心理学的により現実的な容認度評定のモデルを求めて

黒田 航 / 杏林大学

Workshop:“見えない” 言語をどう “見る” か —

言語知識へ至る方法論に関する考察と議論

at 17th Annual Meeting of JCLA, 2016/9/10,
明治大学中野キャンパス

始めに

- 発表の目的
  - 容認性判断 (acceptability judgement) の実態を探り，現実に合うモデル化を提案する
    - 言語学者の大半が，それが何なのかちゃんとわかっていない
- 提案
  - 容認性判断は，容認度評定 (acceptability rating) の特殊な場合
- 容認度評定は $A = E \times C \times R$ としてモデル化できる
  - $C^* = C \times R$ は広義の文脈
- 補足
  - 文法性判断 (grammatical judgement) は，容認性判断の特殊な場合
    - であるか，あるいはまったく別種の判断で，本来なら容認性判断と関連づけられないもの

本発表の論点と潜在的な貢献

- 事実の確認
  - 論点 1: 評定者は（理論言語学の素朴な想定に反して）容認度評定で使われる刺激に一様な応答をしない.
  - 論点 2. かと言って，彼らの応答はカオス的な訳ではなく，特徴的な応答パターンを幾つか認める事ができる．別の言い方をすれば，評定者は特定の応答パターンを内在化させている.
- 今後の課題
  - 論点 3. 特定の応答パターンとは課題に対する個人の最適方略の事であり，この数は有限である.
  - 論点 4. この意味での特徴的な応答パターン=バイアスの存在を認識し，それに対する補正を施さない限り，容認度評定の結果は内在化されたシステム=言語知識の実体を明らかにするものだとは言えない.
  - 論点 5. 必要な補正の一つは，容認度評定の結果が一定の仕方でバイアスされた多変量データだと理解する事である.
- 根本的主張
  - 論点 6. このような理解がない形で利用された容認度(評定)は言語知識に対する妥当な観察データを与えず，逆に観察的妥当性を歪めるデータとなる.
  - 追加の論点7: 言語の知識が存在するならば，それは個体群に分散的に体現された集合知である

