「見えないデータ」を推定する

持橋大地
統計数理研究所
daichi@ism.ac.jp

小石川中等教育学校スーパーサイエンス・ハイスクール
2022-9-7 (水)

自己紹介

- 持橋大地
  情報・システム研究機構 統計数理研究所 准教授
  専門：計算言語学、自然言語処理、人工知能

統計学の
国立研究所
(立川市)

- 総合研究大学院大学 統計科学専攻
  - 大学院から入学することが可能

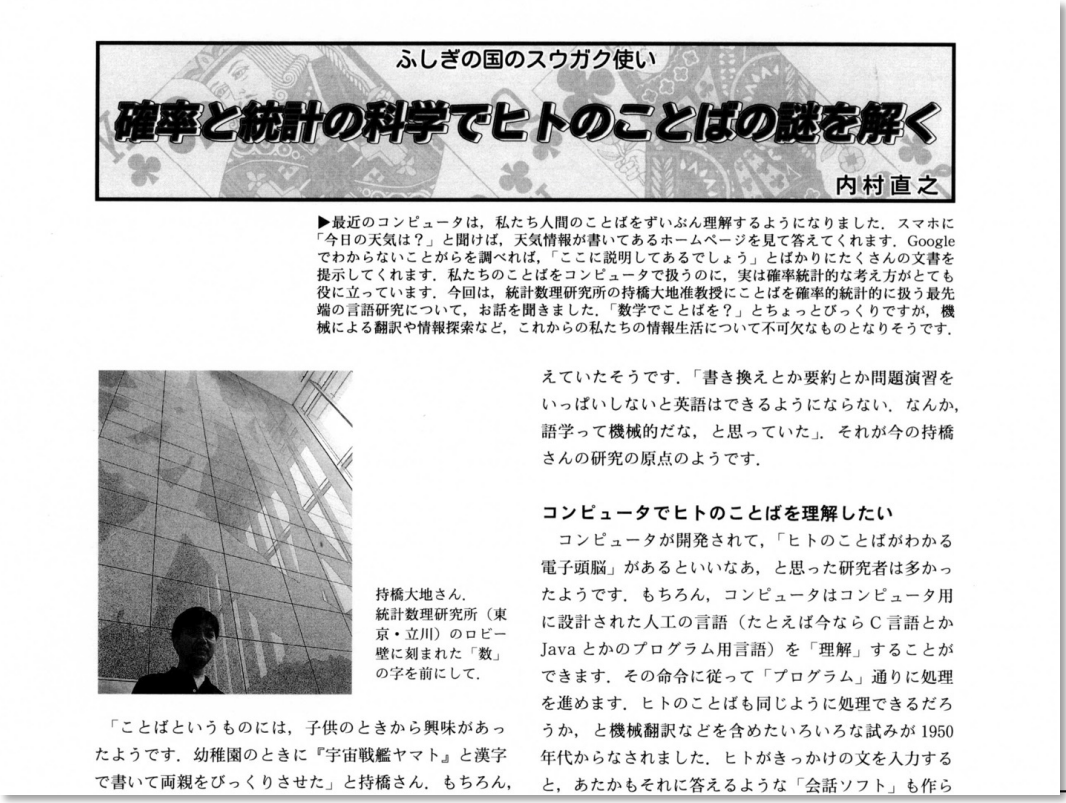
経歴

NAISTの様子

- 小石川高校卒 44期
  - 音楽研究会でした (指揮者)
- これまで：
  - 東大文III入学 (文学部進学コース)
  - 教養学部 基礎科学科第二に進学 (理転、文科から2名)
  - 東大の院を辞退して、奈良先端科学技術大学院大学 (NAIST) に進学
  - ATR 音声言語研究所
  - NTT コミュニケーション科学基礎研究所
  - 統計数理研究所 (2011年〜)

小石川高校時代 (1)

- 予備校等は行かず、Z会と月刊『大学への数学』などで勉強
- 『大学への数学』2014年9月号
  “ふしぎの国のスウガク使い”で紹介していただきました



ふしぎの国のスウガク使い

確率と統計の科学でヒトのことばの謎を解く

内村直之

▶最近のコンピュータは，私たち人間のことばをずいぶん理解するようになりました．スマホに「今日の天気は？」と聞けば，天気情報が書いてあるホームページを見て答えてくれます．Google でわからないことがらを調べれば，「ここに説明してあるでしょう」とばかりにたくさんの文書を提示してくれます．私たちのことばをコンピュータで扱うのに，実は確率統計的な考え方がとても役に立っています．今回は，統計数理研究所の持橋大地准教授にことばを確率的統計的に扱う最先端の言語研究について，お話を聞きました．「数学でことばを？」とちょっとびっくりですが，機械による翻訳や情報探索など，これからの私たちの情報生活について不可欠なものとなりそうです．

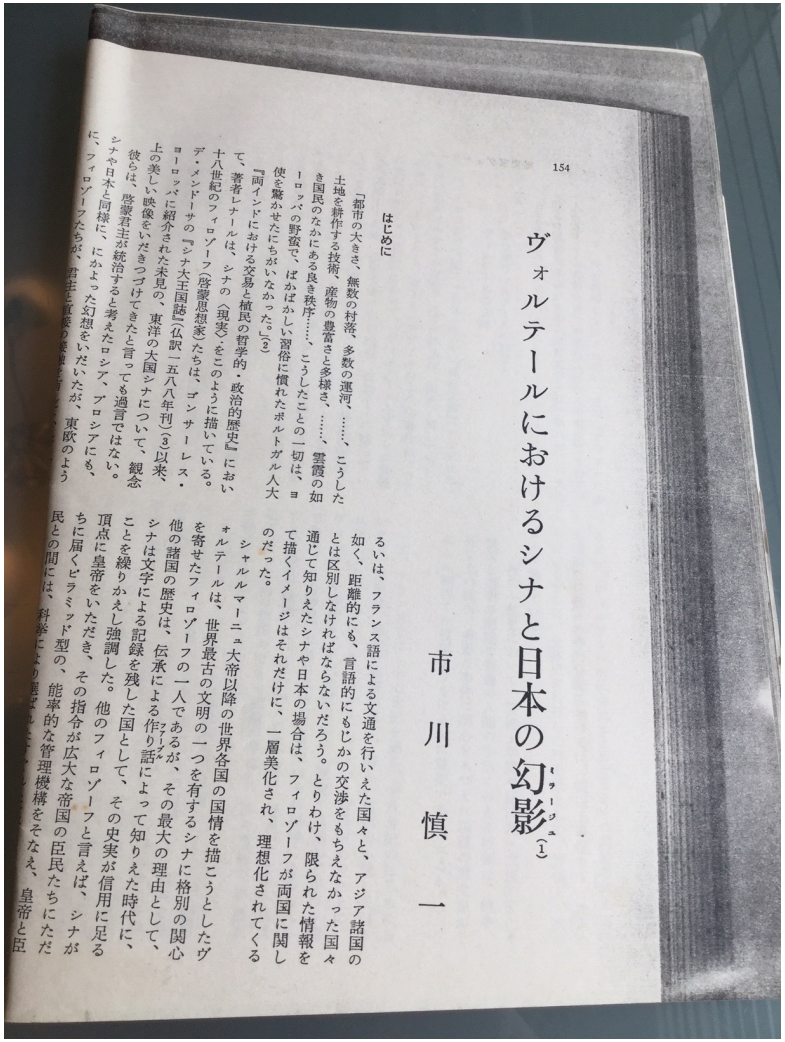
「ことばというものには，子供のときから興味があったようです．幼稚園のときに『宇宙戦艦ヤマト』と漢字で書いて両親をびっくりさせた」と持橋さん．もちろん，

えていたそうです．「書き換えとか要約とか問題演習をいっぱいしないと英語はできるようにならない．なんか，語学って機械的だな，と思っていた」．それが今の持橋さんの研究の原点のようです．

コンピュータでヒトのことばを理解したい

コンピュータが開発されて，「ヒトのことばがわかる電子頭脳」があるといいなあ，と思った研究者は多かったようです．もちろん，コンピュータはコンピュータ用に設計された人工の言語（たとえば今ならＣ言語とか Java とかのプログラム用言語）を「理解」することができます．その命令に従って「プログラム」通りに処理を進めます．ヒトのことばも同じように処理できるだろうか，と機械翻訳などを含めたいろいろな試みが 1950 年代からなされました．ヒトがきっかけの文を入力すると，あたかもそれに答えるような「会話ソフト」も作ら

小石川高校時代 (2)



154

ヴォルテールにおけるシナと日本の幻影(ファンタスム)(1)

市川慎一

はじめに

「都市の大きさ、無数の村落、多数の運河、……、こうした土地を耕作する技術、産物の豊富さと多様さ、……、雲霞の如き国民のなかにある良き秩序……、こうしたことの一切は、ヨーロッパの野蛮で、ばかばかしい習俗に慣れたポルトガル人大使を驚かせたにちがいなかった。(2)

