

データ解析

第15回：実世界

渡辺澄夫

実世界 との戦い



実世界

観測データ



評価法



回帰・判別分析

推定と検定

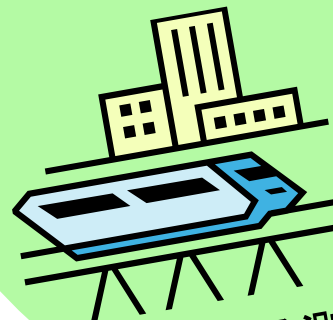


統計的検定



ベイズ法
・階層ベイズ法

解析方法



時系列予測

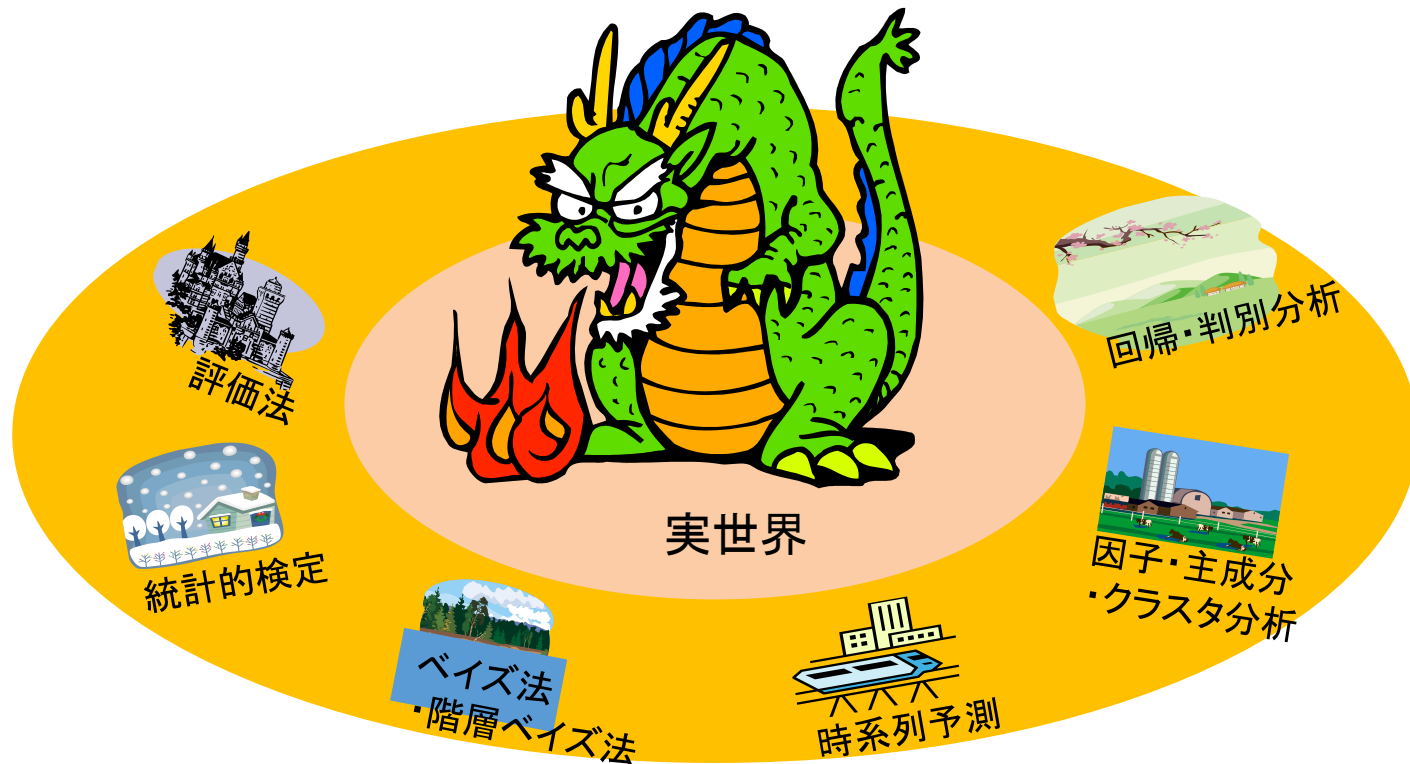


因子・主成分
・クラスタ分析

圧倒的に強大な実世界

実世界は、超高次元、際限なく複雑、非定常...

