統計学ガイドツアー

2020.03.03

Yuma Uchiumi

（uchiumi@ailab.ics.keio.ac.jp）

発表の目的

Presenter

1. 統計学の魅力や可能性を伝える
2. 正しく有用な情報を伝える
3. 初心者から入門者への橋渡しをする

Audience

1. 発表を自分なりに楽しむ
2. 疑問や質問は積極的に聞いてみる

今日の流れ

#01 Lecture

1. なぜ、いま統計学を学ぶべきなのか？
2. 最先端のデータサイエンス事情

#02 Presentation

3. 統計学とは何か？学ぶ意義は？
4. 統計学を学ぶためのヒント

#01 Lecture

統計学の背景について

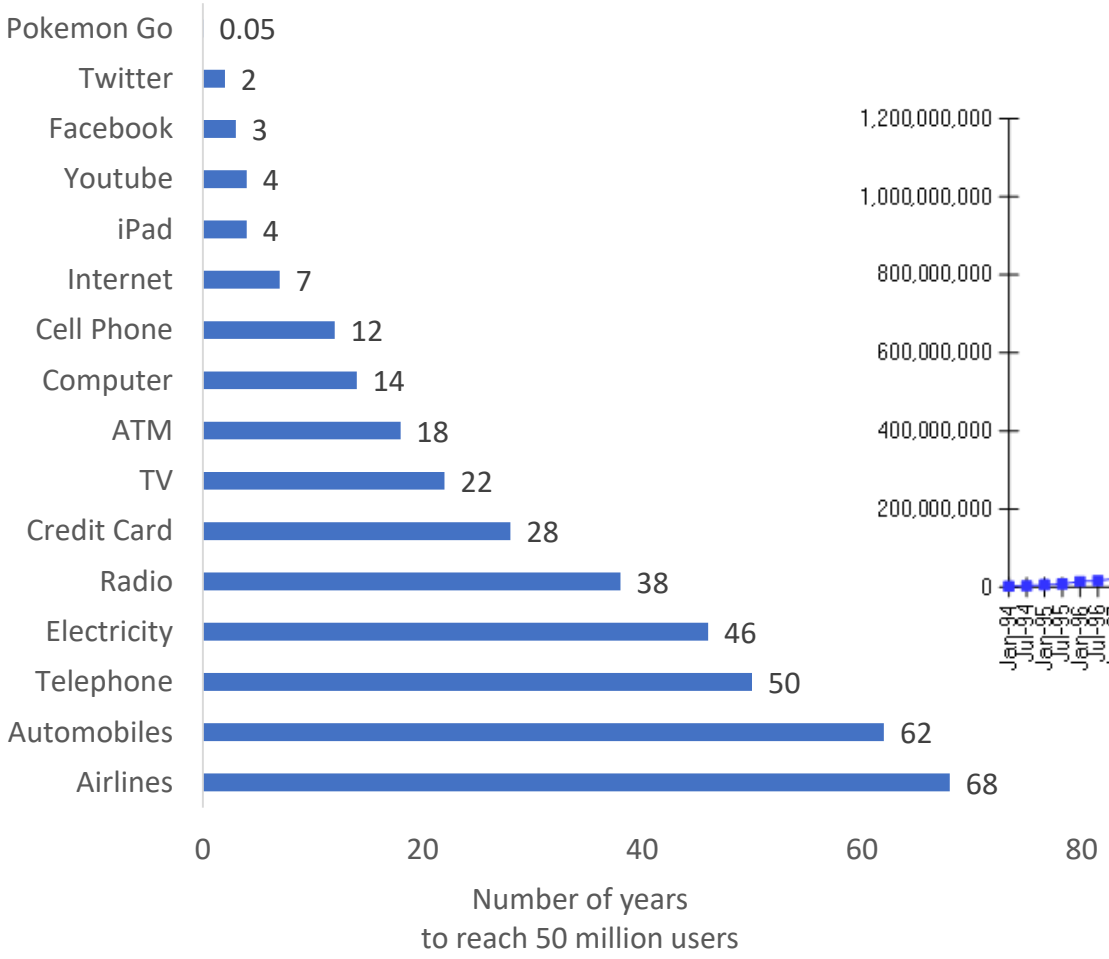
1. なぜ、いま統計学を学ぶべきなのか？
2. 最先端のデータサイエンス事情

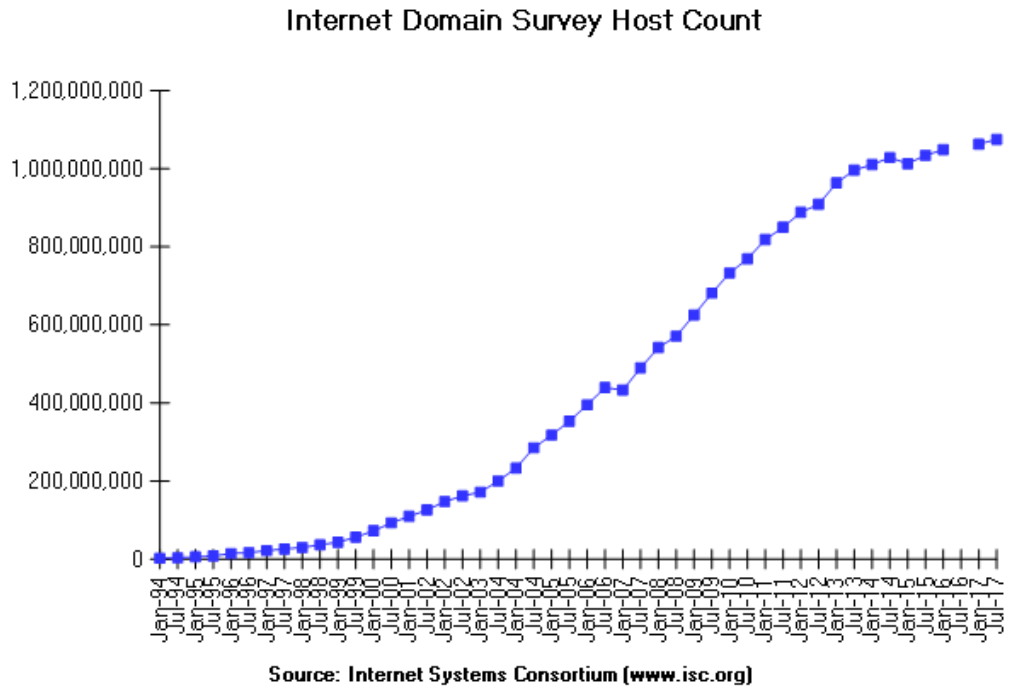
#01 Lecture

1. なぜ、いま統計学を学ぶべきなのか？

BigData/IoT：ヒトのあらゆる経済活動はオンラインへ

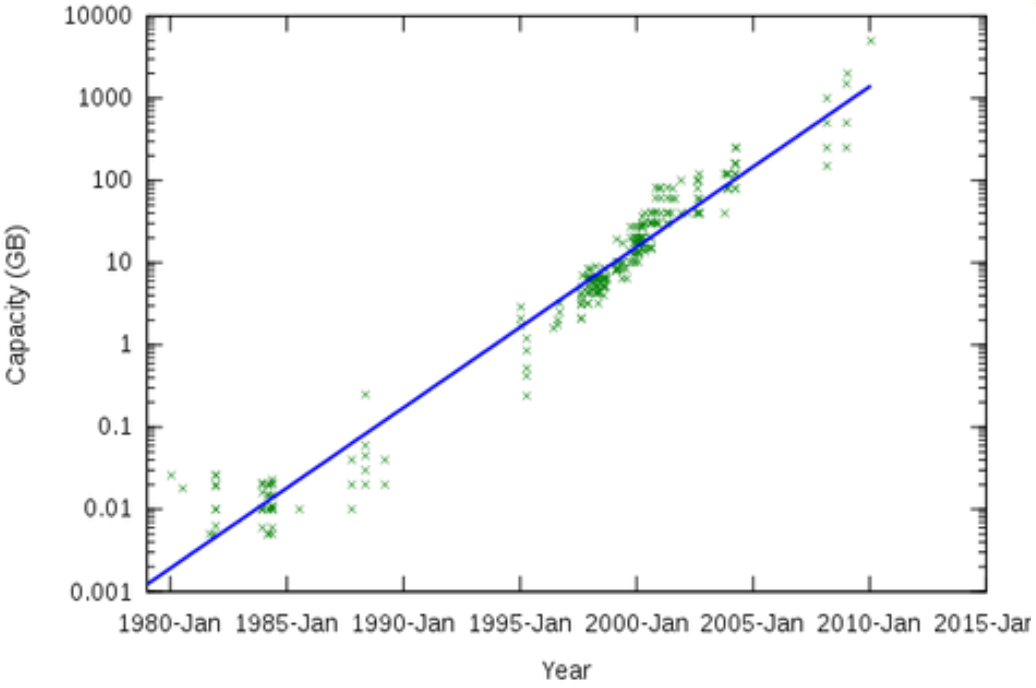
取引・広告・消費を含めたほぼすべての経済活動がオンラインで可能な社会へ