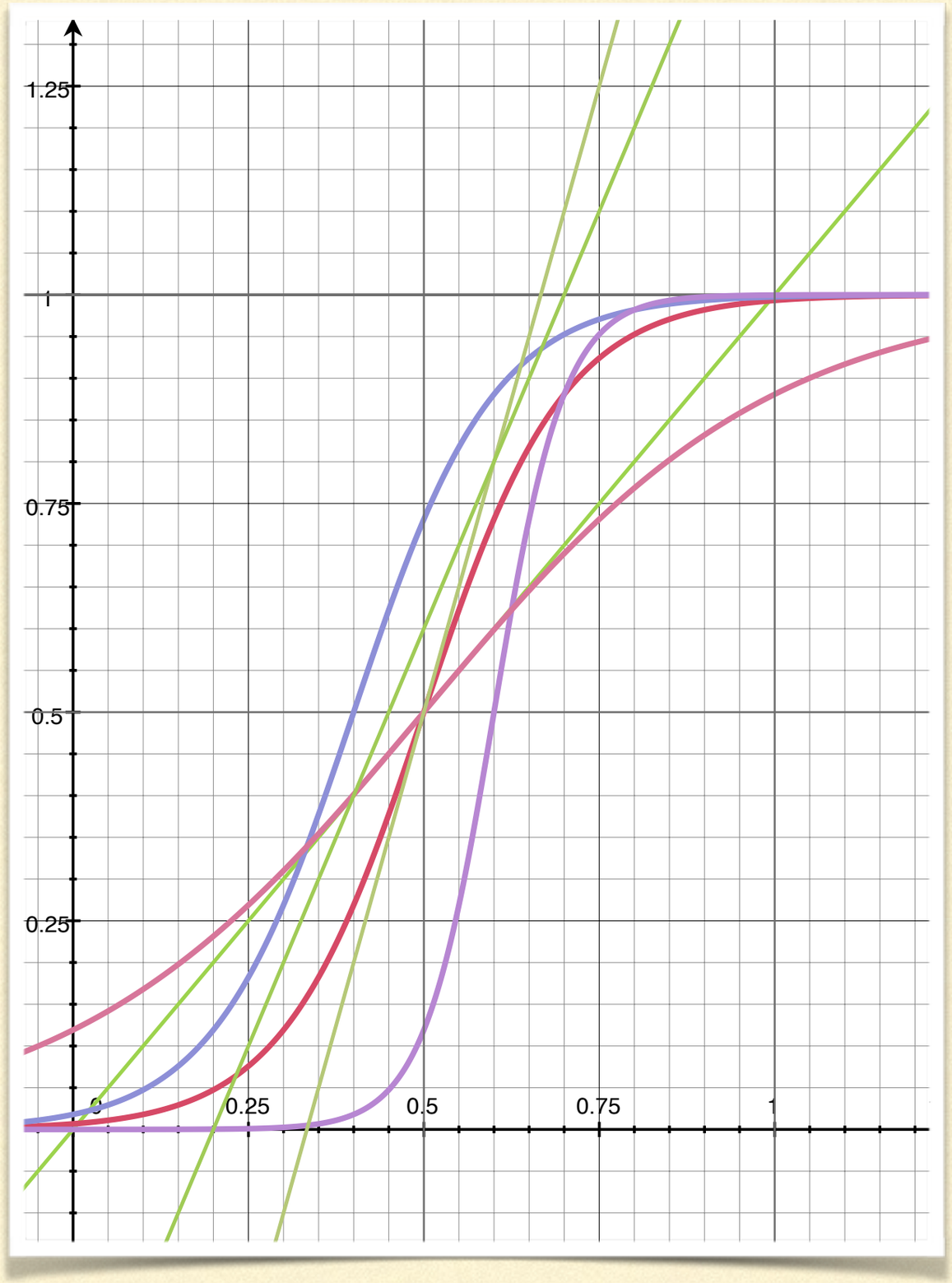
(一言でなく) 三言で言うと

- 1. 言語知識 (~=文法) が特定の個人に完全に体現されていると考えるのは，経験科学の観察妥当性の基準に反する
  - ideal speaker/hearer は明白な虚構
- 2. 従って，(言語学が経験科学であるならば) 言語研究者が,
  - 専門家である自分の容認性判断が絶対で,
    - 他人が与えた自分のと違う判断が間違いだと確信している状態は,
- 言語学の科学性に対する深刻な侵害で，かつ社会的害毒
  - 控え目に言っても単なる自惚れ
- 3. だが，言語学の “証拠” は少なからずこの種の害毒で汚染されていて，証拠の質を上げるために対処しないといけない
  - JCLA 16 での私の発表の論点

補足

- 本発表は発表者の次のオンライン論文の概要です
  - 言語表現の容認度とは何か? また何であるべきか? — 言語学者であるはずなのに，容認度判断が何であるかに自信をもって答えられない (大半の) 人々への手引き
    - http://clsl.hi.h.kyoto-u.ac.jp/~kkuroda/papers/on-acceptability.pdf