『両インドにおける交易と植民の哲学的・政治的歴史』において、著者レナールは、シナの〈現実〉をこのように描いている。十八世紀のフィロゾーフ(啓蒙思想家)たちは、ゴンサーレス・デ・メンドーサの『シナ大王国誌』(公訳一五八八年刊)(3)以来、ヨーロッパに紹介された未見の、東洋の大国シナについて、親念上の美しい映像をいだきつづけてきたと言っても過言ではない。彼らは、啓蒙君主が統治するシナや日本と同様に、にかよった幻想をいだいたが、東欧のようるいは、フランス語による文通を行いえた国々と、アジア諸国の如く、距離的にも、言語的にもじかの交渉をもちえなかった国々とは区別しなければならないだろう。とりわけ、限られた情報を通じて知りえたシナや日本の場合は、フィロゾーフが両国に関して描くイメージはそれだけに、一層美化され、理想化されてくるのだった。

シャルルマーニュ大帝以降の世界各国の国情を描こうとしたヴォルテールは、世界最古の文明の一つを有するシナに格別の関心を寄せたフィロゾーフの一人であるが、その最大の理由として、他の諸国の歴史は、伝承による作り話によって知りえた時代に、シナは文字による記録を残した国として、その史実が信用に足ることを繰りかえし強調した。他のフィロゾーフと言えば、シナが頂点に皇帝をいただき、その指令が広大な帝国の臣民たちにただちに届くピラミッド型の、能率的な管理機構をそなえ、皇帝と臣民との間には、科挙により選ばれた

- 高3で世界史を勉強している際、フランス啓蒙思想のヴォルテールが中国に興味を持った、という話を教科書で読み、世界史科を訪ねると、この「ヴォルテールにおけるシナと日本の幻影」をいただいた
- 高3の5月くらいに、小石川の藤棚の下で読みました

小石川高校時代 (3)

- 地学を勉強しているとき、深海底にあるというマンガン団塊の成因に興味を持ち、地学科を訪ねた
- 英語の専門書を借してもらい、受験勉強の傍に読んで、どうしてマンガン団塊ができるのかを理解
  - precipitation=沈殿+αによって発生

海底の
マンガン団塊

今日の話

「見えないデータ」を推定する

- 文系の人にも関係のある話です

アンケート分析

質問	Aさん	Bさん	C さん	Dさん
1. 東京は暮らしやすい?	はい	いいえ	はい	はい
2. 東京の物価は高い?	はい	いいえ		いいえ
3. 東京は安全?	はい	はい	いいえ	いいえ
4. 東京の交通は便利?	はい	いいえ	はい	はい
:				
10. 東京は好きな都市?	はい	いいえ	はい	はい

- 皆さんの多くの研究で、アンケート結果の分析が必要
- 簡単のため、はい/いいえ (はい=1, いいえ=0) の場合を考える
- ナイーブに行うと、色々な問題がある

アンケート分析 (2)

質問	Aさん	Bさん	Cさん	Dさん
1. 東京は暮らしやすい?	1	0	1	1
2. 東京の物価は高い?	1	0	—	0
3. 東京は安全?	1	1	0	0
4. 東京の交通は便利?	1	0	1	1
:				
10. 東京は好きな都市?	1	0	1	1

- 何でも1を答える人や、ほとんど0を答える人がいる
  → 同じ1や0でも、人によって重要性が異なる
- 答えていない場合(欠損値といいます)がある場合がある
  → 単純に平均してもよいか?

テストの場合で考えてみる

	Aさん	Bさん	Cさん	Dさん
問題1	1	0	1	0
問題2	1	1	0	0
問題3	1	0	—	1
問題4	0	0	0	0
：				
問題20	1	0	1	—

- Aさんは、ほとんどの問題に正解
- Bさんは多くの問題に不正解だが、他の人ができない問題2には正解した
- Cさんは問題3が未回答、Dさんは後半は早退して欠席

テストの場合で考えてみる (2)

	Aさん	Bさん	Cさん	Dさん
問題1	1	0	1	0
問題2	1	1	1	1
問題3	1	0	—	1
問題4	0	0	0	0
:				
問題20	1	0	1	—

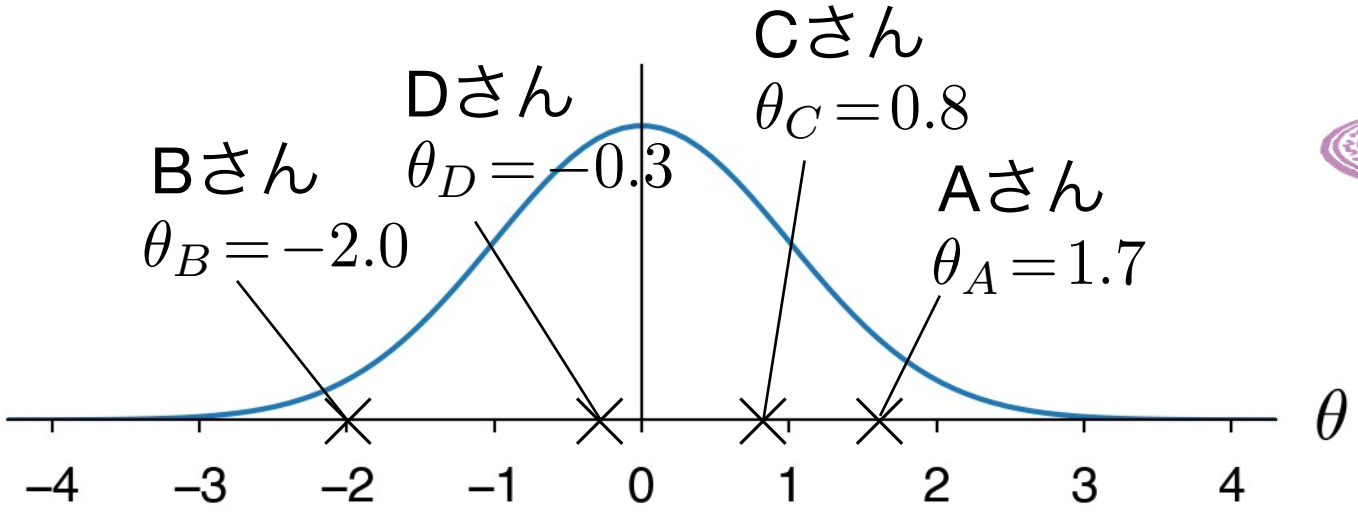
→ 易しい問題

→ 難しい問題

↓ 能力高　↓ 能力低

- 人には、見えない能力があるのでは...?
- 問題には、見えない難しさがあるのでは...?