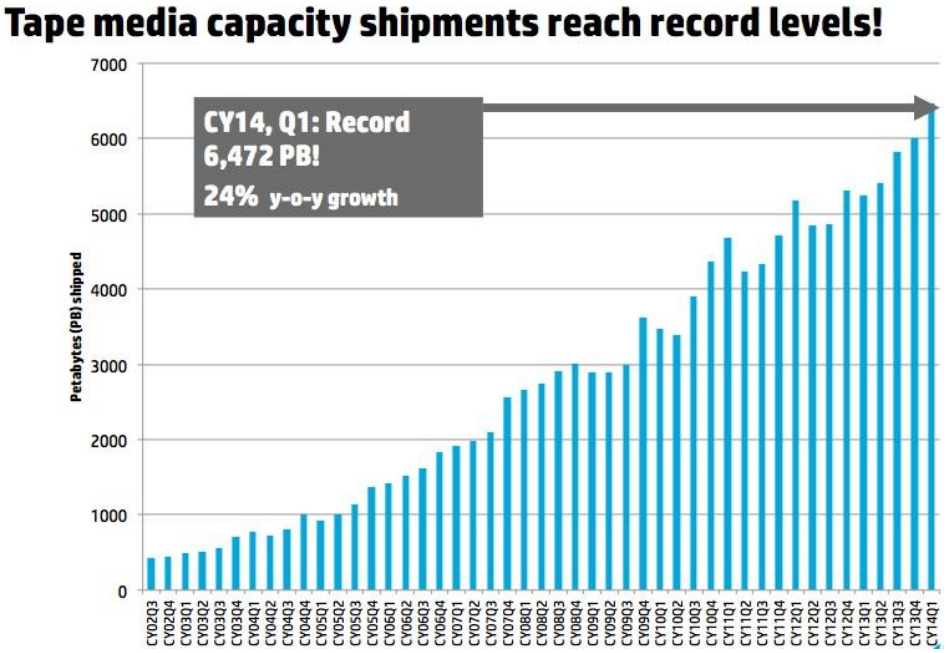




BigData/IoT：基盤技術の向上(Ex. Large Storage)

ITは，他の技術に比べて7倍速で成長することから“Dog Year”と例えられる．





source:
https://www.digitaltonto.com/2011/our-emergent-digital-future/

まとめ：情報過多な社会では、データ分析の価値があがる

さまざまな要因によって，統計学に明るい人材が求められている．

情報化社会

1. データの蓄積（Big-Data, Internet of Things）
2. 分析環境の発達（Computing Resource, Network Infrastructure）
3. 新しい手法（Open Source Software, Machine Learning, Deep Neural Network）

統計学は、情報化社会における「教養力」となる

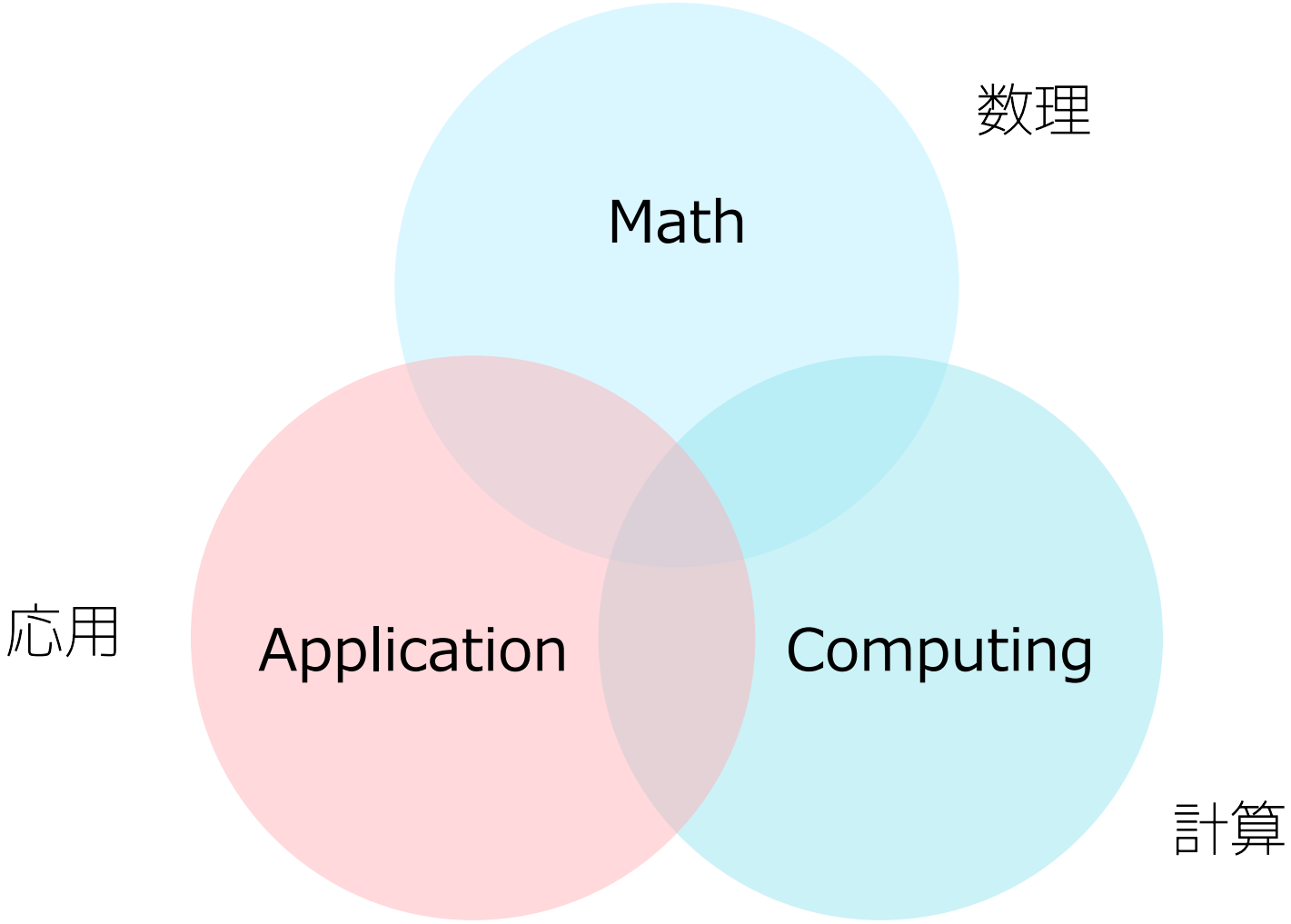
#01 Lecture

2. 最先端のデータサイエンス事情

#Introduction

データアナリティクス、データサイエンスとは何か？

現代的な分析手法には, 応用・数理・計算すべてに対する理解が必要.