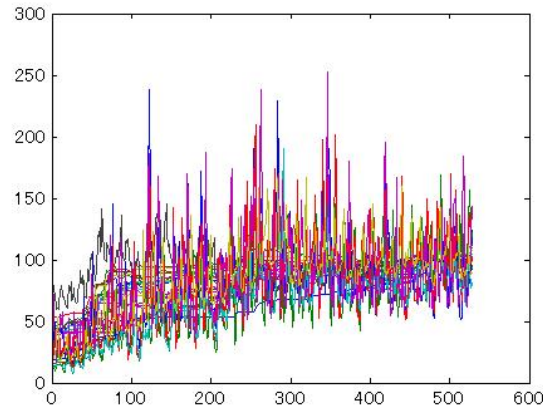
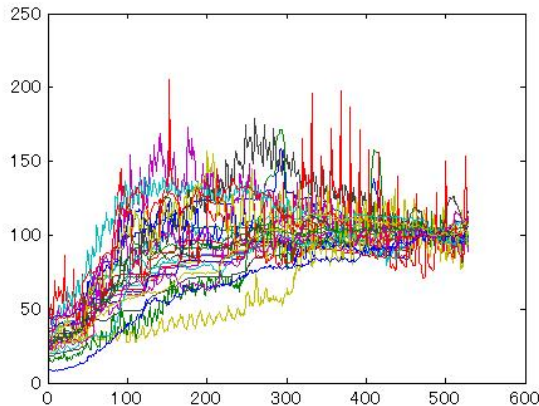
簡単なデータ解析だけでは、もう戦えない...



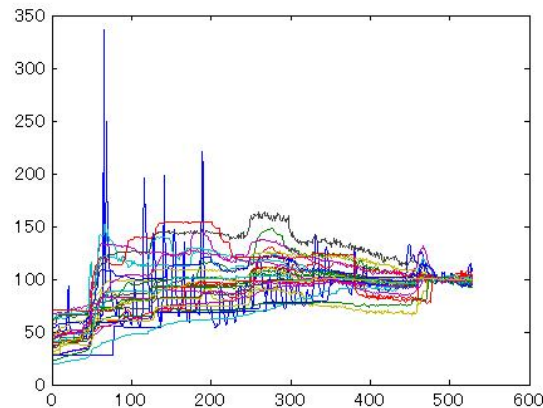
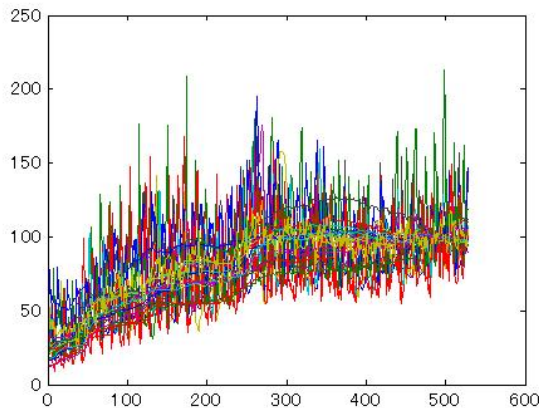
実世界の例

1970年1月から2013年12月までの108品目の値段
「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>