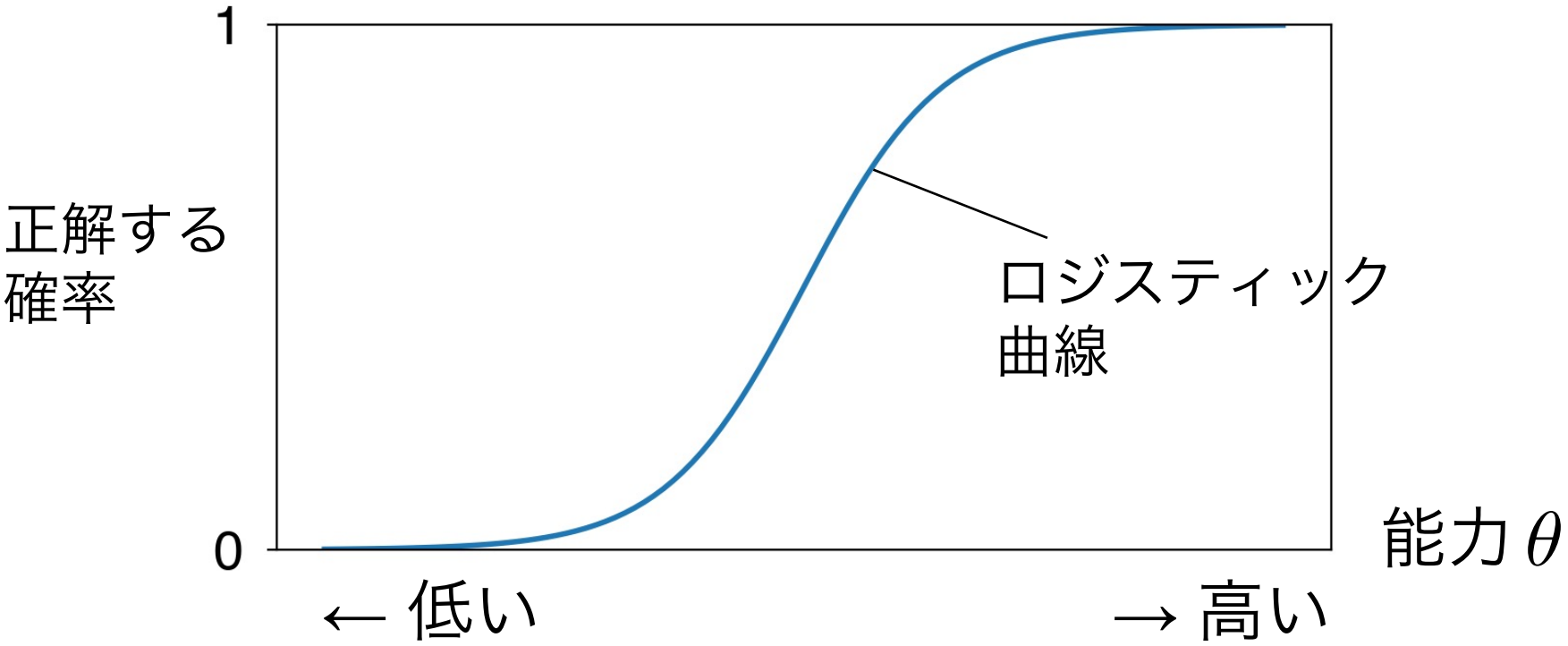
「能力」のモデル化



こういう能力ではない

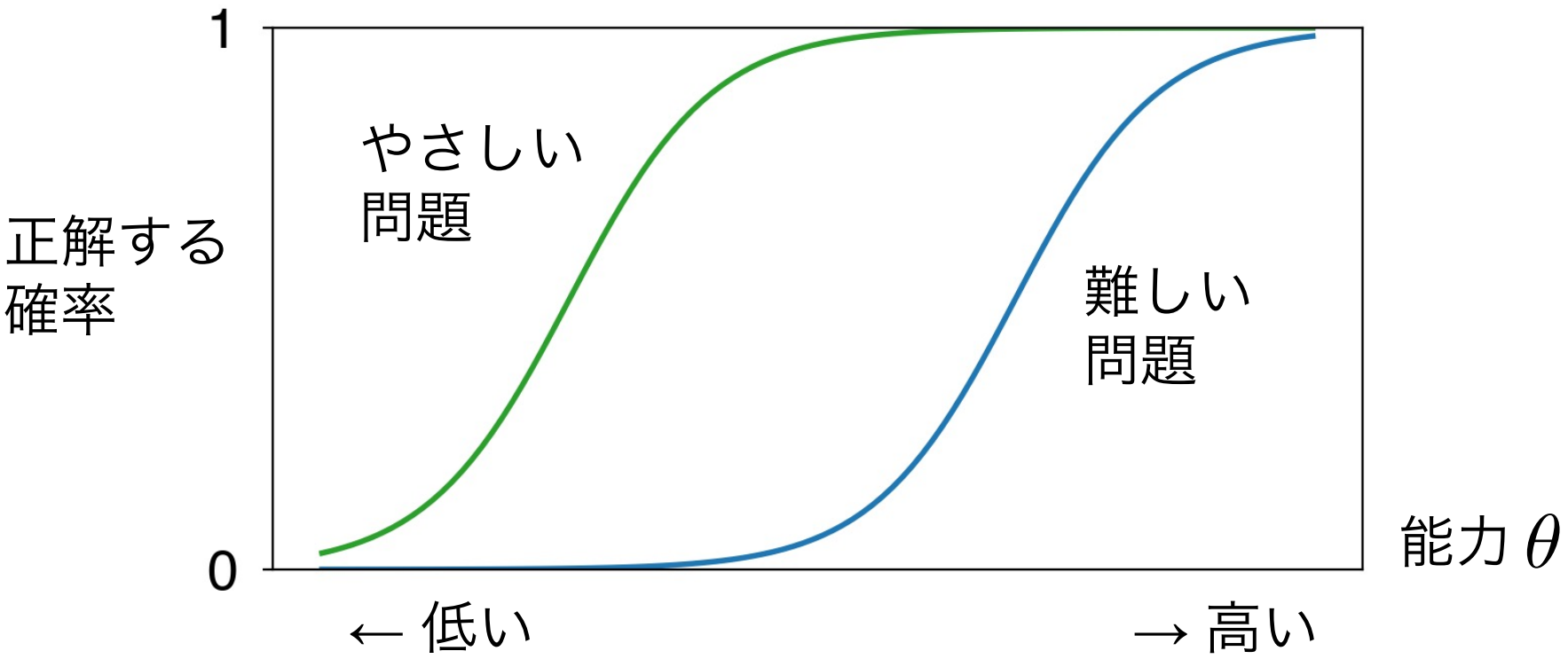
- 人には見えない能力 θ がある (未知)
  - $\theta_A, \theta_B, \theta_C, \theta_D$ は本当はわれわれには未知
- θ は平均が0で、分散が1の標準正規分布とよばれる釣り鐘型の分布に従っている (一般的な仮定)
- 偏差値のようなもの (0＝偏差値50, 2=偏差値70)

能力が高いと、問題に正解しやすい



- 人間なのでミスもあり、結果は確率的

問題には「難しさ」がある



- やさしい問題は、能力が低くても正解できる
- 難しい問題は、能力が高くないと正解できない

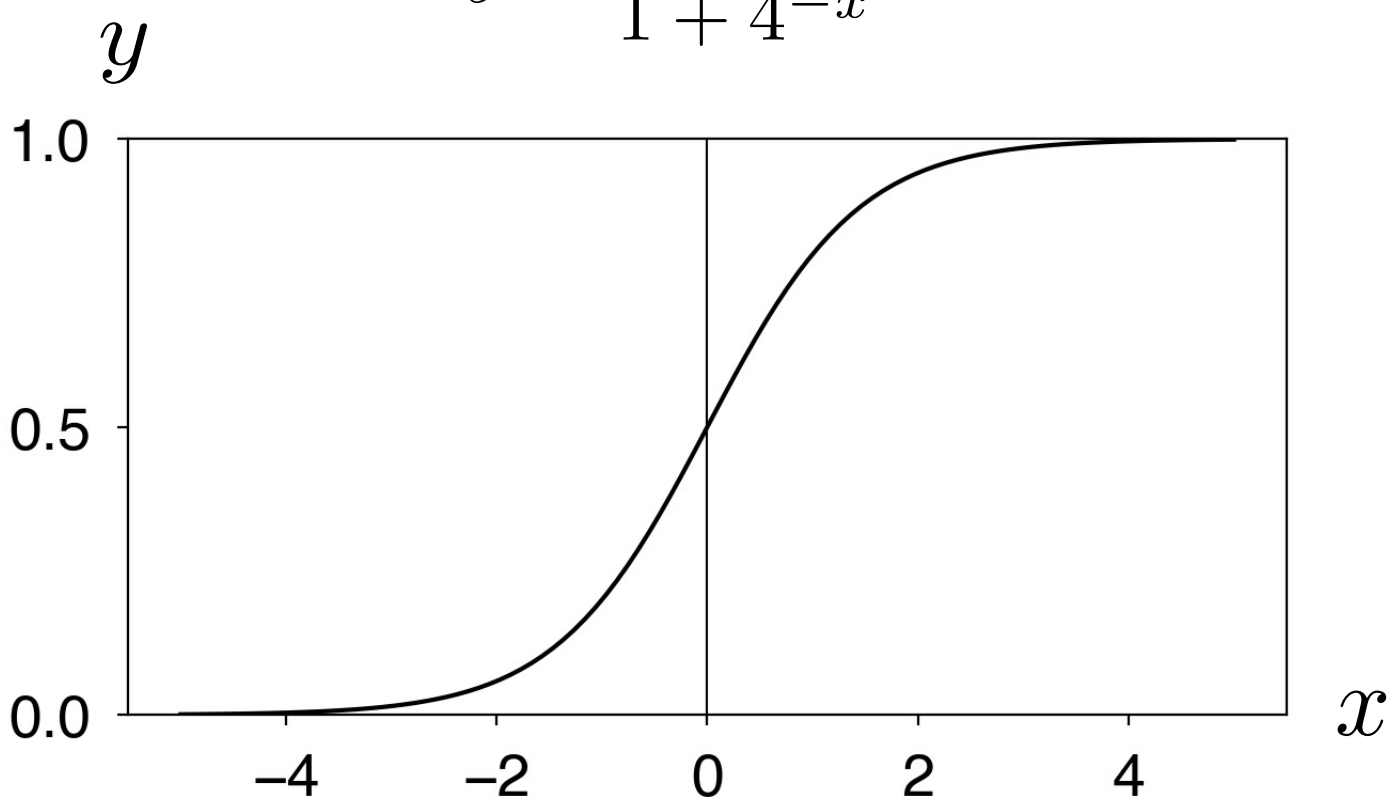
数学的に書くと...