#01 Lecture

データアナリティクス、データサイエンスとは何か？

現代的な分析手法には, 応用・数理・計算すべてに対する理解が必要.

Application	Math	Computing
ABテスト	解析学	決定理論
クラスタリング	線形代数	情報理論
時系列と傾向	複素関数論	離散数学
パネルデータ	ベクトル解析	アルゴリズム
モデルと実証	確率・測度論	可視化
実験計画法	数理統計	モンテカルロ法
計量分析	最適化	プログラミング
要因分析	線形計画法	データモデリング

「応用」:統計モデリングによる因果推定

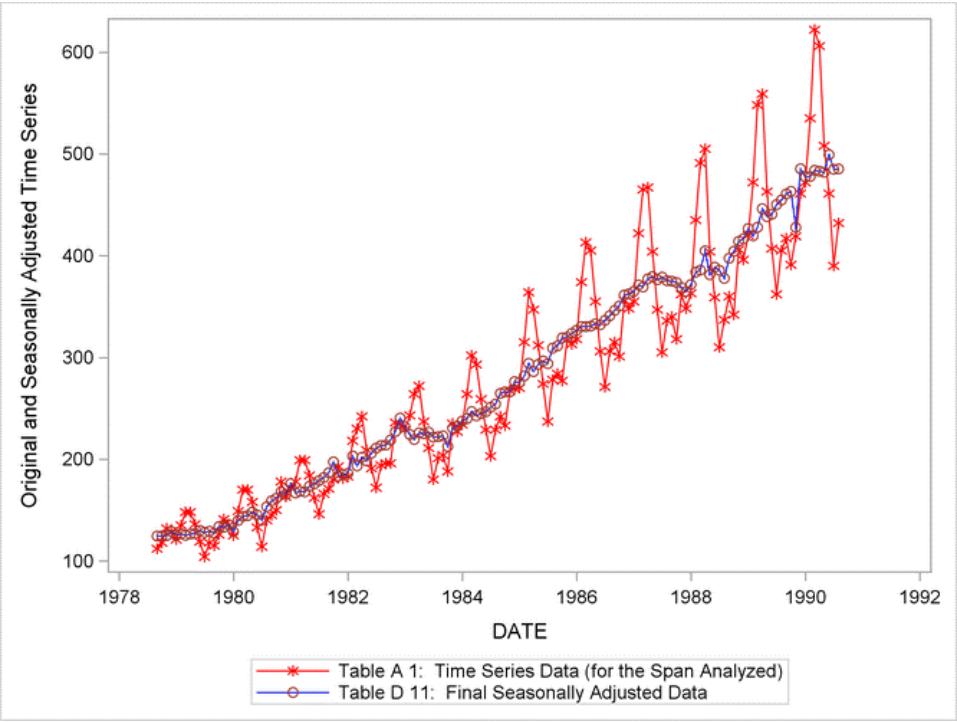
伝統的な統計学の応用分野として, 計量経済学があげられる.

パネルデータモデル

Panel data Models: Dependent variable is GDPGROWTH; t-value in parenthesis

Independent variables	Model 1 OLS	Model 2 FE-CS	Model 3 FE-CSW	Model 4 RE-CS	Model 5 RE-CS with AR(1)
BF	-0.000246 (-0.75209)	0.000144 (0.48256)	0.000175 (1.0774)	7.05×10^{-5} (0.2523)	0.000428 (1.29)
FC	-0.00083*** (-3.889)	-0.00097*** (-3.065)	-0.0007*** (-3.391)	-0.00099*** (-3.99)	-0.0007564** (-2.01)
FDI	0.000161 (0.271544)	-0.00269*** (-5.307)	8.02×10^{-5} (0.1259)	-0.0022*** (-4.549)	-0. 002473*** (-4.50)
FisF	0.000735*** (3.2262)	0.000954*** (3.5469)	0.00066*** (4.039)	0.00089*** (3.7892)	0. 0001892 (0.53)
FinF	-0.000205 (-0.9378)	0.000402 (1.55148)	0.000248 (1.3435)	0.000189 (0.871112)	0.0004986* (1.78)
GCF	0.001780*** (4.3539)	0.00252*** (5.5315)	0.00187*** (6.565)	0.00247*** (6.3008)	0. 001932*** (3.36)
IF	5.22×10^{-5} (0.203758)	-0.000278 (-1.2131)	-4.75×10^{-5} (-0.379)	-0.00024 (-1.09563)	-0.0002142 (-0.78)
NOD	-1.33×10^{-11} (-0.9867)	1.57×10^{-11} (0.9805)	5.12×10^{-12} (0.5837)	1.91×10^{-12} (0.145953)	-5.31×10^{-12} (-0.31)
D(NOD)	2.59×10^{-12} (0.30585)	-2.44×10^{-12} (-0.3273)	2.24×10^{-12} (0.7205)	1.49×10^{-12} (0.224740)	4.47×10^{-12} (0.36)
NOD×NOD	1.98×10^{-21} (0.339555)	-6.95×10^{-21} (-1.3086)	-5.26×10^{-21}* (-1.940)	-4.36×10^{-21} (-0.8948)	-1.53×10^{-21} (-0.29)
D(NOD) ×D(NOD)	-4.08×10^{-21} (-0.69879)	2.23×10^{-21} (0.4717)	1.93×10^{-21} (1.1714)	4.30×10^{-22} (0.094844)	-8.62×10^{-22} (-0.20)