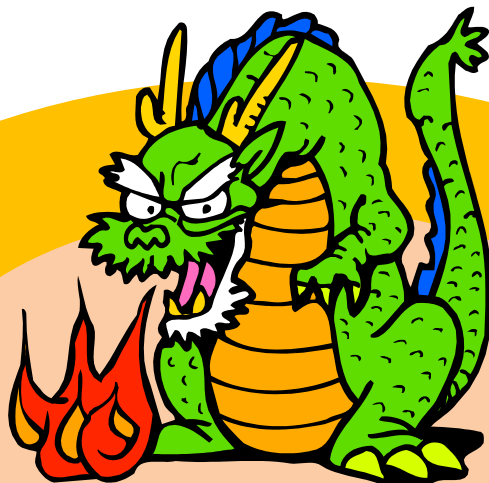


このデータは
108×528であり
小規模であるが
それでも十分に
複雑である。



どのような目的で
どのような方法で
どのようなモデルを
用いて解析を
行なっていくことにな
るのだろうか

新しい
世界へ



実世界

観測データ



深層学習

機械学習



ベイズフィルタ
空間統計



カーネルマシン

学習理論



経験過程



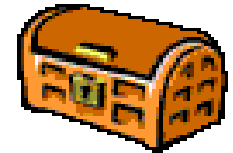