- ロジスティック曲線

$$y = \frac{1}{1 + 4^{-x}}$$

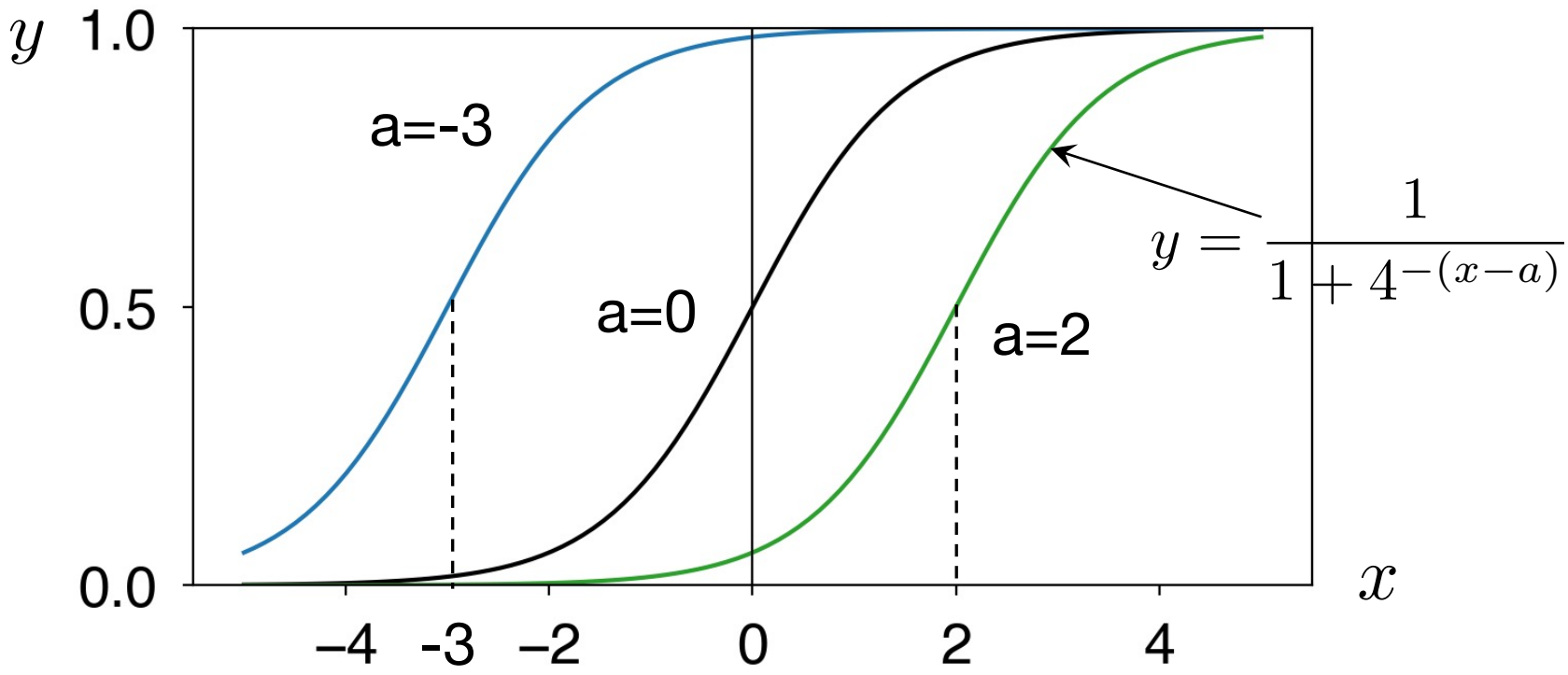
大学以降は

$$y = \frac{1}{1 + e^{-ax}}$$



数学的に書くと... (2)

- $y = \dfrac{1}{1+4^{-x}}$ を x 方向に a だけ動かした曲線は、
  $x \to (x-a)$ を代入して $y = \dfrac{1}{1+4^{-(x-a)}}$

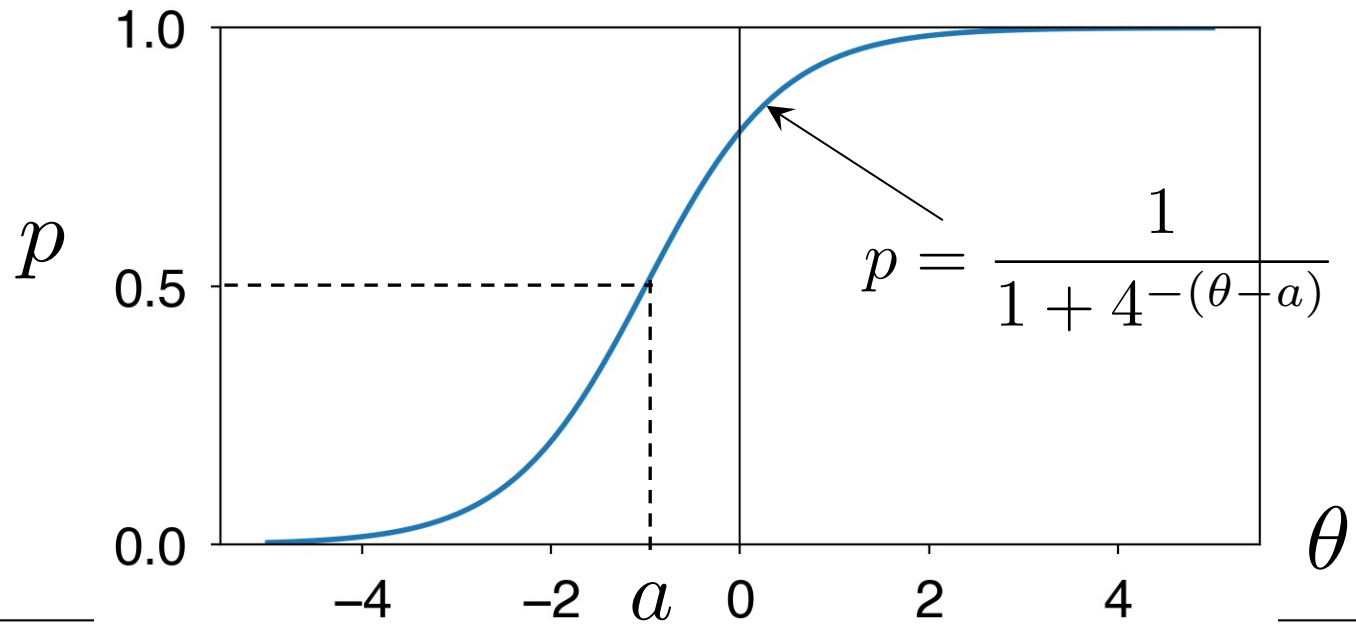


数学的に書くと... (3)

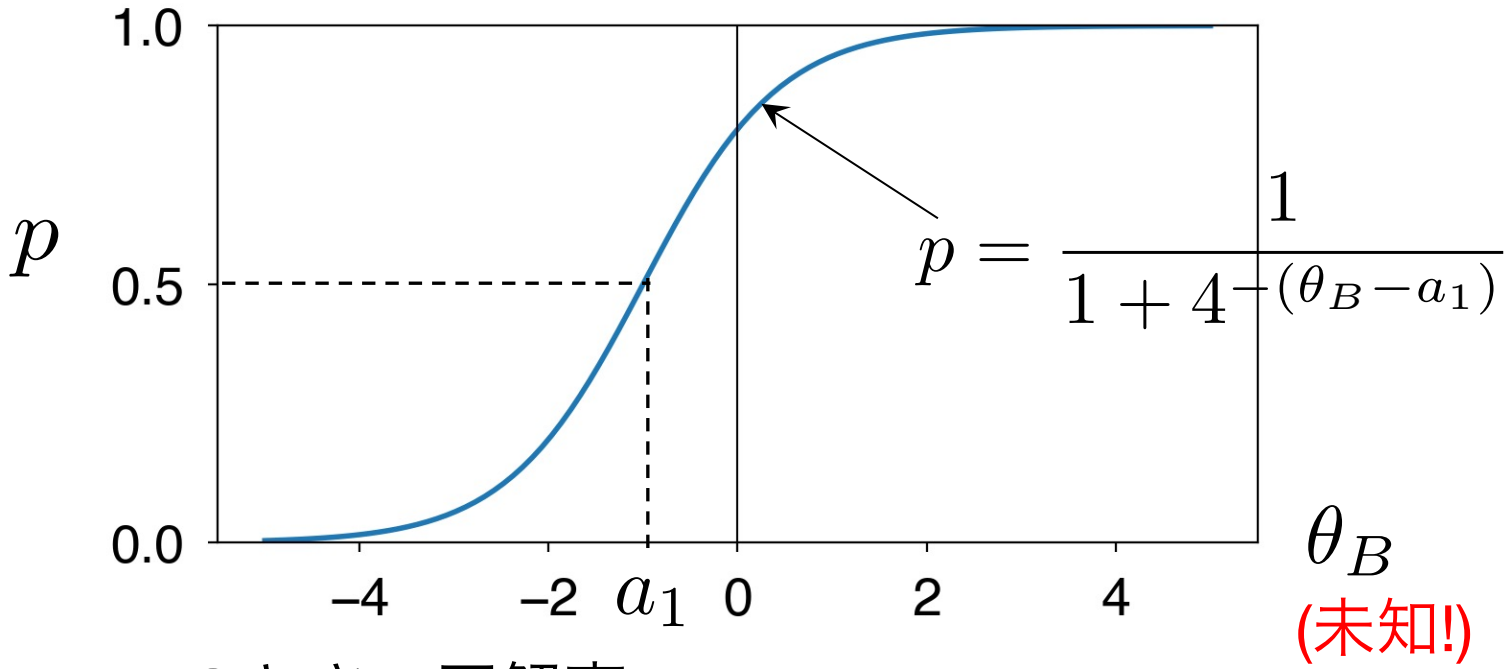
- いま、横軸xは人の能力 θ、縦軸yは正解率 p なので、

$$p = \frac{1}{1 + 4^{-(\theta - a)}}$$

と表せる。 (能力が高ければ正解しやすい)