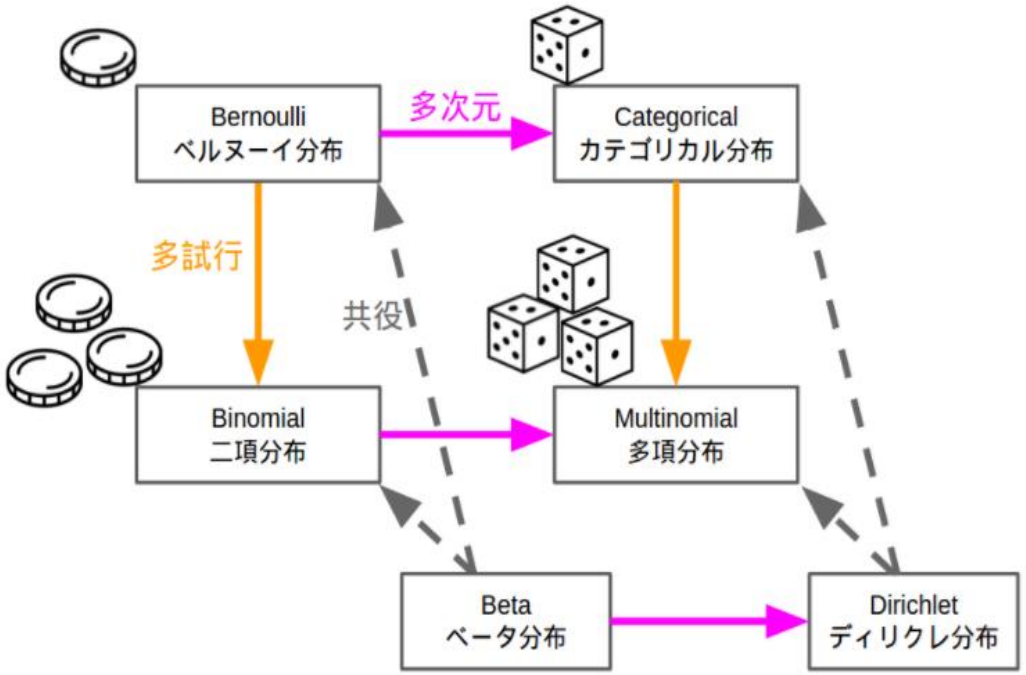
時系列過程の傾向分析



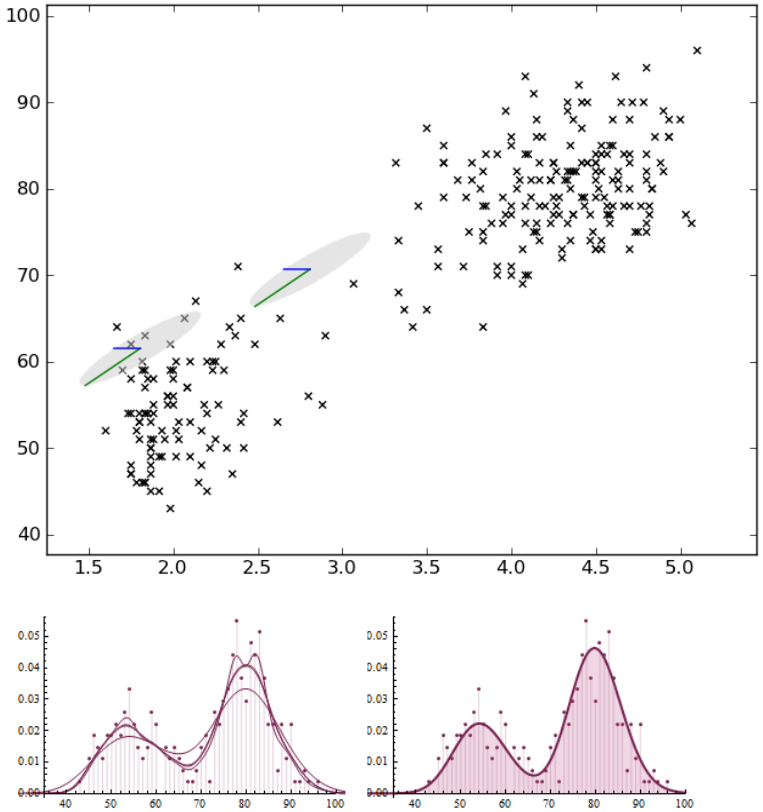
「数理」: 確率分布(密度)に関する性質の解明

数理統計学の例として, 確率分布にまつわる議論を紹介する.

指数型分布族(共役性, 多次元化)



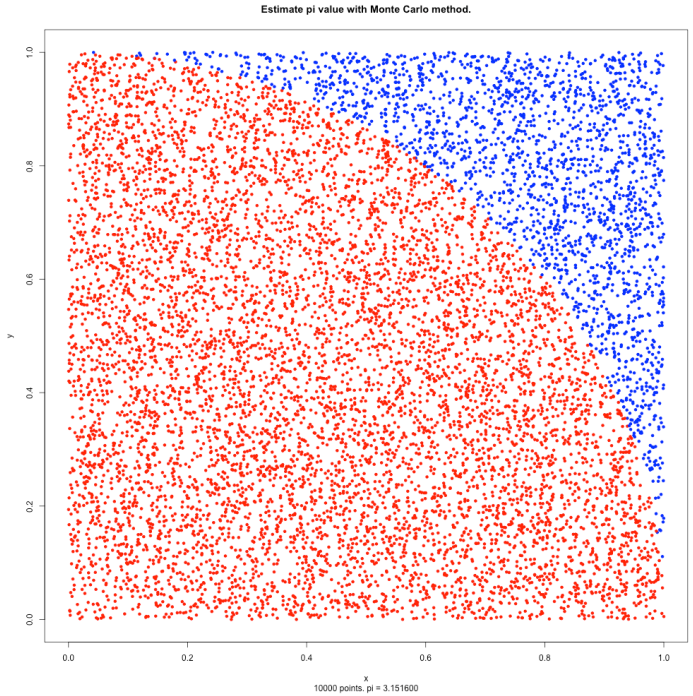
混合モデルと主軸変換



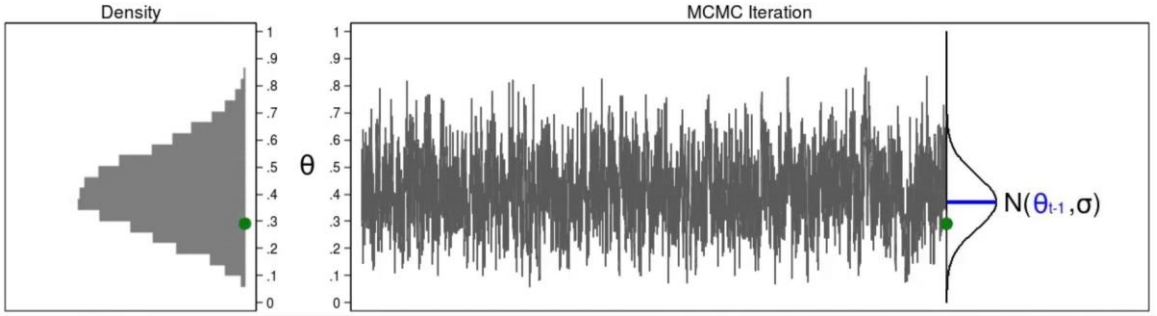
「計算」: 疑似乱数を用いた計算機シュミレーション

計算機をもちいた大規模な高速演算によって, 手法の見直しや再発見がおきている.