代数幾何



微分幾何

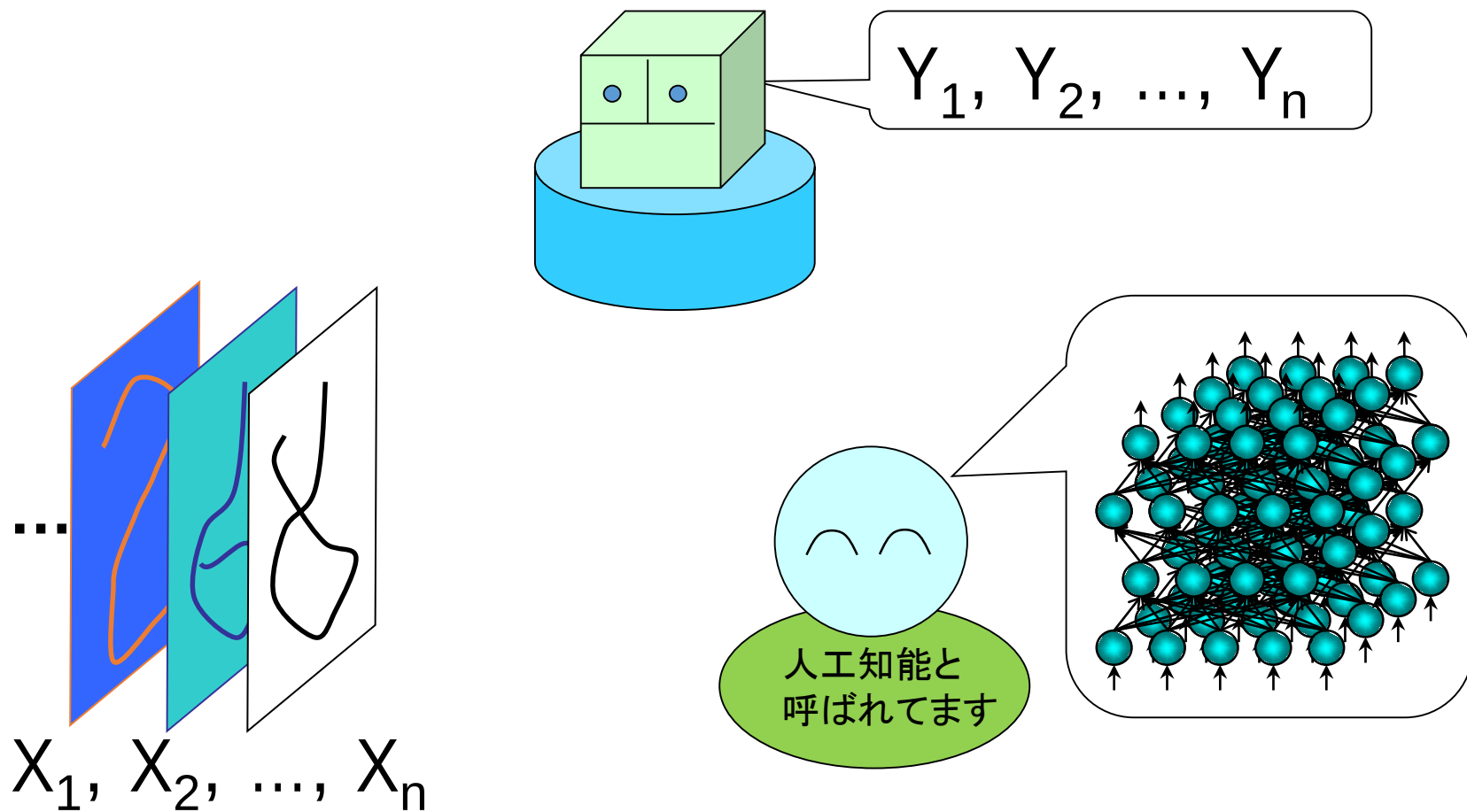
実世界と戦う

1. 実世界は、超高次元、際限なく複雑、非定常・・・
昔ながらのデータ解析では、もう戦えない・・・
2. 真に強大な敵と戦うために作られたのが機械学習・・・
深層学習、各種ベイズ法、カーネルマシン・・・
3. 人工知能は部分的には人間の能力も超えつつある・・・
しかし、それらを解析するための理論的な基盤がない・・・
4. もう昔の統計学は使えない・・・
新しい数学が必要だ・・・ ← いまここ