容認度(評定)とは何であり，また何でないか？



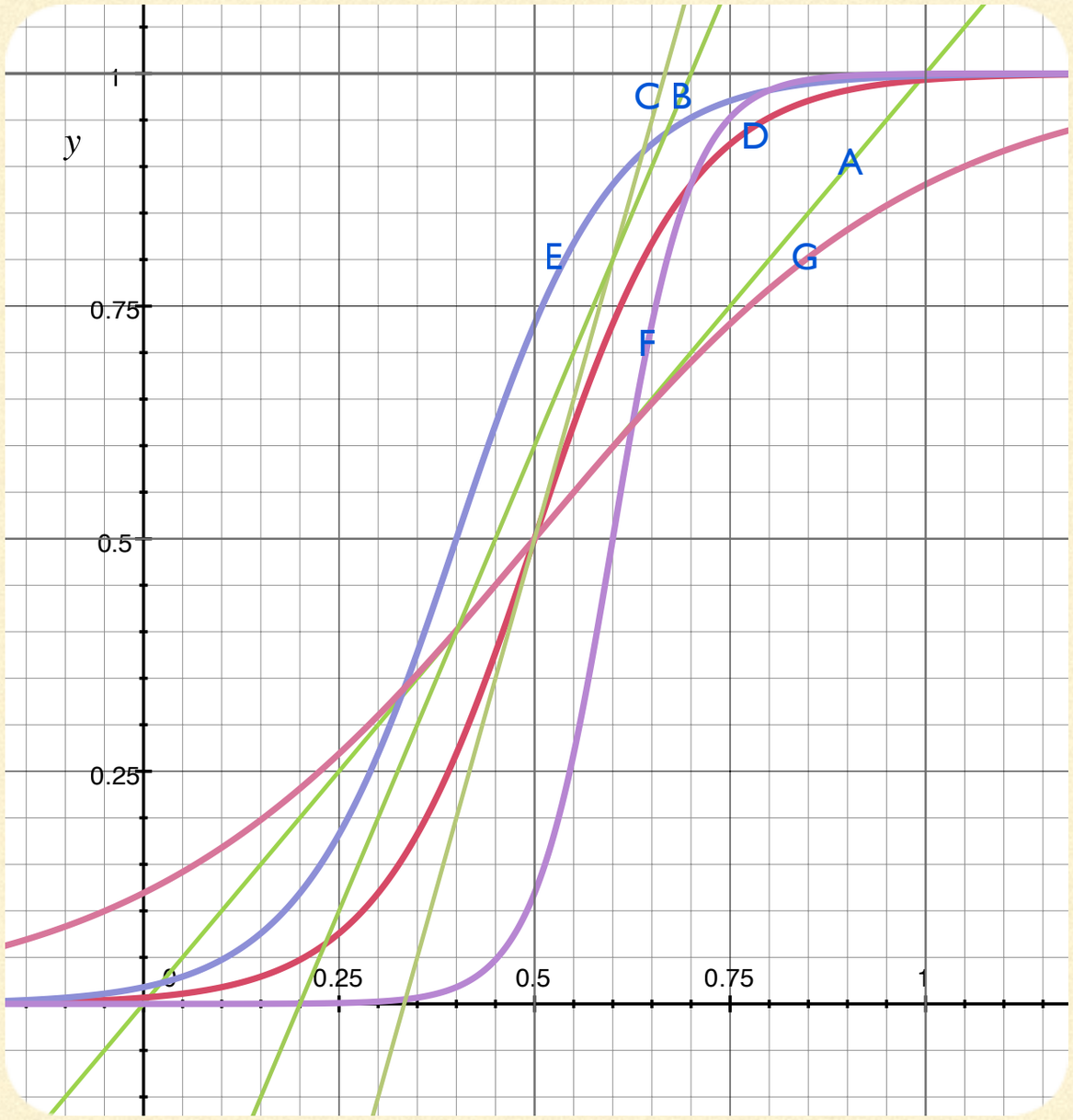
予稿の訂正

- 評定者数: 191名 ⟹ 185名
  - 数え間違いがありましたが，結果に影響ありません
- 評定文: s1, . . ., s1945 ⟹ s1, . . ., s1935 と d1, . . ., d10

容認度評定に関する“俗説”

	理論言語学（想定）	理論言語学（想定）	実態	論点	優位
Q1. ほぼカテゴリー判断か？	YES	NO	YES	分布の形状	生成系
Q2. 連続的か？	NO	YES	YES	1.0 から 0.0 の間で連続的に変化	認知系
Q3. 文脈は影響するか？	NO?	YES	YES	容認度は表現の属性ではない	認知系？
Q4. 判定者は等質か？	YES	YES	NO	個人差=応答戦略の違いがある	どっちも不適切