乱択アルゴリズム



マルコフ連鎖モンテカルロ法（MCMC）



Step 1: $r(\theta_{new}, \theta_{t-1}) = \frac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \frac{\text{Beta}(1,1,0.290) \times \text{Binomial}(10,4, 0.290)}{\text{Beta}(1,1,0.371) \times \text{Binomial}(10,4, 0.371)} = 0.773$

Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.773, 1\} = 0.773$

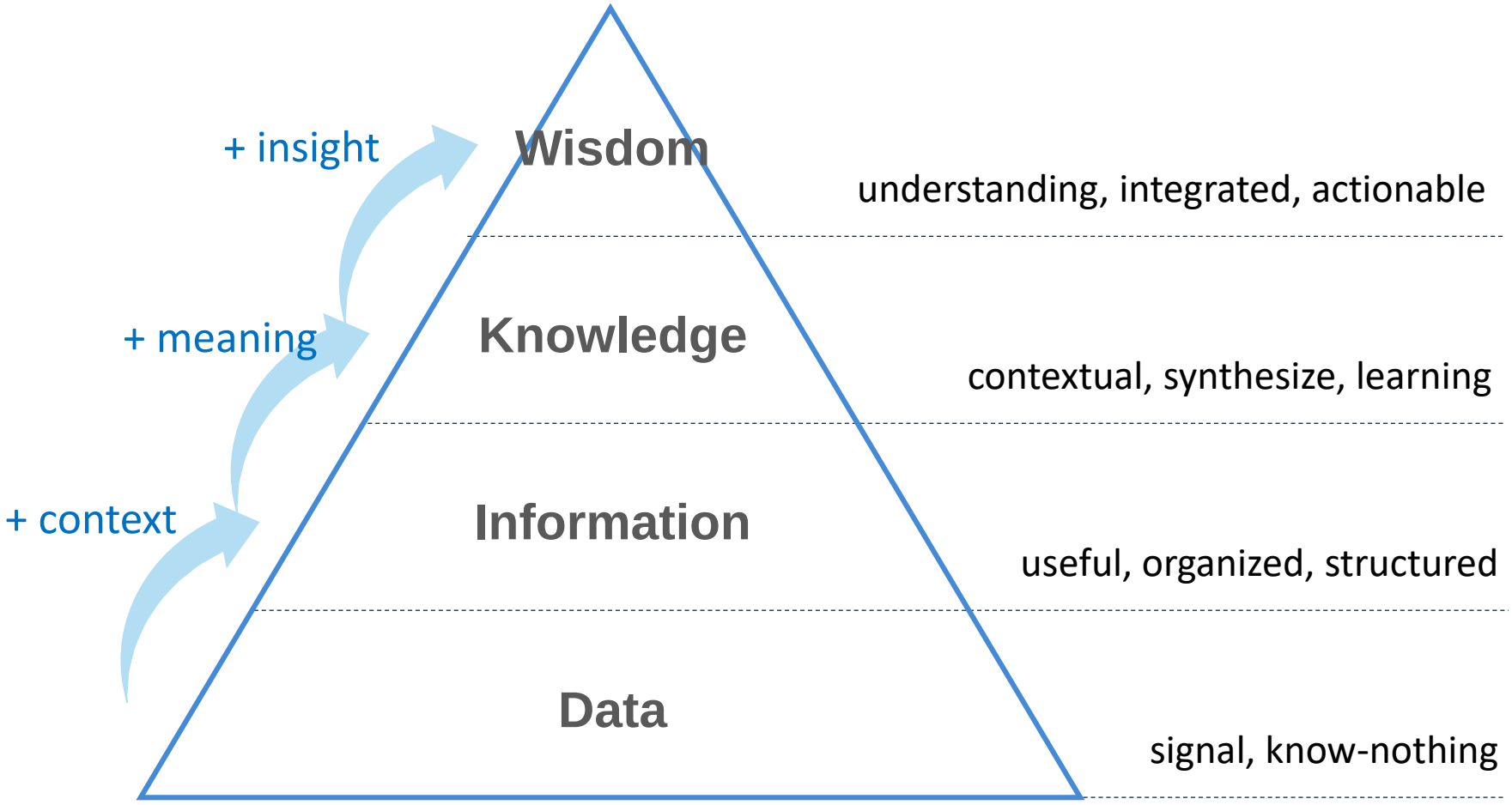
Step 3: Draw u ~ Uniform(0,1) = 0.420

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1})$ → If $0.420 < 0.773$ Then $\theta_t = \theta_{new} = 0.290$