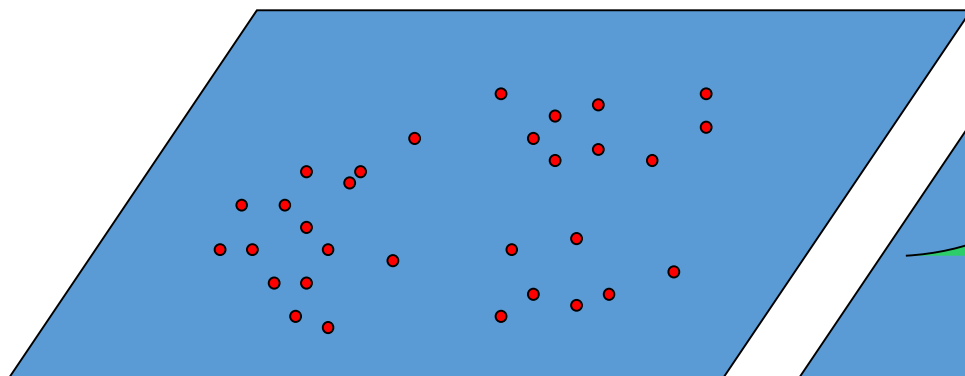
実世界に向かって

回帰・判別 → 深層学習

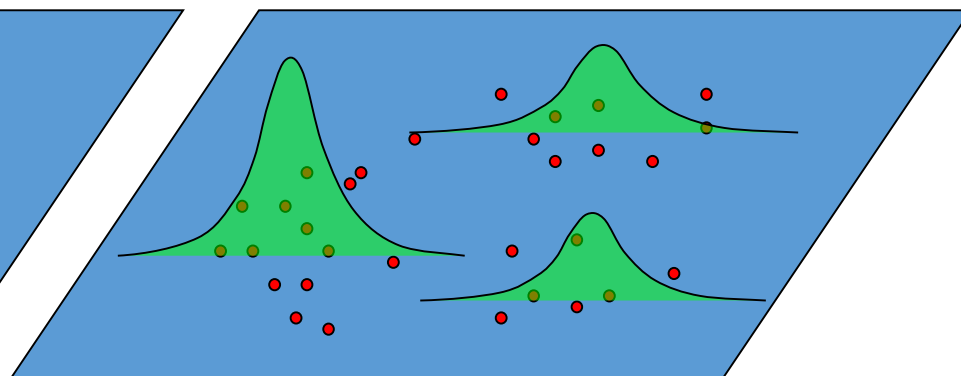


因子・主成分・クラスタ → ノンパラベイズ・トピックモデル

データ



構造



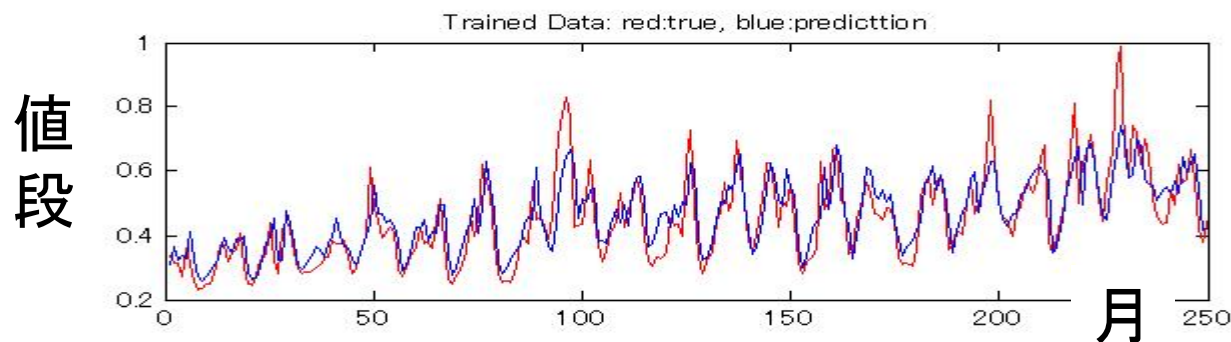
一個の点がひとつの文書であり、その文書にはトピックがある。
また文書が属する分野(政治、科学など)がある。こうした構造を
条件つき確率で取り出して、階層ベイズ法を適用する。