例: Bさんが問1に正解する確率



- $\theta_B = a_1$ のとき、正解率0.5
- θ_B が大きいと、正解率が上がる

データを見てみる

	Aさん	Bさん	Cさん	Dさん
問題1	1	0	1	0
問題2	1	1	1	1
問題3	1	0	—	1
問題4	0	0	0	0
：				
問題20	1	0	1	—

→ 難しさ a_1

→ 難しさ a_2

↓ 能力 θ_A

↓ 能力 θ_B

正解する確率

$$p = \frac{1}{1 + 4^{-(\theta_B - a_1)}}$$

$a_1, a_2, \cdots$ の大きさによるが、
θ_B は小さいのでは...?

データの確率

	Aさん	Bさん	Cさん	Dさん
問題1	1	0	1	0
問題2	1	1	0	0
問題3	1	0	—	1
問題4	0	0	0	—
：				
問題20	1	0	1	—

- 右の表から、$D_{11}=1$, $D_{21}=0$, $D_{31}=1$, $D_{41}=0$, ... なので、
- データの確率＝

$$\frac{1}{1+4^{-(\theta_1-a_1)}} \times \left(1-\frac{1}{1+4^{-(\theta_2-a_1)}}\right) \times \frac{1}{1+4^{-(\theta_3-a_1)}}$$

正解　　不正解　　正解

$$\times \left(1-\frac{1}{1+4^{-(\theta_4-a_1)}}\right) \times \cdots$$

不正解

データの確率を最大化する
$\theta_1, \theta_2, \theta_3, \theta_4, a_1, a_2, \cdots, a_{20}$
を見つけたい

データの確率 (2)

	Aさん	Bさん	Cさん	Dさん
問題1	1	0	1	0
問題2	1	1	0	0
問題3	1	0	—	1
問題4	0	0	0	—
:				
問題20	1	0	1	—

- データ全体の確率は、個々の1/0が出る確率の積
- *i* さんが問題 *n* に正解する確率は、

$$\frac{1}{1+4^{-(\theta_i - a_n)}}$$

- データ全体の確率は、

Πは、和を表すΣの積版

$$p(D) = \prod_{i=1}^{4}\prod_{n=1}^{20} p(D_{in})$$

$$= \prod_{i=1}^{4}\prod_{n=1}^{20} \underbrace{\left(\frac{1}{1+4^{-(\theta_i - a_n)}}\right)^{D_{in}}}_{\text{正解した場合}} \underbrace{\left(1-\frac{1}{1+4^{-(\theta_i - a_n)}}\right)^{1-D_{in}}}_{\text{間違った場合}}$$

どうやって求める?

- データの確率

$$p(D) = \prod_{i=1}^{4} \prod_{n=1}^{20} p(D_{in})$$

$$= \prod_{i=1}^{4} \prod_{n=1}^{20} \underbrace{\left(\frac{1}{1+4^{-(\theta_i - a_n)}} \right)^{D_{in}}}_{\text{正解した場合}} \underbrace{\left(1 - \frac{1}{1+4^{-(\theta_i - a_n)}} \right)^{1-D_{in}}}_{\text{間違った場合}}$$

- 確率を最大にするパラメータ $\theta_1, \theta_2, \theta_3, \theta_4, a_1, a_2, \cdots, a_{20}$?
- 高校の授業のような閉じた形では解けない！
  → 適当な初期値からはじめて、数値的に求める
  (EMアルゴリズム、MCMC法)

実際に計算してみる

	Aさん	Bさん	Cさん	Dさん	推定した難易度 a
問題1	1	0	1	0	0.077
問題2	1	1	1	1	-1.154
問題3	1	0	—	1	-0.262
問題4	0	0	0	—	1.158
問題5	1	1	0	—	-0.249
問題6	1	0	1	—	-0.266
推定した能力 θ	0.881	-0.546	0.197	0.134	

- Aさんの能力が高く、Bさんが低いなどが連続値でわかる
- 全員正解の問題2の難易度が低く、問題4の難易度が高いことも、数値的にわかる　(＝見えないデータ)

項目反応理論 (item response theory)

- これは、心理統計学で項目反応理論とよばれるモデル
  - TOEICの採点は、項目反応理論で行われています
  - 教育学や政治学、自然言語処理などでも応用されている

	Aさん	Bさん	Cさん	Dさん	推定した難易度 a
問題1	1	0	1	0	0.077
問題2	1	1	1	1	-1.154
問題3	1	0	—	1	-0.262
問題4	0	0	0	—	1.158
問題5	1	1	0	—	-0.249
問題6	1	0	1	—	-0.266
推定した能力 θ	0.881	-0.546	0.197	0.134	

アンケート分析に戻ると...

質問	Aさん	Bさん	Cさん	Dさん
1. 東京は暮らしやすい?	1	0	1	1
2. 東京の物価は高い?	1	0	—	0
3. 東京は安全?	1	1	0	0
4. 東京の交通は便利?	1	0	1	1
:				
10. 東京は好きな都市?	1	0	1	1

- Aさん、Bさん、...のパラメータ $\theta \rightarrow$ その人が1と答えやすい傾向の強さ (全員同じではない!)
- 質問1、質問2、...のパラメータ $a \rightarrow$ その質問で1と答えやすい傾向の強さ (どの質問も同じではない!)

政治学での応用

質問	議員A	議員B	議員C	議員D
1. 憲法改正に賛成?	1	0	1	1
2. 夫婦別姓に賛成?	1	0	—	0
3. 自衛隊海外派遣は合憲?	1	0	0	0
4. 外国人参政権を認める?	0	1	1	1
:				
20. 職業教育をすべき?	1	0	1	1

- 自民党→ほとんど賛成、共産党→ほとんど反対 のバイアス
- 賛成/反対のデータから、各議員の右派←→左派の思想的な位置が連続値でわかる
- 理想点 (ideal point) として、計量政治学で分析されている

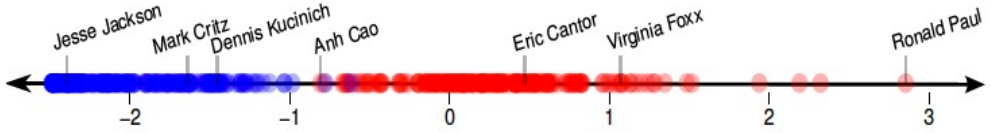


Figure 1: Traditional ideal points separate Republicans (red) from Democrats (blue).

ここまでのまとめ

質問	議員A	議員B	議員C	議員D
1. 憲法改正に賛成?	1	0	1	1
2. 夫婦別姓に賛成?	1	0	—	0
3. 自衛隊海外派遣は合憲?	1	1	0	0
4. 外国人参政権を認める?	1	0	1	1
:				
20. 職業教育をすべき?	1	0	1	1

- 上のようなアンケートデータの背後に、見えないデータ $\theta_A, \theta_B, \theta_C, \theta_D, a_1, \cdots, a_{20}$ を仮定して分析
  - 単純に平均を取るより、本質的な分析ができる
- データをもっともよく説明する $\theta_A, \theta_B, \theta_C, \theta_D, a_1, \cdots, a_{20}$ を、数学的なモデルを作って計算する

現代短歌の評価データ (研究中)

- 脳科学を用いて情動を理解する共同研究の一環として、現代短歌の評価を7段階で行ってもらった
  - 京大短歌会および早稲田短歌会の、合計40名 (現在)