“ほぼカテゴリー判断” の意味



- 要点
  - 見かけカテゴリカルである性質は連続性と矛盾せず
    - ただし x 軸の解釈は難しい
- 線型関数族 $y = ax - b$ (ただし $y < 0$ では $y = 0$ で $y > 1$ では $y = 1$) と対比
  - A: $y = x$; B: $y = 2x - 0.4$; C: $y = 3x - 1$
- 相転移的変化を表わす Sigmoidal 関数族 $y = 1/(1 + \exp(a(b - x)))$
  - D: $y = 1/(1 + \exp(10\ (0.4 - x)))$
  - E: $y = 1/(1 + \exp(10\ (0.5 - x)))$
  - F: $y = 1/(1 + \exp(20\ (0.6 - x)))$
  - G: $y = 1/(1 + \exp(4\ (0.5 - x)))$

以下の方針

- $A = E \times C \times R$ のモデル化の詳細の提供
  - A: 容認度評定値の分布は
    - E: 個々の表現の固有の傾向
    - C: 狭義の文脈の効果
    - R: 評定者の評定バイアス=戦略
  - の関数である
- ただし
  - $C \times R$ が広義の文脈 C^* の効果
- この想定の下で次の二つの問題を同時に探求
  - 文脈は容認度評定 A にどう影響するか？ (cf. Q3)
  - 評定者の反応パターンの違いをどう扱うか？ (cf. Q4)

論点 I

容認度の程度差はなぜ生じるか？

用語と概念の整理 1/3

- 用語
  - 表現 e_i の評定者 r_k による文脈 c_j を想定した容認度評定 $a(e_i, c_j, r_k) = a_{i,j,k}$ とは，r_k が c_j を想定し e_i に容認度 a の値を一定の範囲 ($0 \leq a \leq 1.0$) で割り当てる課題
    - すべての評価は標準化すれば $0 \leq a \leq 1.0$ の範囲に収まる
    - 容認度は基本的に順位尺度で，一部に間隔尺度と解釈できる場合もある
  - 容認度評定値とは，容認度評定課題の結果
  - e, c, r のグループ化
    - E は表現の集合 $\{e_1, e_2, \ldots, e_n\}$
    - R は評定者の集合 $\{r_1, r_2, \ldots, r_N\}$
    - C は利用可能な文脈の集合 $\{c_1, c_2, \ldots, c_m\}$ とする

用語と概念の整理 2/3

- 定式化
  - $A = E \times C \times R$
- 問題の構造
  - 実際の問題では，E と A が観測可能なデータで，R と C が説明すべき潜在変数
- “文脈” の概念の精緻化
  - 従来の研究で “文脈” と呼ばれていた要素は $C \times R$ という二つの効果の混合
    - 広い意味での文脈 $C^* = C \times R$
    - 狭い意味での文脈 C
- 貢献
  - 従来モデルでは R を C から分離する事を想定していない

容認度の程度差が生じる理由

- 文脈効果
  - 表現 e_i は，異なる文脈 $c_1, c_2, \ldots, c_m$ で異なる容認度 $a_{i,1}, a_{i,2}, \ldots, a_{i,m}$ をもつ
    - $C = \{c_1, c_2, \ldots, c_m\}$，$A_{i,m} = \{a_{i,1}, a_{i,2}, \ldots, a_{i,m}\}$ とする
- この意味での容認度 (の違い) は，e_i の (Darwin 的な意味での) C への適応度 (の違い)

$$A = E \times C$$

	c	…	c	…	c
e	a	…	a	…	a
⋮	⋮	⋱	⋮	⋱	⋮
e	a	…	a	…	a
⋮	⋮	⋱	⋮	⋱	⋮
e	a	…	a	…	a

文脈効果について考えるべき事 1/2

- 疑問
  - 言語学者が表現 e_i の容認度 a_i と言う時，意味しているのは次のどれか？
- 答え 1 (先の定義)
  - $A_i = e_i$ の (Darwin 的な意味での) C への適応度
  - A_i は分布をもつ構造
- 答え 2
  - $a_i = A_{i,m}$ の最大値

文脈効果について考えるべき事 2/2

- 答え 2
  - “e_i の容認度 = $A_{i,m}$ の最大値”
- は理論的につまらない
  - C さえ操作すれば，表現 e_i の容認度は (e_i が文法的である限り)，幾らでも大きくできる
- し実用性も低い
  - 結果を言語処理や心理学で利用しづらい