時系列予測 → ベイズフィルタ

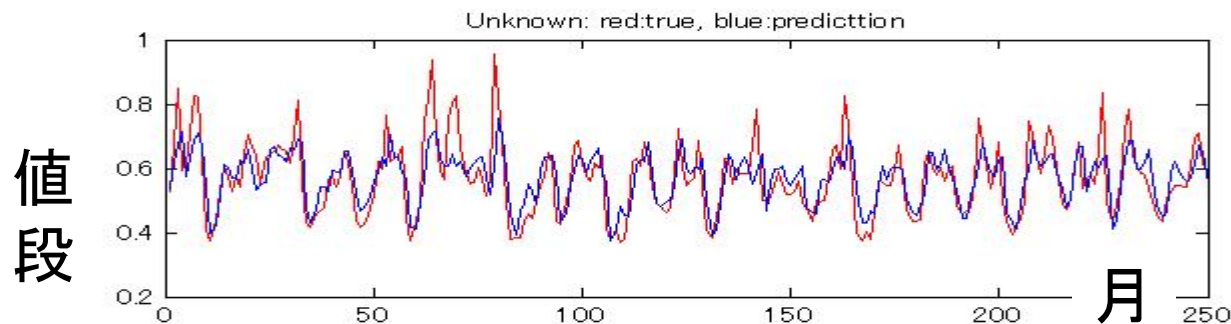


例: 1970年1月から2013年12月までの白菜の値段
「政府統計の総合窓口」のデータを使用しています。

<http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do>



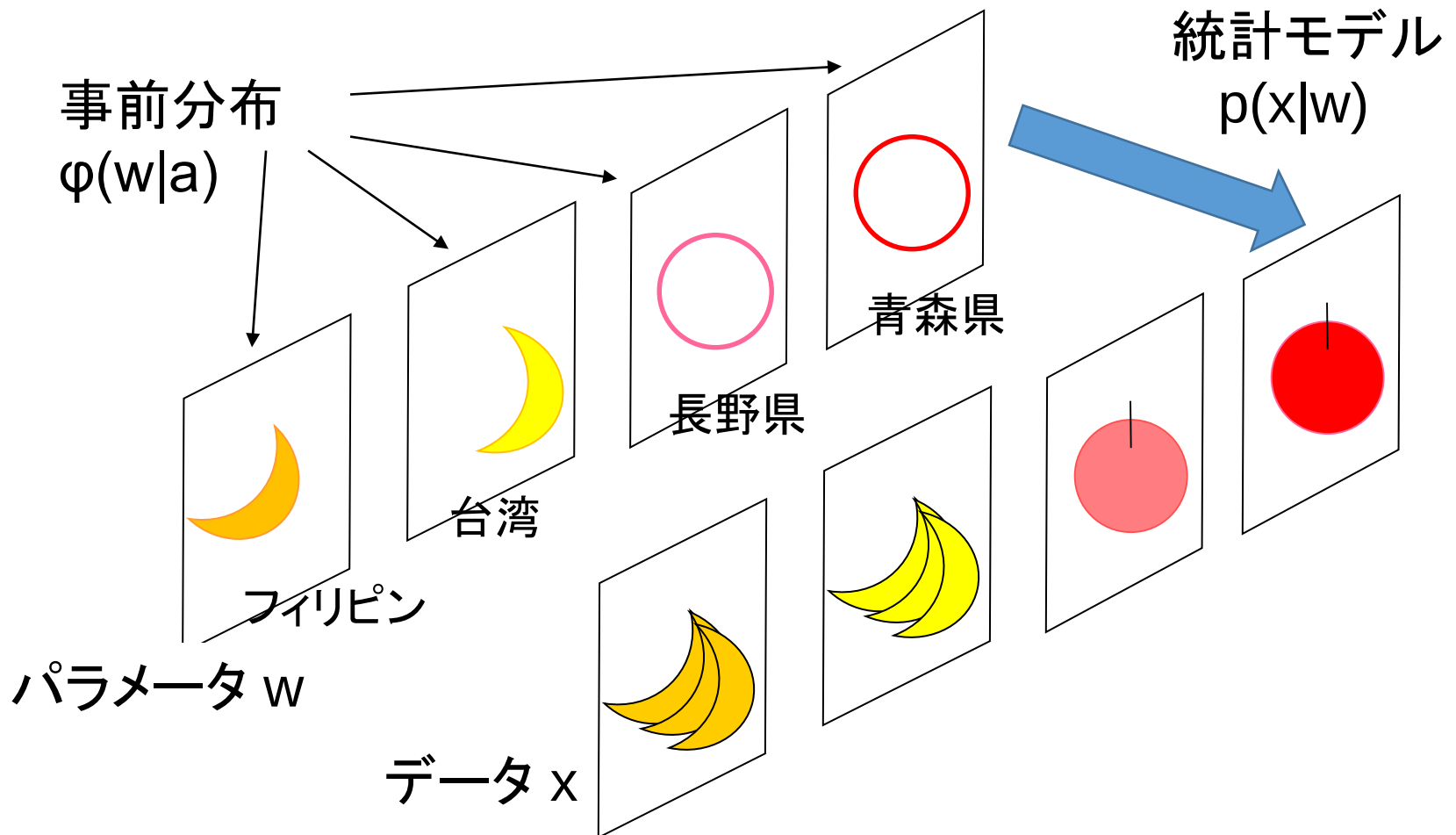