	A	B	C	D	E	F	G	H	I
1	抜かれても雲は車を追いかけない雲には雲のやり方がある	2	5	4	5	6	6	5	4
2	人間のための明かりを消ししのち闇にはうごく機械七台	4	5	3	5	6	6	5	5
3	少女群 紺の水着の胸うすくみづにあるときひとたばの葦	4	4	3	5	7	7	4	6
4	がらんどうの海は冷えゐて此処に立つ吾らのほかに彩をもたない	5	5	3	5	6	5	7	6
5	カップ焼きそばにてお湯を切るときにへこむ流しのかなしきしらべ	5	5	3	6	6	5	3	5
6	郊外のショッピングモールへ近づけば満州国に来た心地する	6	4	2	3	6	4	1	3
7	瞬間のやはらかき笑み受くるたび水切りさるるわれと思へり	5	5	5	6	7	5	3	5
8	ブラインド下りたる昼の図書館を浸す水中のやうなる時間	3	4	4	5	5	3	5	5
9	もしぼくが男だったらためらわず凭れた君の肩であろうか	4	5	2	3	7	7	7	6
10	生殖とかかわりのない愛なども容れてどこへもゆかぬ方舟	6	5	3	3	6	5	6	7
11	逢えばくるうこころ逢わなければくるうこころ愛に友だちはいない	4	5	5	4	5	6	7	4
12	すきなひとに干してもらえた下着たち来世はきっと梨になれるよ	3	5	2	4	5	5	3	5
13	中央線に揺られる少女の精神的外傷(トラウマ)をバターのように溶かせ夕焼け	2	5	2	6	5	6	7	4
14	天井まで「少年ジャンプ」積んでいた小坂の部屋から見た夕焼け	3	4	2	6	6	5	1	2
15	どの犬も目を合わせないこれまでもすきなだけではだめだったから	4	5	2	5	6	5	2	5
16	花火ってひらくばっかり剥き出しのただたくさんの副詞となって	4	5	5	6	6	6	2	6

現代短歌の評価データ

Samejima (1969)

- 評価値が1~7なので、「段階反応モデル」というモデルで計算すると、各歌の潜在的な θ (評価値)が求められる

θ	歌
0.9208	観覧車回れよ回れ想ひ出は君には一日我には一生
0.9002	くれないの二尺伸びたる薔薇の芽の針やはらかに春雨のふる
0.7390	金色のちひさき鳥のかたちして銀杏ちるなり夕日の丘に
0.7356	生年と没年結ぶハイフンは短い誰のものも等しく
0.6808	向日葵は金の油を身にあびてゆらりと高し日のちひささよ
0.6437	切り終へて包丁の刃の水平を見る眼の薄き水なみだちぬ
:	
-0.1455	「研修中」だったあなたが「店員」になり真剣な眼差しがいい
-0.1514	まなつあさぶろあがりてくれば曙光さすさなかはだかの感傷機械
-0.1742	天井まで「少年ジャンプ」積んでいた小坂の部屋から見た夕焼け
-0.2121	恋をすることになるのだこの夏に出逢いたかったひとに出逢って
-0.5339	2x-5y=0 ピーチミントのガム噛みながら

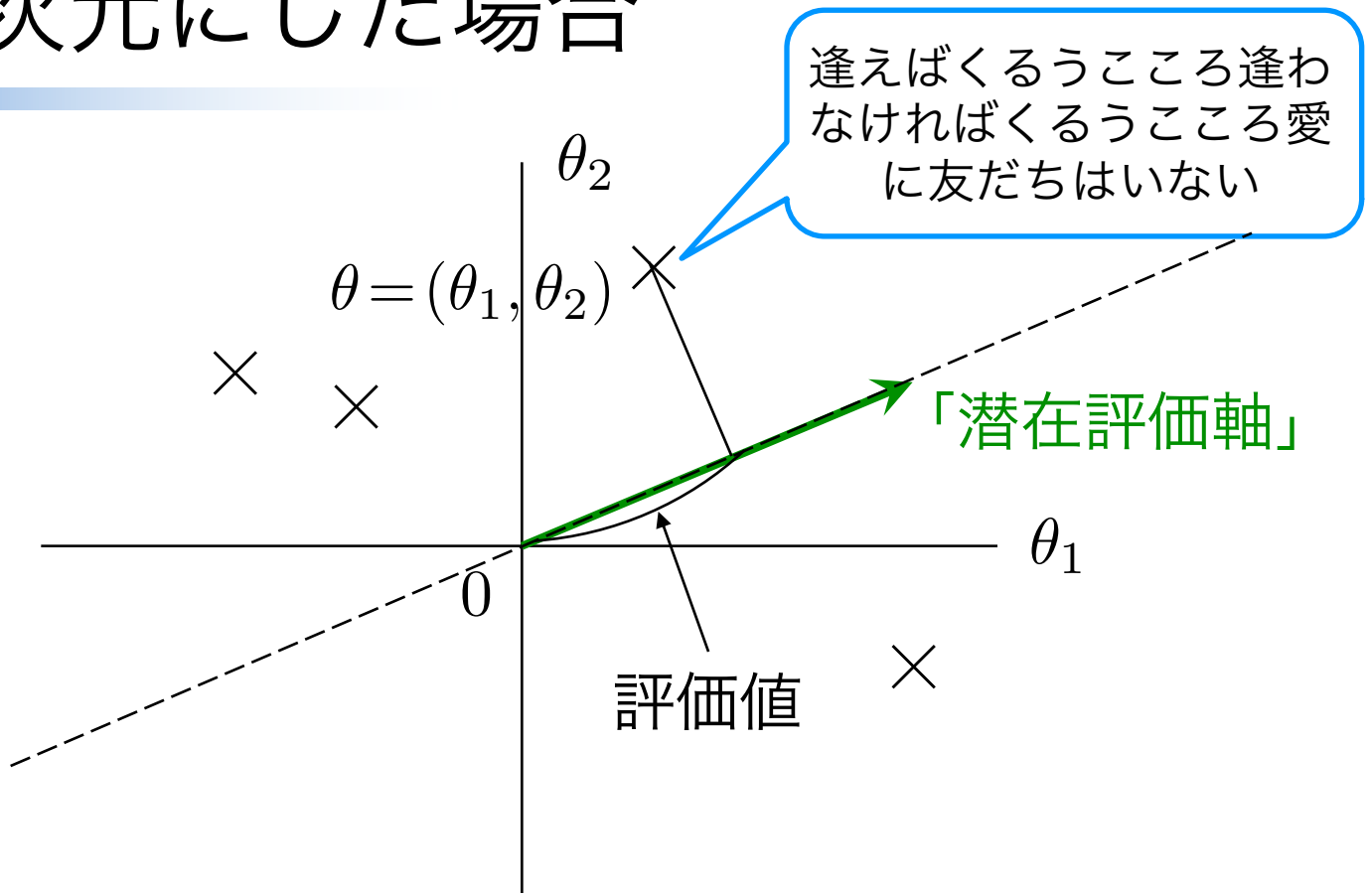
現代短歌の評価データ

- 人によって、かなり評価の傾向が異なる?

	A	B	C	D	E	F
1	抜かれても雲は車を追いかけない雲には雲のやり方がある	2	5	4	5	6
2	人間のための明かりを消ししのち闇にはうごく機械七台	4	5	3	5	6
3	少女群 紺の水着の胸うすくみづにあるときひとたばの葦	4	4	3	5	7
4	がらんどうの海は冷えゐて此処に立つ吾らのほかに彩をもたない	5	5	3	5	6
5	カップ焼きそばにてお湯を切るときにへこむ流しのかなしきしらべ	5	5	3	6	6
6	郊外のショッピングモールへ近づけば満州国に来た心地する	6	4	2	3	6
7	瞬間のやはらかき笑み受くるたび水切りさるるわれと思へり	5	5	5	6	7
8	ブラインド下りたる昼の図書館を浸す水中のやうなる時間	3	4	4	5	5
9	もしぼくが男だったらためらわず凭れた君の肩であろうか	4	5	2	3	7
10	生殖とかかわりのない愛なども容れてどこへもゆかぬ方舟	6	5	3	3	6
11	逢えばくるうこころ逢わなければくるうこころ愛に友だちはいない	4	5	5	4	5