- 答え 1
  - “e_i の容認度は (Darwin 的な意味での) e_i の C への適応度”
- の方が理論的におもしろいし，現実の複雑なデータとより良く合致
- 発展的問題
  - 潜在変数 C への適応度という形で容認度をモデル化するには，どうしたら良いか
  - C のモデル化で個体差はどう扱えば良いか？

脱線的補足

- 私は
  - 安易なモデル化=理想化に安住するのが大嫌いで
    - ウケる結果が出やすいのは，怪しいという事
  - 複雑な現実を，本質的な複雑性を残しながらモデル化するという泥臭い仕事が大好きで
  - 応用的価値のある理論/モデルが最良の理論/モデルだ
    - 科学モデルは実地のデータ=生データに合致してこそ意味がある
  - と考える人です

文脈の効果の実態は何か？

R の効果を記述するための記述 *A* の拡張

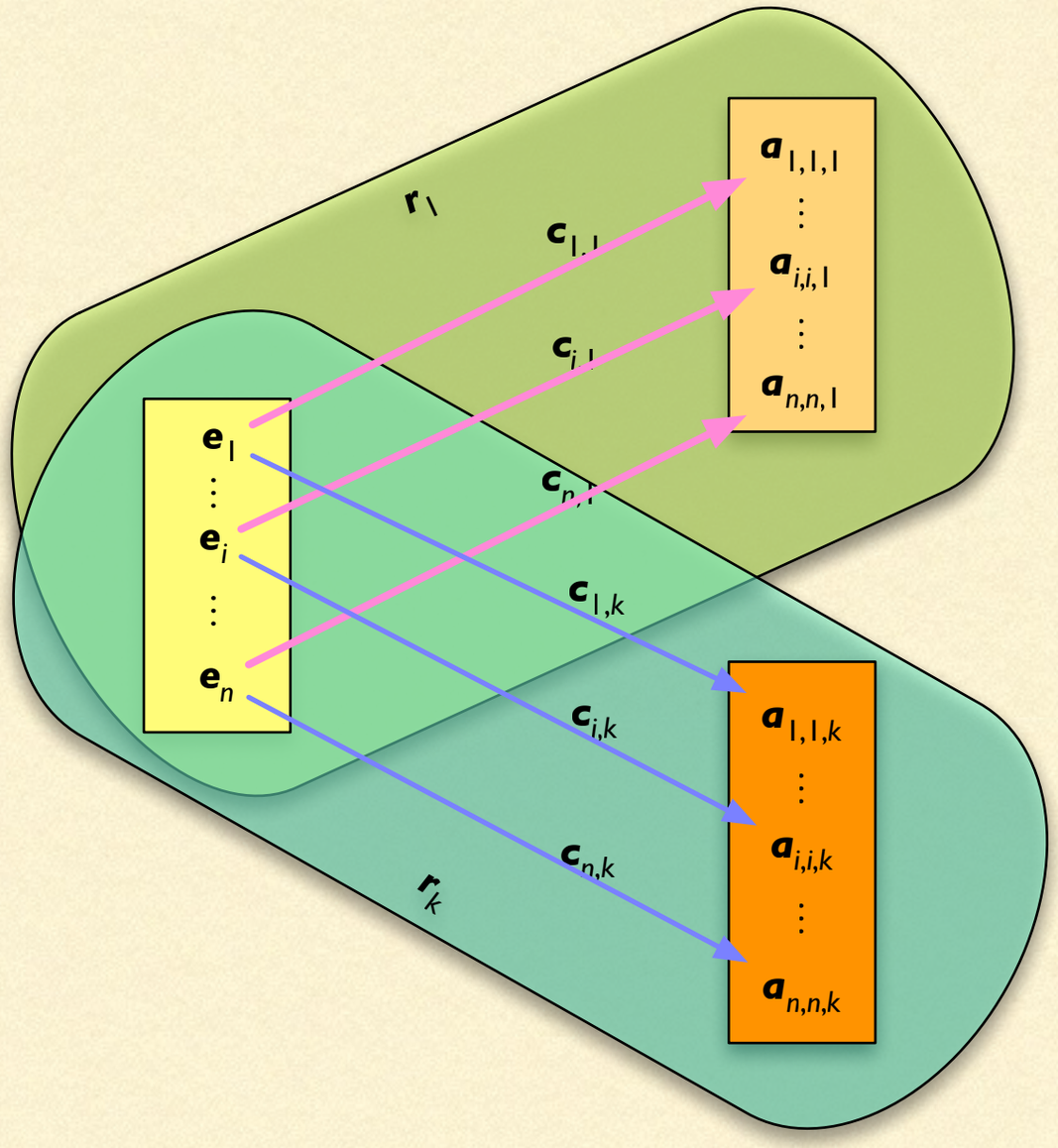
文脈があれば…？

- 副作用
  - 適当な文脈を想定すれば，ほぼ何でも説明できる
    - 実を言うと，これは科学的に全然嬉しくない
- 根本問題
  - そもそも文脈の定義が漠然としている
  - 少なくとも客観的な意味での文脈 (e.g., 順序効果)と評定者の個体差 (e.g., 主観性の変異) とが区別されていない

反応の等質性の仮定

- 問題点
  - 先のモデル化では，個体差が考慮されていない