Otherwise $\theta_t = \theta_{t-1} = 0.371$

DIKWモデル，意思決定の難しさ

人々の意思決定を変えない分析結果に価値はない！



#02 Presentation

統計学の概観をつかむ

3. 統計学とは何か？学ぶ意義は？
4. 統計学を学ぶためのヒント

#02 Presentation

統計学とは何か？学ぶ意義は？

統計学の成り立ち, 記述統計学と推測統計学

なぜ、統計学は有効なのか？学ぶ意義は？

学問としての関心以外にも，実用上たいへん役立つ能力が身につく．

1.「**推論**」: 正解がない問題に対して、できうる限り正しい答えを出せる

（Ex.）明日10時にあなたがYouTubeで聞いている音楽は何？

→不確かな状況について、定量的な「評価」や「予測」ができる

2.「**知識**」: データの解釈に対する”目利き力”が上がる

（Ex.）選挙調査のサンプル数として、信頼性がある範囲は？

→人は、１つの「データ」から複数の「図表」「解釈」「主張」を生み出せてしまう

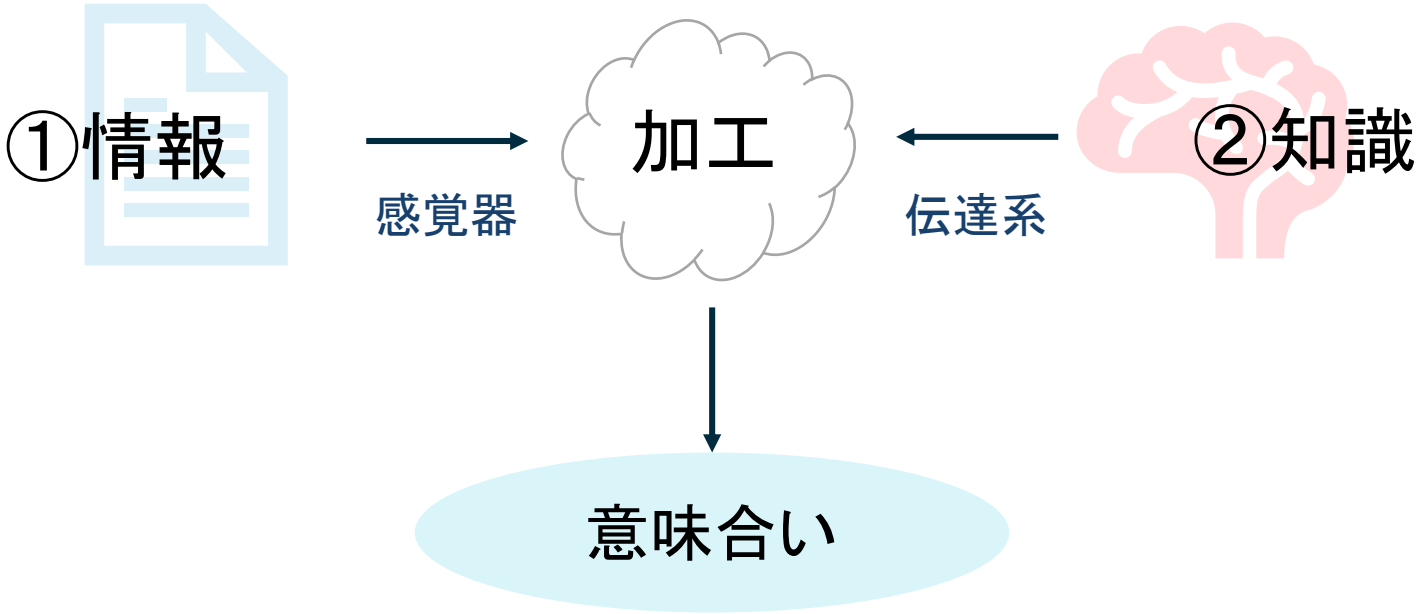
3.「**判断**」: ランダムネスと秩序を区別するクセがつく

（Ex.）オリンピック開催国で物価が上がることは偶然か？必然か？

→情報過多な社会では、「本質的な情報」と「それ以外」を判別すべき

余談:そもそも思考とは？

→「思考対象について何らかの**意味合い**を得るため、**情報**と**知識**を加工すること」



情報と知識 → 思考のための必要条件

統計学 → 思考のための十分条件

統計とは何か？統計学とは何か？

統計や統計学は幅広い目的・用途があるが，大まかな定義は以下のようになる．

統計：

・近代統計学の父，カール・ピアソン（1857-1936）

“The Grammar of Science”：「統計とは、科学の文法である.」

→「現象を，（何らかの調査によって）数量で把握すること」

統計学：

・インドの統計学者，ラオ,C.R.（1920-）

「不確実性を数量化することで、自然や社会にあふれる偶然に立ち向かう科学」

→「統計の作り方、それによる判断・推論の方法を研究する学問」

統計学はどのように生まれたのか？

17Cまでに、多くの分野で独自の統計理論が誕生．18C以降これらが統合される．

A） ゲームのテーブルから始まった確率論

B） 常備軍や国家財政上の必要から起こった国家状態の統計

C） 古代地中海貿易における海上保険の計算

D） 17世紀のペストを機とする死亡率の研究

E） 天文観測で生じる観測誤差の理論

F） 生物・生態学における諸量の相関関係の理論

G） 農地で実験を計画するための方法論

H） 経済学や気象学における時系列の理論

I） 心理学における要因分析やランキングの理論

J） 社会学におけるχ2乗統計量の方法

近代統計学を築いた10人

18C以降、近代科学の発展とともに統計理論が統一され「統計学」が誕生した.

1. ペティ (William Petty, 1627-1687) 社会経済現象の数量的観察
2. アッヘンヴァル (Gottfried Achenwall, 1719-1772) 国勢学派，統計調査
3. ラプラス (Simon Laplace, 1749-1827) 古典確率論の大勢， 近代確率論の基礎
4. ガウス (Carl Friedrich Gauss, 1777-1855) 誤差理論と正規分布，最小2乗法
5. ケトレー (Adolphe Quetelet, 1796-1874) 大量観察と統計的法則性，平均の概念
6. ゴルトン (Francis Galton, 1822-1911) 遺伝学の数理的理論，回帰の導入
7. カール・ピアソン (Karl Pearson, 1851-1936) 近代統計学の数理的基礎，母集団，相関係数
8. ゴセット (William Gosset, 1876-1937) t分布の導入，小標本理論
9. フィッシャー (Ronald Fisher, 1890-1962) 統計的推測理論，標本分布論，実験計画法，F分布
10. ワルド (Abraham Wald, 1902-1950) 統計的決定理論，検定理論と推定理論の統一