- 「評価の傾向」を数学的に表せないか?
  → 項目反応理論で、θ が1次元ではなく多次元の場合

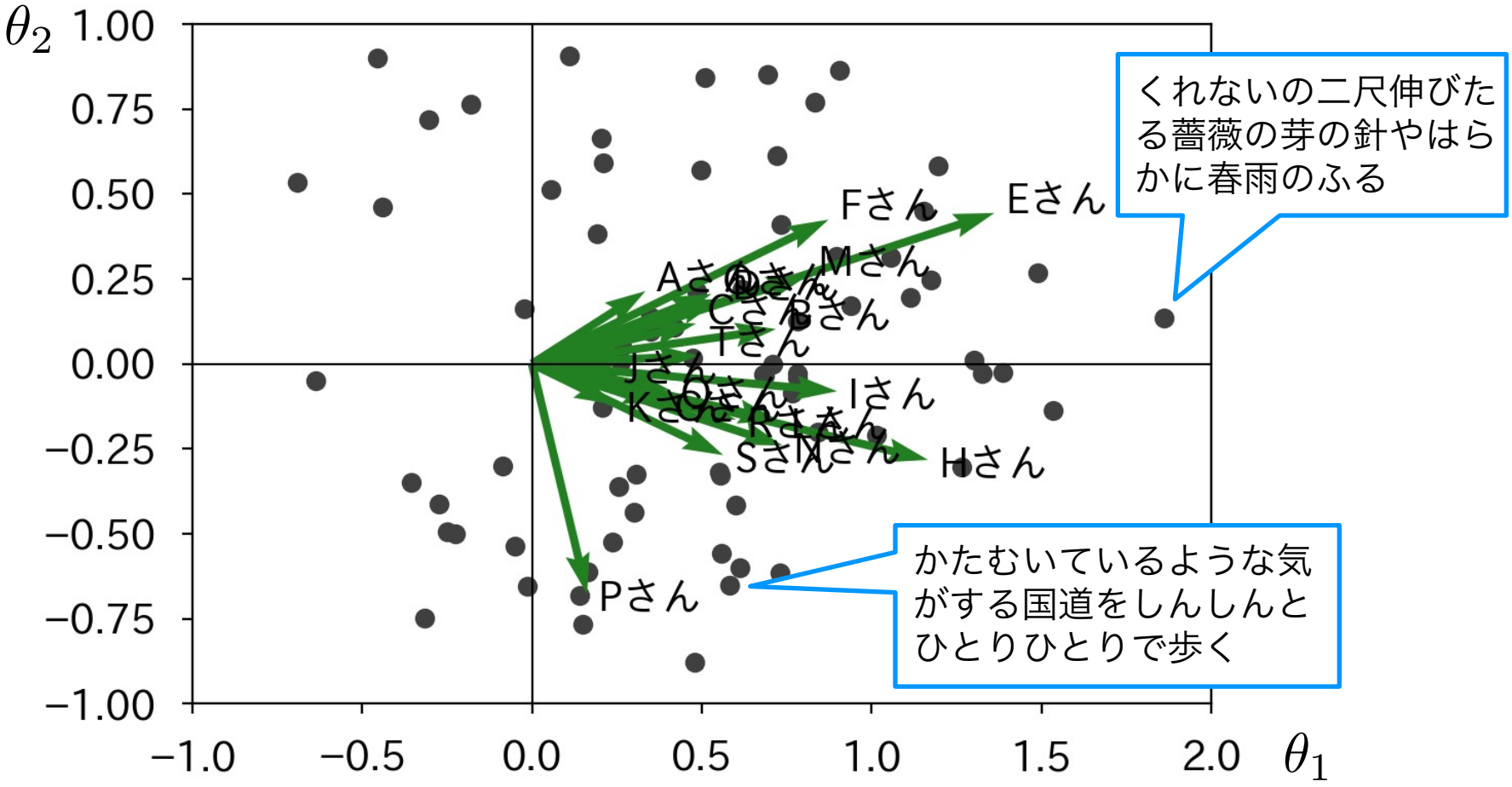
θ を2次元にした場合



- 2次元の θ は、「潜在評価軸」に下ろした垂線の足で評価される

θ を2次元にした場合 (結果)

- 各点が短歌、緑の矢印が各評価者の潜在的な「評価軸」



国会議事録の分析

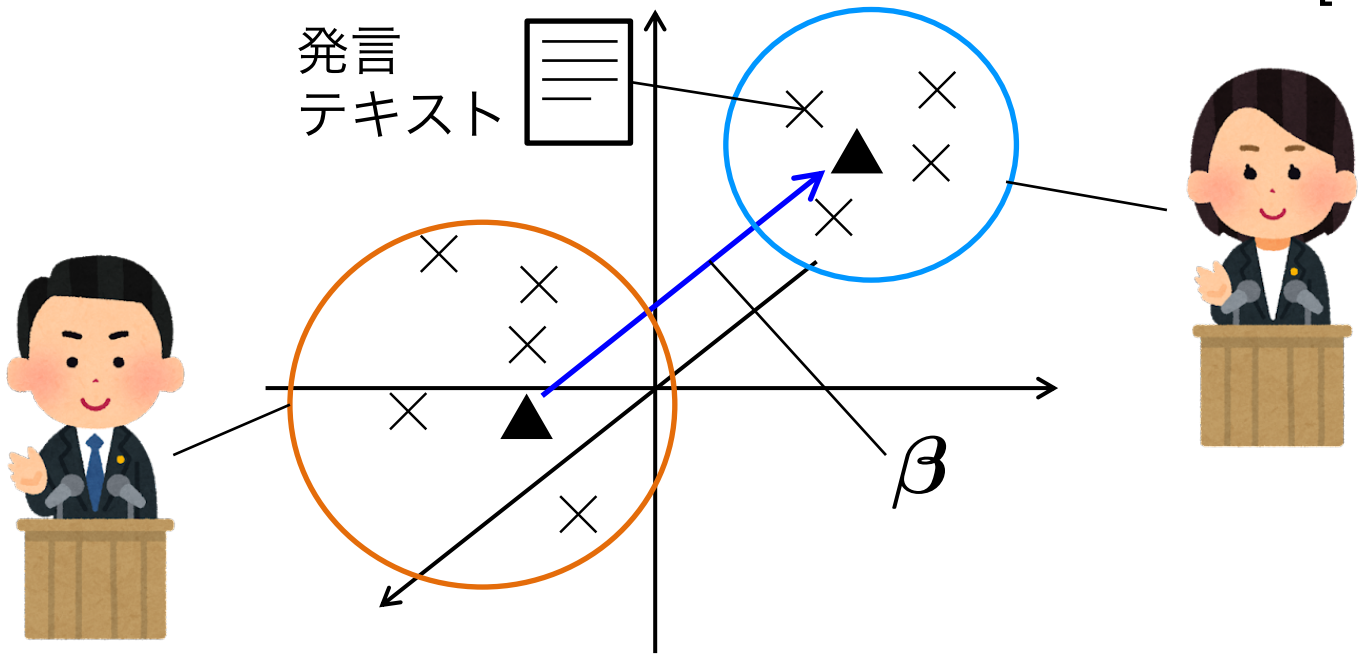
- 2021年春の通常国会の衆議院・農林水産委員会の議事録をWebから保存
  - 議員の論点の違いが比較的明らかな委員会
  - 12回の開催、1,324個の発言
- 発言数の多い玉木雄一郎議員(国民民主党)の発言を $\theta > 0$ 、田村貴昭議員(日本共産党)の発言を $\theta < 0$ として、それぞれランダムに20発言を抽出して「議論の軸」を抽出する
- PLSS：確率的潜在意味スケーリングという方法(持橋2021)

国会議事録の例 (農林水産委員会)

○玉木委員
今、飲食店はすごくコロナの影響を受けていて、店舗型からデリバリー型に変えようとして、必死になって頑張っているところもあるんですね。そういうところはどうしてもプラスチックのスプーンとかを使っていて、更に追加の負荷がかかるのではないのかと心配されているところもいるんです。
ただ、そういった環境対策は必要なので否定するものではないんですけれども、果たしてやることがＣＯ2の削減に本当にどれだけつながるのかとか、説明責任が大事だと思うんですよ。...
○田村（貴）委員
また続きをやらせていただきたいと思いますけれども、次の質問は、米価下落問題についてであります。
新型コロナウイルスの感染拡大による外食需要の激減、インバウンド需要の消滅によって、米余りと米価下落が一層深刻化しています。
民間在庫は四か月連続で三百万トンを超えており、一月二十六日に行われた二〇二一年産の政府備蓄米入札では、全農が、六十キロ、...

PLSS (確率的潜在意味スケーリング)

[持橋2021]



- Word2vecから得られる意味空間上に、2人の発言を両軸にとる「論点軸」β を考えて最適化する
  - この見えない軸上で項目反応理論を考える

国会議事録の分析 (結果)