- と言うより，
  - 評定者の反応は質的に同じだとアプリオリに仮定されている
- これを反応の等質性の仮定 (assumption of homogeneous response) と呼ぶ
- この仮定を放棄して，反応の個人差を明示的に記述する R を導入すると，どうなる？

文脈概念の精緻化



- すべき事
  - 広義の文脈 C^* ≠ 狭義の文脈 C
  - 狭義の文脈 C ⊂ 広義の文脈 C^*
    - R と C は相関するが，同一ではない
- ただし
  - 共通の E に対する，十分な数の R の反応を分類しないと，R の想定する C の実態は把握しようがない
    - R が一様でない限りは

Conflation of *A* = *R*x*E*x*C* (*Case 1*)

$A_{k,j} = [a_{i,1,k}, a_{i,2,k}, \ldots, a_{i,m,k}]$ (*m* は *c* のインデックス)

	e	…	e	…	e
r	A	…	A	…	A
⋮	⋮	⋱	⋮	⋱	⋮
r	A	…	A	…	A
⋮	⋮	⋱	⋮	⋱	⋮
r	A	…	A	…	A

Conflation of $A = R \times E \times C$ (Case 2)

$A^*_{i,k} = [a_{i,1,k}, a_{i,2,k}, \ldots, \ldots, a_{i,m,k}]$ (m は c のインデックス)