学習データ
赤: 真
青: 学習結果



テストデータ
赤: 真
青: 予測

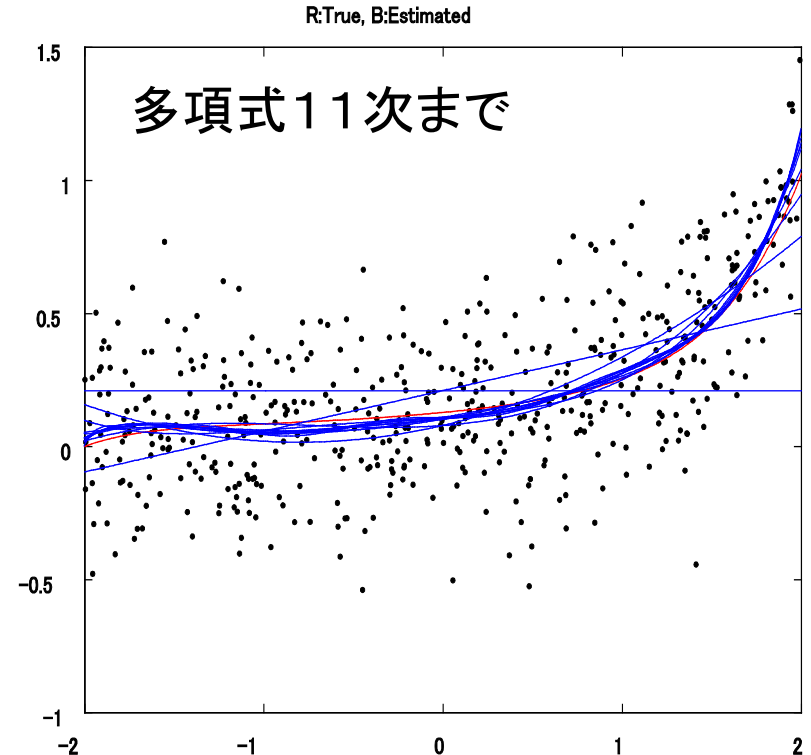
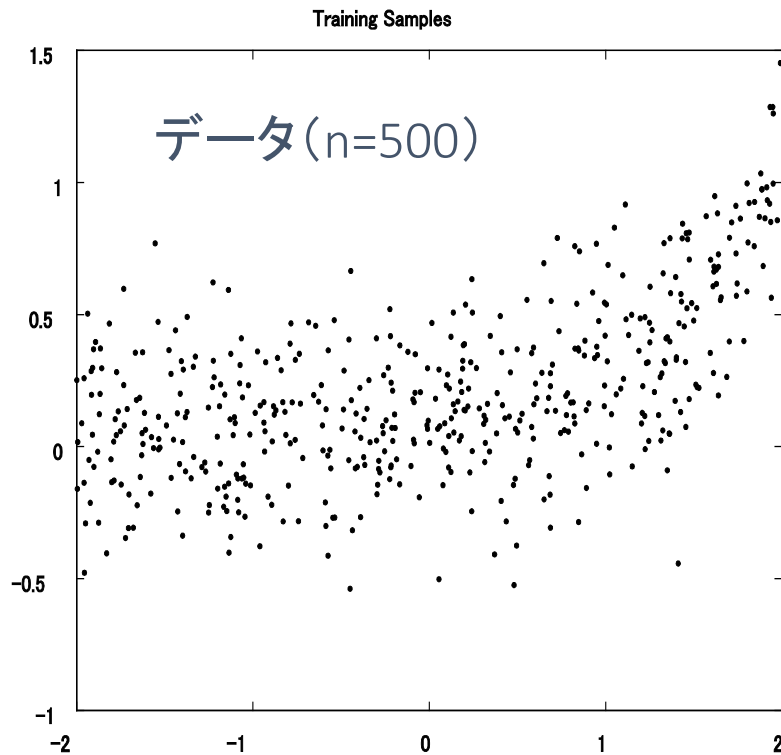
時系列の背後に観測できない時系列がある問題を考えて、階層ベイズを時間とともに変動するモデルにして解析する。今日の標準手法。