記述統計学と推測統計学

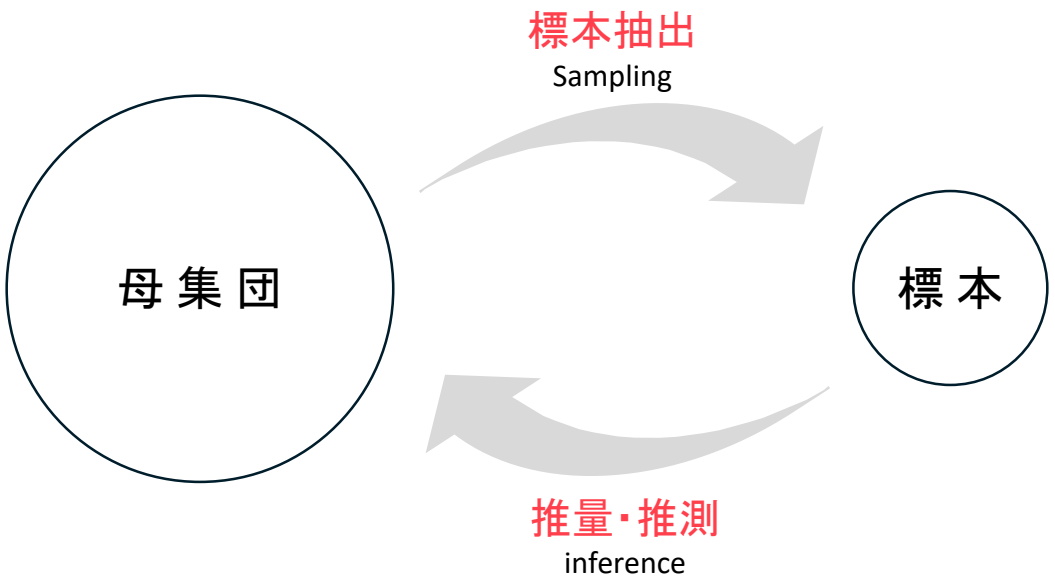
記述統計は、大数の法則と中心極限定理に支えられて推測統計学へ拡張される。

- 記述統計学・・・与えられたデータをすべて観察し、整理・要約する方法

  全部(=母集団)を丹念に調べ、規則性から法則を見出す

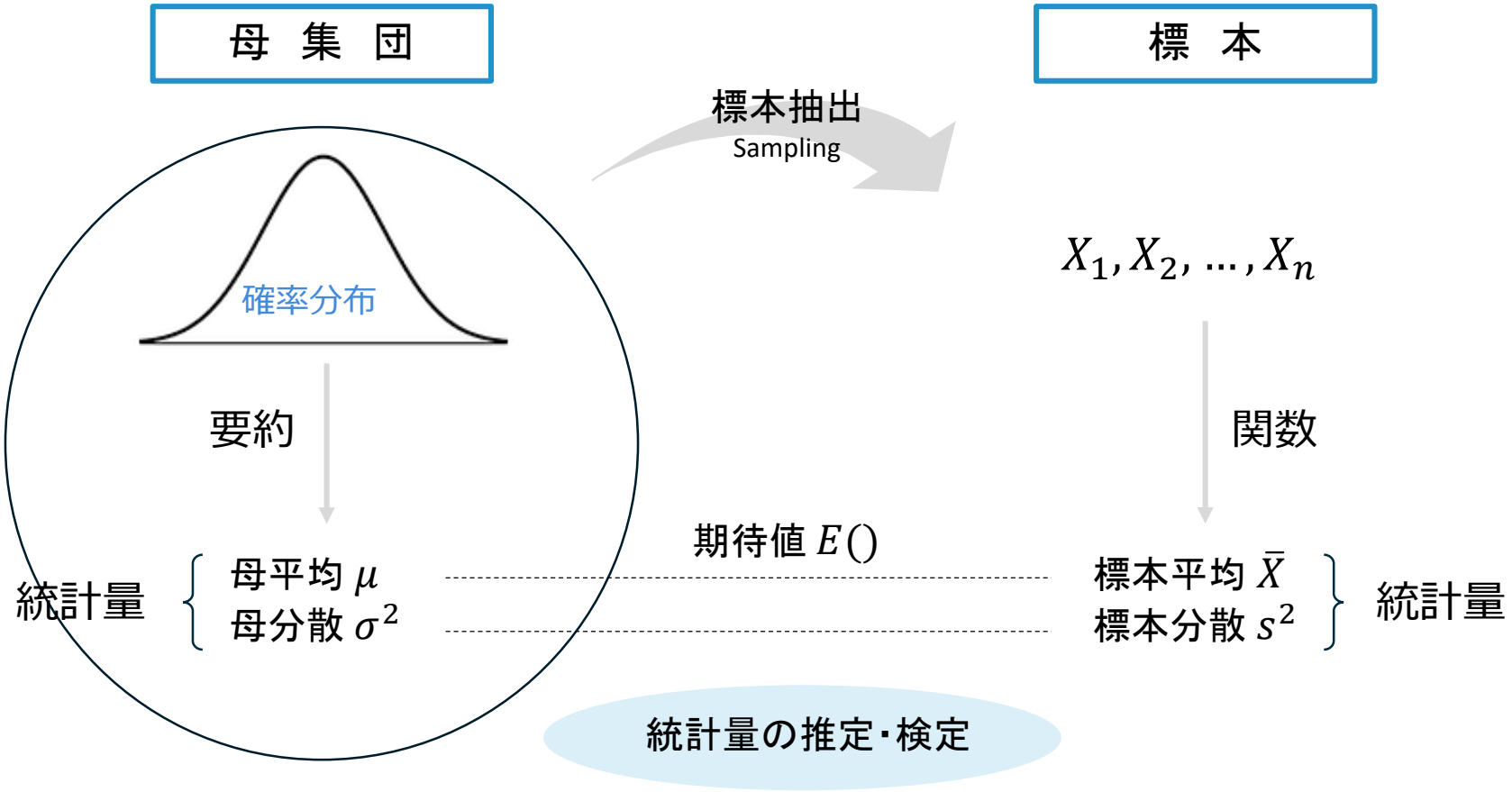
- 推測統計学・・・与えられたデータから全体について推量・推測する方法

  一部(=標本)を観察し、論理性のある推測で全体の法則性を発見する



推測統計の全体図

「標本から母集団を推測できる」→不確かな現象に対する定量的アプローチ



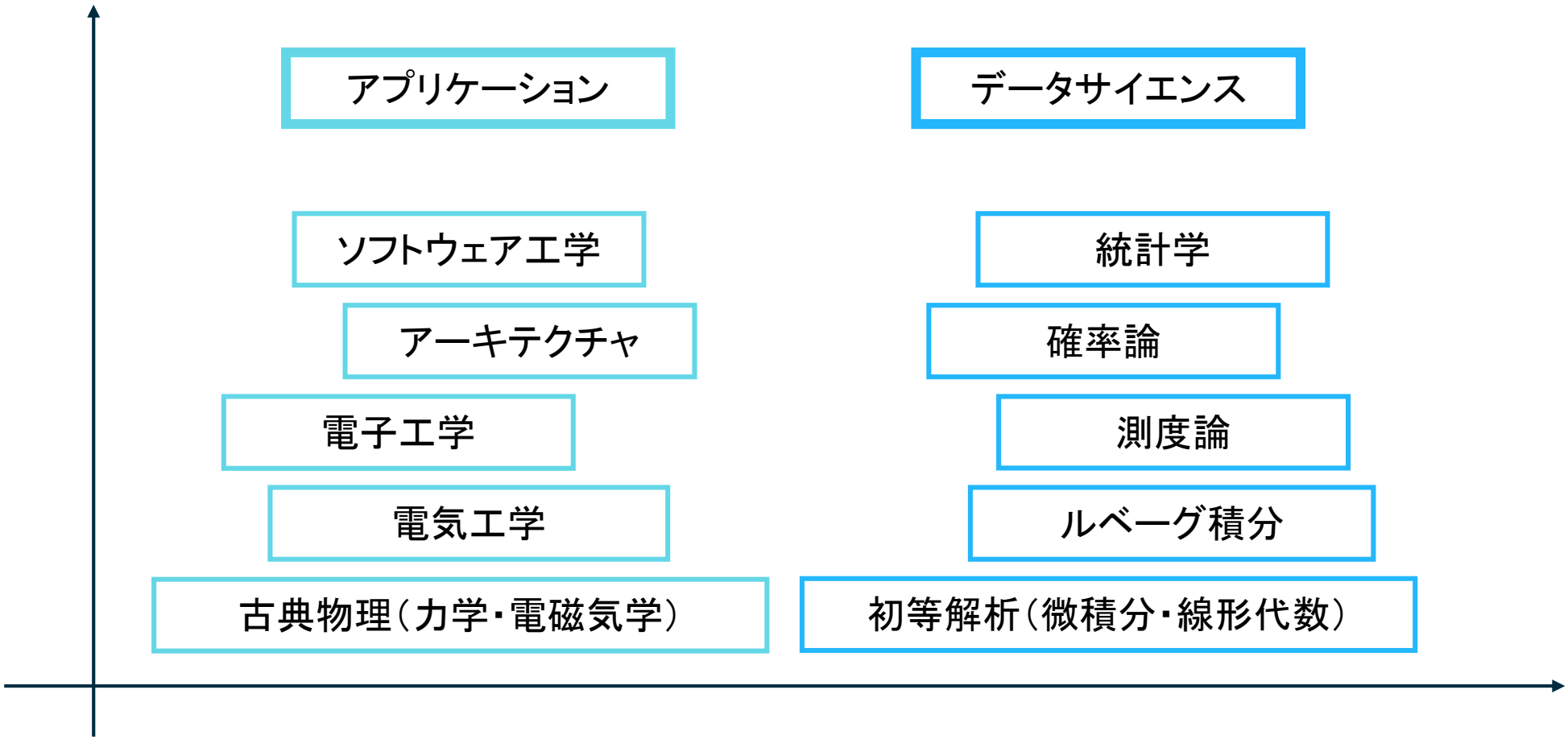
#02 Presentation

統計学を学ぶためのヒント

数学を学ぼう！＋確率論入門

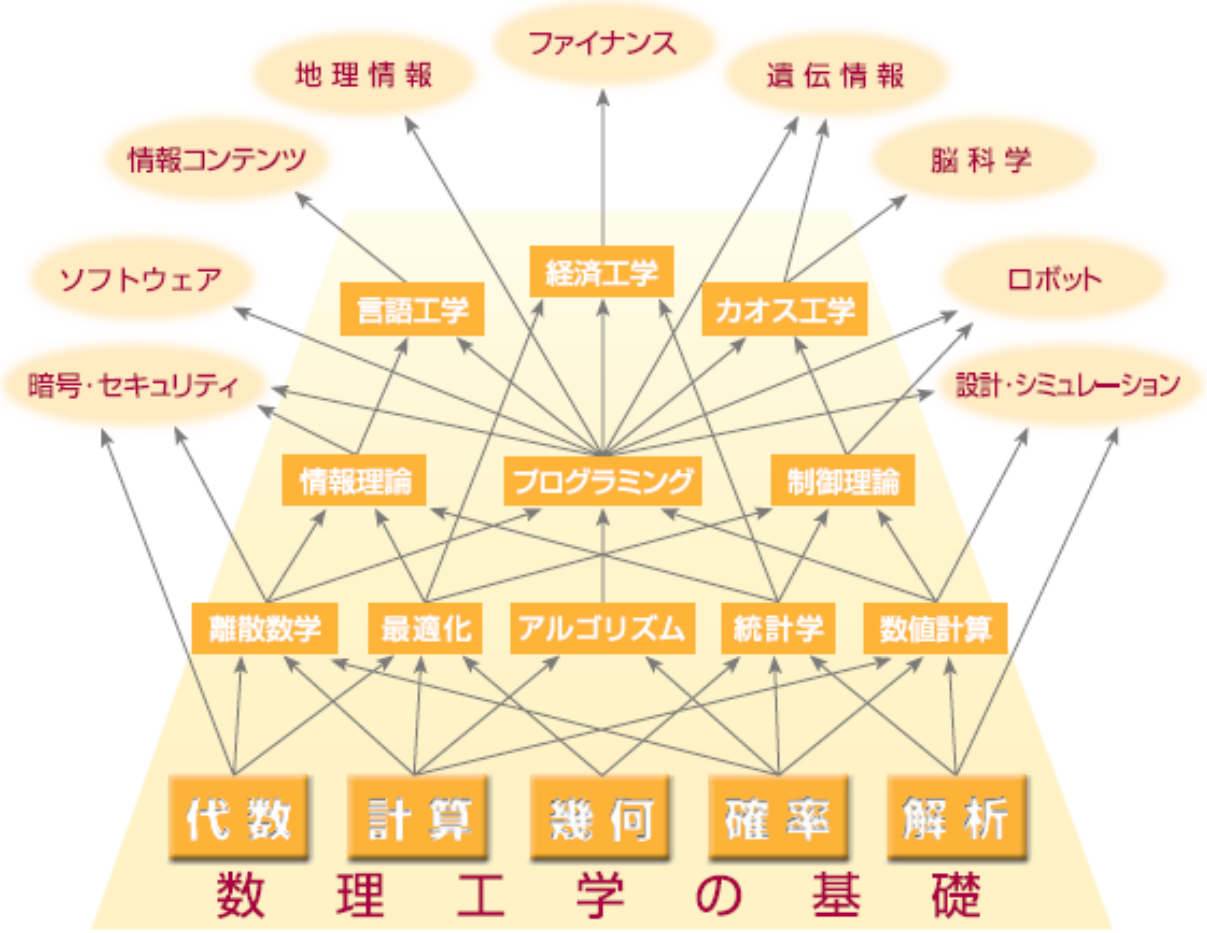
どんな分野でも、まず”レイヤー”を意識せよ

何を始めるにせよ, 基礎が肝心.



どんな分野でも、源流には数学がある！

関連するすべての数学を理解する必要はないが, ”位置情報”は意識するべき.



余談：数学を勉強するなら、その作法を知ることが大切

①→②→③という順序ではなく，応用分野から理論を学ぶことも大切．

高校の数学

→目に見える現実と結びついた議論。具体的。
（実数中心、図形による証明が多い）

大学の数学

→目に見えない理論が中心。抽象的。

数学の作法を知らないと，**難しい**と感じる！

〈数学の作法〉

①まずは数（元）と演算を定義。

②つぎに定理を証明しながら論理体系を固める。

③実は、その体系に対応する応用分野がある。

余談：測度論から確率論へ

確率はΩ・$\mathcal{B}$・Pの3要素によって決まる．$(\Omega, \mathcal{B}, P)$の組み合わせを"確率空間"という．

①標本空間Ω

②可測集合族$\mathcal{B}$の定義

(M1) $\emptyset \in \mathcal{B},\ \Omega \in \mathcal{B}$

(M2) $A \in \mathcal{B} \Rightarrow \bar{A} \in \mathcal{B}$

(M3) $A_k \in \mathcal{B}, k = 1,2, \ldots \Rightarrow (\bigcup_{k=1}^{\infty} A_k) \in \mathcal{B}$

測度空間 $(\Omega, \mathcal{B})$

③確率測度Pの定義

(P1) すべての$A \in \mathcal{B}$に対して$P(A) \geq 0$

(P2) $P(\Omega) = 1$

(P3) $A_i \cap A_j = \emptyset\ (i \neq j) \Rightarrow P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$

確率空間 $(\Omega, \mathcal{B}, P)$

余談：公理からはじまり、定理が導かれる

確率が絡む現象ならば，その”元”と”演算”をつねに意識する．

“元”

＋

“演算”

↓

“定理”

〇確率Pの定義

(P1) すべての$A \in \mathcal{B}$に対して$P(A) \geq 0$

(P2) $P(\Omega) = 1$