	r	…	r	…	r
e	A	…	A	…	A
⋮	⋮	⋱	⋮	⋱	⋮
e	A	…	A	…	A
⋮	⋮	⋱	⋮	⋱	⋮
e	A	…	A	…	A

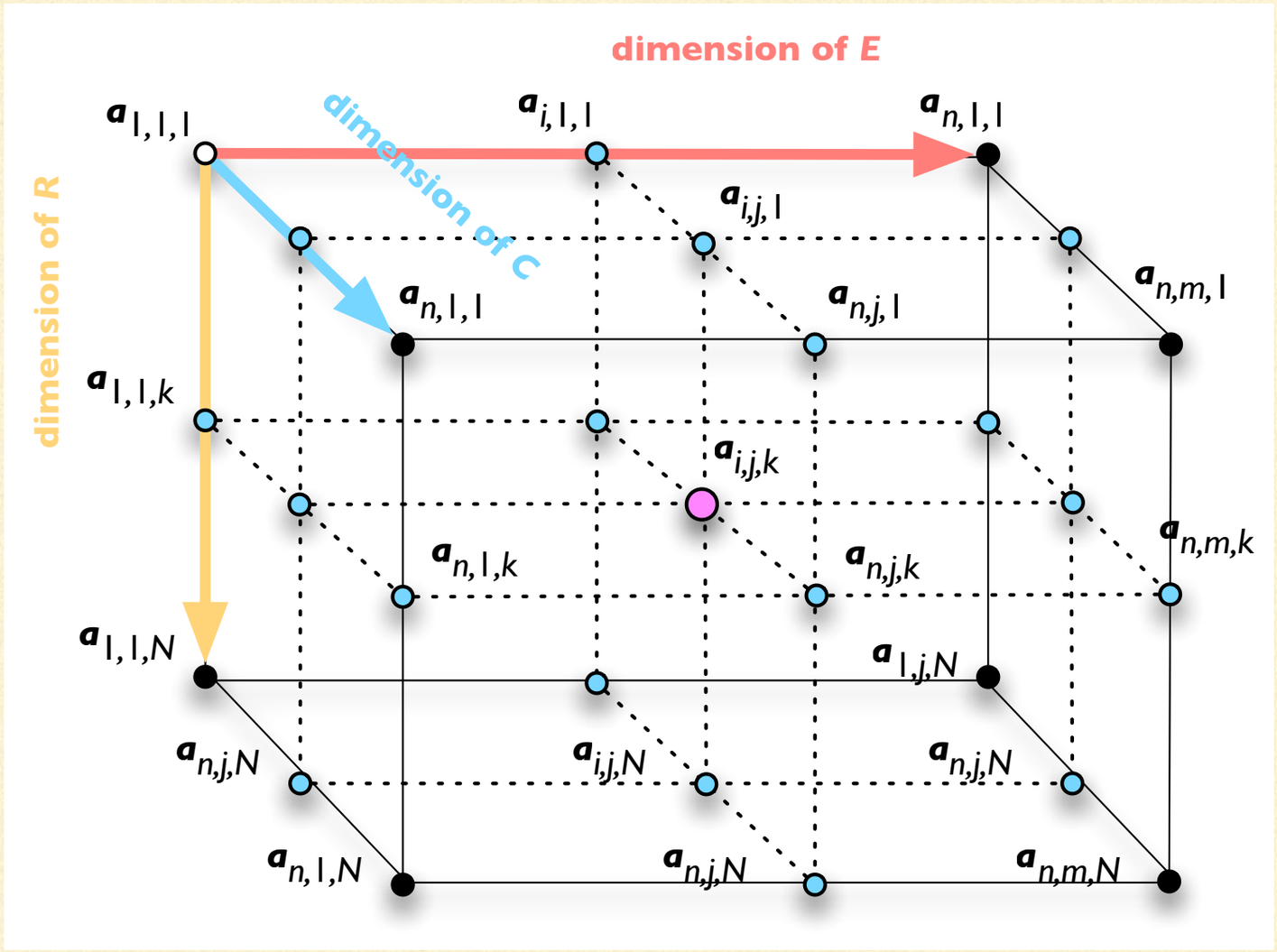
Conflation of *A* = *R*x*E*x*C* (Case 3)

$A^{**}_{k,j} = [a_{1,j,k}, a_{2,j,k}, \ldots, a_{n,j,k}]$ (*n* は e のインデックス)