θ	発言者	発言内容
1.6410	大串（博）委員	立憲民主党・無所属の大串です。　早速質疑に入ります。　貯保法ですけれども、私は、
1.4988	矢上委員	時間の関係で次の質問に移らせてもらいますけれども、低コスト化対策ですね。二問あっ
1.4838	重徳委員	だから、農水省に何の非もないのかと言っているんですよ。　例えば、大臣、富山県の御
1.3139	大串（博）委員	全くちぐはぐですね。一時的な要因で余っているんだったら、一時的に市場から切り離さ
1.1895	本郷政府参考人	木材流通に関してでございます。需給のミスマッチを起こさないように、生産、加工の事
1.0874	玉木委員	国民民主党の玉木雄一郎です。　本法案についてまず質問いたします。　先ほどから、規
1.0784	玉木委員	コロナにはいろいろなことを教えてもらったなと思ったんですが、例えば、マスク一つ国
1.0716	近藤（和）委員	石川県能登半島の近藤和也でございます。よろしくお願いいたします。　ＣＯＶＩＤ―1
1.0316	金子（恵）委員	今、イノベーションの話もされたので、済みません、順番を変えて、林業労働力の育成、
0.8315	神谷（裕）委員	そうしますと、遡れる限り遡るということだと思うんですが、そこで、先ほど議論になっ
⋮		
−0.5228	野上国務大臣	御指摘のございました主要農作物種子法につきましては、昭和二十七年に、戦後の食料増
−0.7074	野上国務大臣	間伐等特措法によりまして、平成二十年の法律制定後、一定以上の森林面積を有します市
−0.8432	葉梨副大臣	お答えいたします。　佐々木先生の資料の二の品目横断的経営安定対策、これが導入され
−1.0359	水田政府参考人	お答えいたします。　委員御指摘の冊子の二ページのところに「ＥＵやアメリカの現状」
−1.5408	高鳥委員長	お諮りいたします。　ただいま議決いたしました法律案に関する委員会報告書の作成につ
−1.5771	高鳥委員長	起立少数。よって、本修正案は否決されました。　次に、原案について採決いたします。
−1.9059	大串（博）委員	今回の農中さんの議論を契機に是非いい議論をしていただきたいと思いますし、間違って
−2.2223	田村（貴）委員	私は、日本共産党を代表して、本法案に反対の立場から討論を行います。　第一に、改正
−2.5166	田村（貴）委員	私は、日本共産党を代表して、畜舎等の建築及び利用の特例に関する法律案に反対の討論
−2.6210	重徳委員	立憲民主党の重徳和彦です。　今日は矢上筆頭、先輩、同僚議員の御了解をいただきまし
−3.5215	新井政府参考人	お答えいたします。　ＯＩＥ連絡協議会は、産業界及び学界における技術者又は学識経験

- 玉木ー田村の極性軸上に、θ で連続的に発言を並べられる

国会議事録の分析 (結果)

θ	発言者	発言内容
1.6410	大串（博）委員	立憲民主党・無所属の大串です。　早速質疑に入ります。　貧
1.4988	矢上委員	時間の関係で次の質問に移らせてもらいますけれども、低コス
1.4838	重徳委員	だから、農水省に何の非もないのかと言っているんですよ。
1.3139	大串（博）委員	全くちぐはぐですね。一時的な要因で余っているんだったら、
1.1895	本郷政府参考人	木材流通に関してでございます。需給のミスマッチを起こさな
1.0874	玉木委員	国民民主党の玉木雄一郎です。　本法案についてまず質問いた
1.0784	玉木委員	コロナにはいろいろなことを教えてもらったなと思ったんです
1.0716	近藤（和）委員	石川県能登半島の近藤和也でございます。よろしくお願いいた
1.0316	金子（恵）委員	今、イノベーションの話もされたので、済みません、順番を変
0.8315	神谷（裕）委員	そうしますと、遡れる限り遡るということだと思うんですが、
		⋮

- 玉木ー田村の極性軸上に、θ で連続的に発言を並べられる

国会議事録の分析 (結果)

:

−0.5228	野上国務大臣	御指摘のございました主要農作物種子法につきましては、
−0.7074	野上国務大臣	間伐等特措法によりまして、平成二十年の法律制定後、
−0.8432	葉梨副大臣	お答えいたします。　佐々木先生の資料の二の品目横断
−1.0359	水田政府参考人	お答えいたします。　委員御指摘の冊子の二ページのと
−1.5408	高鳥委員長	お諮りいたします。　ただいま議決いたしました法律案
−1.5771	高鳥委員長	起立少数。よって、本修正案は否決されました。　次に、
−1.9059	大串（博）委員	今回の農中さんの議論を契機に是非いい議論をしていた
−2.2223	田村（貴）委員	私は、日本共産党を代表して、本法案に反対の立場から
−2.5166	田村（貴）委員	私は、日本共産党を代表して、畜舎等の建築及び利用の
−2.6210	重徳委員	立憲民主党の重徳和彦です。　今日は矢上筆頭、先輩、
−3.5215	新井政府参考人	お答えいたします。　ＯＩＥ連絡協議会は、産業界及び

- 玉木一田村の極性軸上に、θ で連続的に発言を並べられる

国会議事録の分析 (結果)

- 玉木ー田村議員を極性軸とした際の、各単語の極性
  $\phi_v = \boldsymbol{\beta}^T \vec{v}$

(玉木側)

v	ϕ_v	v	ϕ_v
まずは	0.5167	訴訟	−0.5769
なので	0.4560	傍聴	−0.5646
一つ	0.4364	毀損	−0.5519
ミリ	0.4246	原告	−0.5481
増やす	0.4178	敗訴	−0.5147
整い	0.4109	控訴	−0.5010
もっと	0.4014	係争	−0.4963
もう少し	0.4012	裁判所	−0.4962
植え付ける	0.3995	判決	−0.4808
切り替える	0.3985	審	−0.4727
一番	0.3982	シベリア	−0.4599
どうにか	0.3940	退け	−0.4491
しっかり	0.3903	高裁	−0.4462
同時に	0.3897	弁護	−0.4431
早く	0.3890	最高裁	−0.4410
戦える	0.3886	紛争	−0.4303

(田村側)

今日のまとめ

- 「見えないデータ」を考えて数学的にモデル化することで、アンケートや議会など、多くのデータを適切に分析することが可能
  - 人にはそれぞれ見えない傾向がある
  - 質問項目や発言にも、見えない傾向がある
- データ解析は、ほとんどすべての分野で今後重要になる
- 数学は重要!!
  - 文系でも、数IIIは統計学の理解には必須
  - ただし、難しい問題が解ける必要はありません
- 統数研では、研究所見学なども受け入れています

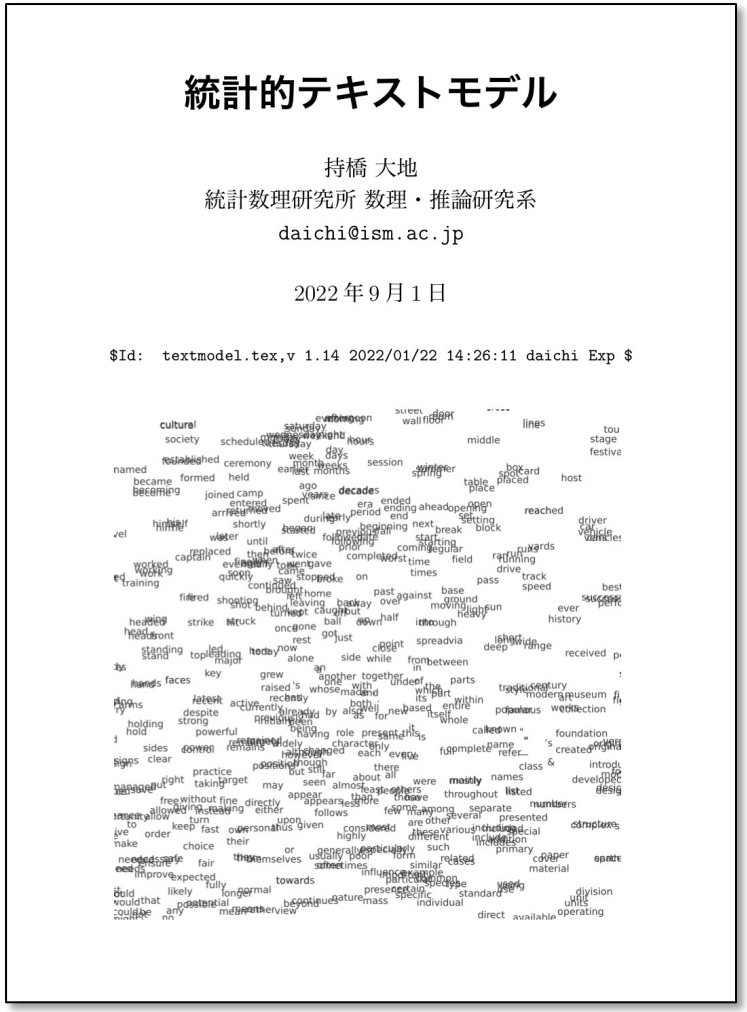
参考図書 (高校生向け)

- 完全に初歩から、わかりやすく説明
- Σから説明しているので大丈夫！
- 大学では、より専門的な教科書もあります

参考図書 (自然言語処理)

- 「岩波データサイエンス Vol.2 統計的自然言語処理」
- 私が特集担当と記事の執筆をしています
- 1500円で読みやすい読み物形式です
- 大きな書店には置いてあるはずです

教科書 (執筆中)



統計的テキストモデル

持橋 大地
統計数理研究所 数理・推論研究系
daichi@ism.ac.jp

2022 年 9 月 1 日

$Id: textmodel.tex,v 1.14 2022/01/22 14:26:11 daichi Exp $

- 岩波書店より、2023年に発売予定
- 皆さんが大学に入る頃には、確実に売っているはずです
- 岩波書店の担当編集者は、小石川の先輩でした (!)

私のホームページ

https://www.ism.ac.jp/~daichi/

- この講演のスライドも、Googleドライブ以外に上記のサイトにも置いておくようにします
- 他にも、様々なスライドや研究資料があります