(P3) $A_i \cap A_j = \emptyset \ (i \neq j) \ \Rightarrow \ P(\bigcup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} P(A_k)$

〇確率Pの基本法則

(加法定理) $P(A) = \sum_B P(A, B)$

(乗法定理) $P(A, B) = P(A \mid B)P(B)$

〇確率Pをつかったさまざまな定理

（ベイズの定理）$P(A \mid B) = \frac{P(B \mid A)P(A)}{\sum_A P(A, B)}$ etc.

公理

定理

Fin

#03 Supplement

3. 補足資料

退屈だと感じたひと向け

#03 Supplement

機械学習(教師あり学習)の基本パーツは期待値

ニューラルネットワークも同様です

- 期待値の式

$$E[f] = \int f(x) \cdot p(x) dx$$

期待値　　重み　　確率

- 教師あり学習（目標:期待損失を最小化すること） $D_{train} = \{x, y\}$ とおく

$$E[L] = \iint L(y, f(x)) \cdot p(x, y) dxdy$$

期待損失　　重み(=損失関数)　　同時確率

「期待値」の概念はとっても大切！！

#03 Supplement

おすすめ図書・文献（入門～初級）

Kindleなら最初の20ページは無料で読めます． URL: Amazonの検索ページ

- 「数学ガールの秘密ノート/やさしい統計」. 結城浩. SBクリエイティブ
- 「データ分析の力−因果関係に迫る思考法」.伊藤公一朗.光文社新書
- 「統計学が最強の学問である」.西内啓.ダイヤモンド社
- 「確率思考の戦略論−USJでも実証された数学マーケティングの力」.森岡毅.角川書店
- 「原因と結果の経済学−データから真実を見抜く思考法」. 中室牧子,津川友介.ダイヤモンド社
- 「異端の統計学 ベイズ」. Sharon Bertsch McGrayne.草思社
- 「完全独習 統計学入門」.小島寛之.ダイヤモンド社
- 「基本統計学(第4版)」. 宮川公男. 有斐閣
- 「統計学入門(基礎統計学Ⅰ)」. 東京大学教養学部統計学教室. 東京大学出版

#03 Supplement

おすすめ図書・文献（中級～上級）

本格的に確率/統計を学びたい人に役立つ内容です．いきなり読むと挫折します．

- 「回帰分析」. 佐藤隆光. 朝倉書店
- 「データ解析のための統計モデリング入門–GLM,階層ベイズ,MCMC」. 久保拓哉. 岩波書店
- 「Pythonによるデータ分析入門」. Wes McKinney. オライリージャパン社
- 「経済・ファイナンスデータの計量時系列分析」. 沖本竜義. 朝倉書店
- 「Econometrics」. Fumio Hayashi. Princeton University Press
- 「Introductory Econometrics: A Modern Approach」. Jeffrey M. Wooldridge. Cengage Learning
- 「数理統計学の基礎」. 久保川達也. 共立出版
- 「パターン認識と機械学習–ベイズ理論による統計的予測」. C. M. Bishop. 丸善出版

Thank you for your attention.

Best regards, Yuma Uchiumi.