実世界から得られたデータに対して、
弱い人間は複数のモデルの候補を作り、複数の推測を行なうことができる。
その中に実世界と一致するモデルは存在しないが、最も適切なモデルは
どれだろうか。さらに、そこで「適切さ」とは何に対する適切さだろうか。
現代統計学の基礎である情報量規準とクロスバリデーションを学ぼう。

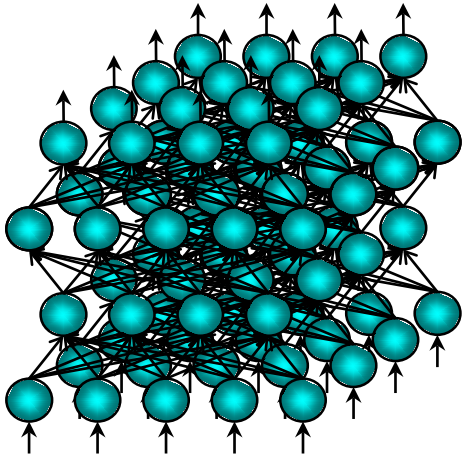
現代から未来へ



1960



1985



2015

2040

この講義では「データ解析」の初歩なことを学びます。今日ではデータの次元も個数も大きくなり、統計モデルも高度化しています。「人工知能」あるいは「機械学習」と呼ばれる分野では極めて大きな複雑さを持つモデルが使われますが、数学的には統計学と等価です。単なる道具として「データ解析」の使いかたを覚えるだけでなく、それを支えている数理について理解することはあなたにとって、変わることのない「確かに拠って立てる基盤」になるでしょう。

旅の地図

実世界

評価法

観測データ

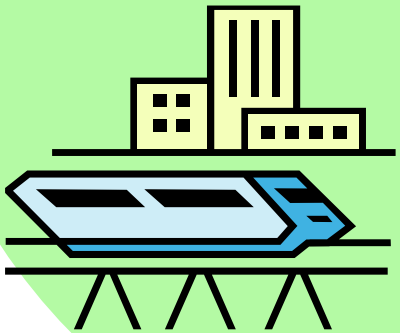
回帰・判別分析

推定と検定

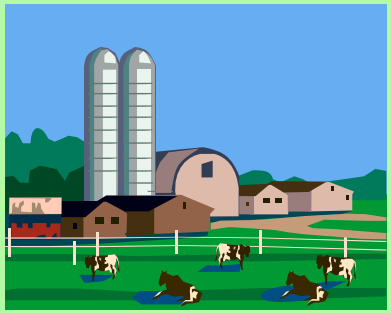
統計的検定

ベイズ法
・階層ベイズ法

解析方法



時系列予測



因子・主成分
・クラスタ分析

（付録）この講義の履修に必要な基礎数学

1. 線形代数

ベクトル, 行列, 線形写像, ランク, 行列式, トレース,
内積, 転置, 固有値, 固有ベクトル, …

2. 微分積分

極限, 収束, 発散, 偏微分, 多重積分, …

3. 初等確率統計

確率分布, 確率密度関数, 確率変数, 条件つき確率,
大数の法則, 中心極限定理, ベイズ推定, 最尤推定, …

※上記の3つをまだ習っていない人は、この講義の履修は
来年以降にして、まずは上記の3つを習いましょう。

※測度論に基づく確率論はこの講義では使いませんが
非常に重要ですので卒業までには勉強しましょう。

線形代数

(1) 任意の M×N 行列 A, B に対して $\mathrm{tr}(AB^T)=\mathrm{tr}(B^TA)$

(2) 任意の N×N 行列 A, B に対して $\det(AB) = \det(BA)$

(3) 任意の N×N 行列 A, B に対して $\det(e^A) = e^{\mathrm{tr}(A)}$

(4) 任意の N×N 行列 A について、ある可逆行列 P とあるジョルダン標準形の行列 J が存在して $A=P^{-1}JP$ と書ける。

(5) 任意の実 N×N 行列 A が $A=A^T$ であるとき（対称行列であるとき）ある直交行列 P とある対角行列 J が存在して $A=P^{-1}JP$ と書ける。

(6) 実 N×N 行列 Aが条件「任意の実N次元ベクトル $x\neq0$ に対して $(x,Ax)>0$」をみたすとき A を正定値行列という。正定値行列Aは対称行列であり、その固有値はすべて正である。

微分積分

(1) N次元実変数 x から実数への関数 f(x) がC^1級関数であるとする。開集合 U 内のある点 x=a で f(x) が極大または極小になるならば
$\nabla f(a)=0$.

(2) f(x) が上記と同じ条件を満たすとする。任意の a と x について
$|x-a|>|x-a^*|$ を満たすある a^* が存在して

$$f(a+x) = f(a) + (x-a)\cdot\nabla f(a^*).$$

(3) N次元実変数 x からN次元実変数 y への関数 y=g(x) が C^1級関数で単射であるとする。この関数のヤコービ行列の行列式の絶対値を
$|g'(x)|$ と書くと

$$\int f(y)\,dy = \int f(g(x))|g'(x)|dx.$$

確率論

(1) 確率変数の定義

(2) 確率変数の独立性の定義

(3) 確率変数の平均と分散共分散行列の定義

(4) 確率変数列の概収束、平均収束、確率収束、分布収束の定義

(5) N次元実ユークリッド空間 に値をとる確率変数 X が有限な
平均 m=E[X] と有限な分散共分散行列 S=V[X] を持つとする。

独立な確率変数 $X_1, X_2, \ldots, X_n$ が X と同じ確率分布に従うとき

$$(1/n)\sum_{i=1}^{n} X_i \text{ は } m \text{ に概収束（従って確率収束）}$$

$$(1/n^{1/2})\sum_{i=1}^{n} (X_i-m) \text{ は正規分布（平均O,分散S) に分布収束}$$