ベイズ法・階層ベイズ法



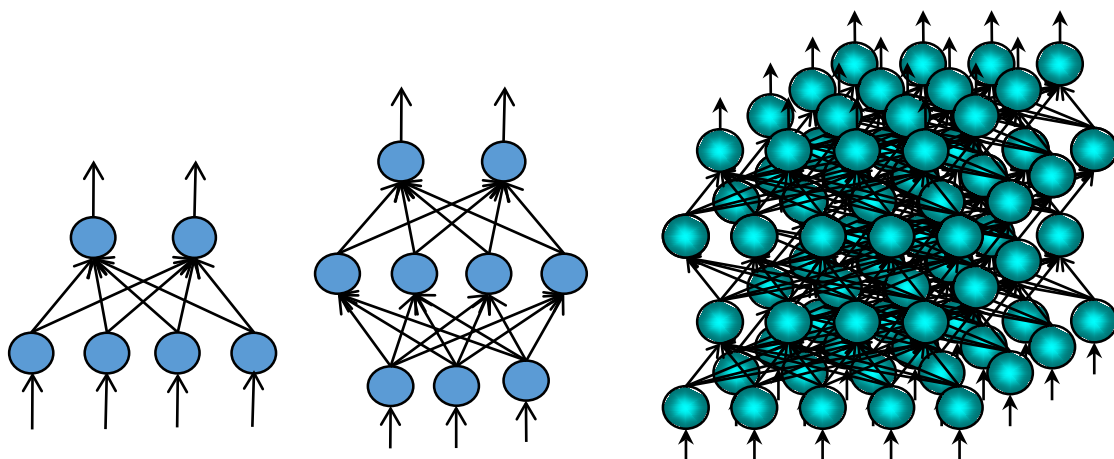
階層ベイズ法で、ハイパーパラメータ同士の相互作用をモデルに取り込むことにより、空間統計学への道が開拓される。

評価法 → 新しい規準



機械学習で現れる高度なモデルを評価するための新しい数学的方法と評価のための指標が構築されつつある。

現代から未来へ



1960

1985

2015



2040

「人工知能」の仕組みは、モデルが複雑になっただけで「統計解析」と同じです。しかしながら、今日の人工知能はパワーアップしてきており、場合によっては人間の職業を奪うのではないかと心配も生まれてきました。情報科学は人間社会を変えます。どんな社会が望ましいかを考える必要があります。

現在

深層学習は召喚魔法？

深層学習は、画像・音声・自然言語のように対象が複雑になればなるほど、問題が大規模で難しくなればなるほど実力を発揮するが、適切に使うにはノウハウが必要になる。このため「**深層学習は召喚魔法(?)**」という意見もある。

- ◎ ネットワーク構造の決め方
- ◎ パラメータ初期値の設計法
- ◎ 正則化項の作り方
- ◎ 最急降下法のとめ方
- ◎ アンサンブルの作り方 など

召喚士 募集中

深層学習を
自由自在に操り
真に強い敵と
戦える才能を持つ
人を探しています。

「誰でも使えるようになってほしい」という要望が多数。

複雑なモデルが必要でないこともある

簡単な問題なら「神経回路網＋ベイズ」を使わなくても「線形回帰＋最尤」で解決することも多いので、そのような場合にまで無理して神経回路網を使う必要はありません。

高度な学習モデルを使いこなすことができる一方で目的に合わせて簡単な方法も選べるのが真の上級者。

最強の方法は、それに相応しい真に困難な問題に適用しましょう。

※ 難しい問題とは：変数間の関係を簡単な方程式で書けないもの。

軍師 募集中

諸葛孔明の
空城の計を
見破ることができる
データサイエン
ティスト求む。

データ解析の応用(1)

神経回路網に限らず、いろいろな学習システムがある。
サポートベクトルマシン、隠れマルコフモデル、混合正規分布、
ボルツマンマシン、ベイズネットワーク、・・・、深層学習。

様々なことに応用されている。
音声画像の処理・認識、ロボット制御、医療情報処理、
ひとのモデリング、自動運転、経済予測、消費者解析、・・・。

応用をする場合、実務の知識と経験が必要です。
どのモデル・方法が適切かは現実をよく考えて決めます。

データ解析のための基本ツールも知っておきましょう。

データ解析の応用(2)

学習理論・統計学・人工知能は、応用の世界から見たら専門的知識や洞察力がいない「道具」であることが望ましい。

しかし、実際は、学習することや推測することについての広い知識と深い理解が応用の場面でこそ必要になる。

(→ 成功・不成功は担当者の能力に強く依存する)

このため、学習理論・統計学・人工知能の応用を実務で行うには「ひと」が必要です。仕事名は、研究開発、データ分析、経営企画、ソフトウェア工学、いろいろです。

最近ではデータの次元や量が非常に大きくなり、広い領域で「データを解析するひと」が必要とされるようになってきました。