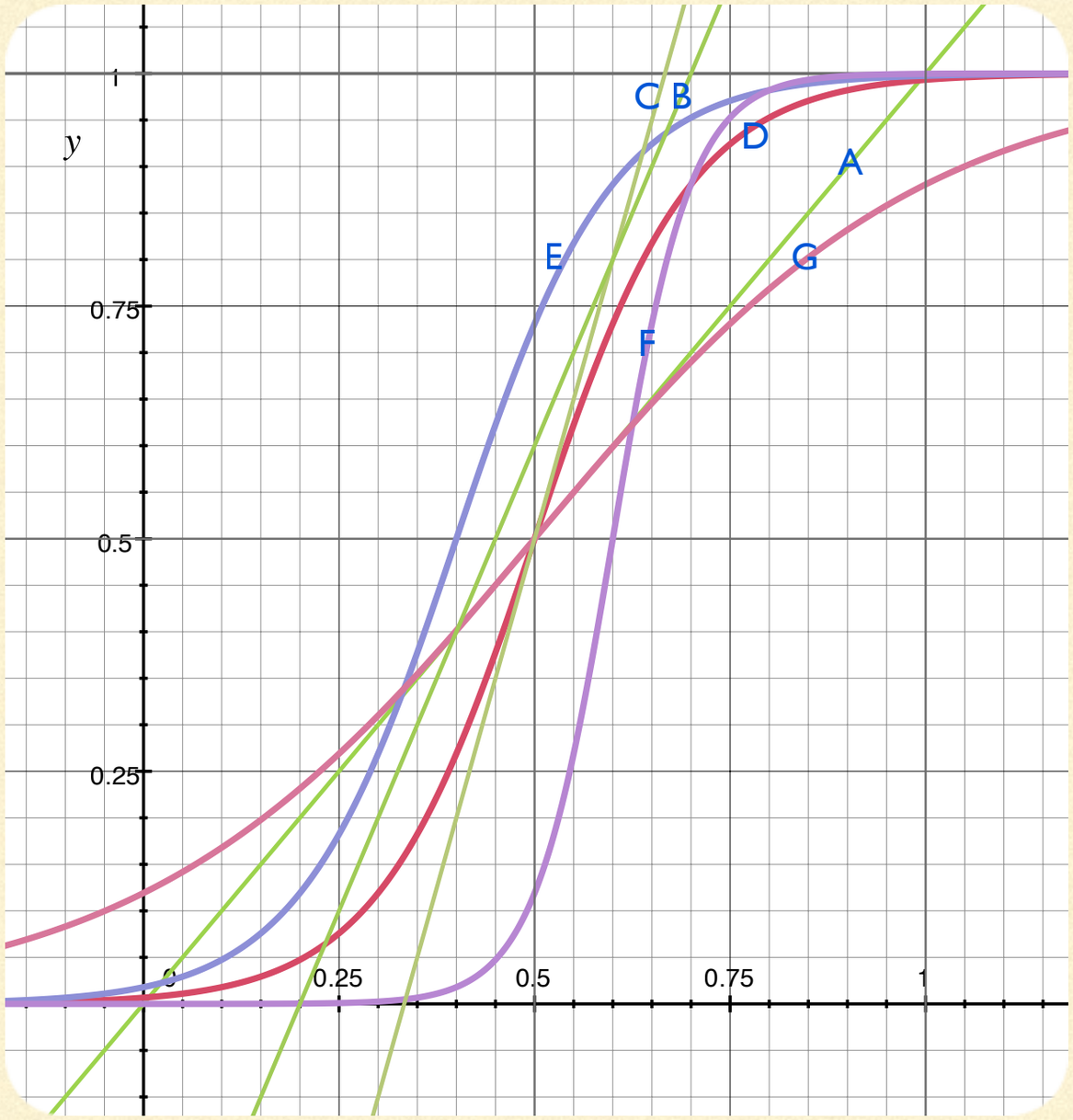
	c	…	c	…	c
r	A	…	A	…	A
⋮	⋮	⋱	⋮	⋱	⋮
r	A	…	A	…	A
⋮	⋮	⋱	⋮	⋱	⋮
r	A	…	A	…	A

TENSOR $A = E \times C \times R$

- Think of a tensor A in which
  - horizontality, width, encodes dimension E, vermilion.
  - distance, or (counter-)depth, encodes dimension C, sky blue, and
  - verticality, or (counter-)height, encodes dimension R, orange.



R の評価バイアスの記述 (グラフは再掲)



- 線型関数族 $y = ax - b$ (ただし $y < 0$ では $y = 0$ で $y > 1$ では $y = 1$) と対比
  - A: $y = x$; B: $y = 2x - 0.4$; C: $y = 3x - 1$
- 相転移的変化を表わす Sigmoidal 関数族 $y = 1/(1 + \exp(a(b - x)))$
  - D: $y = 1/(1 + \exp(10\ (0.4 - x)))$
  - E: $y = 1/(1 + \exp(10\ (0.5 - x)))$
  - F: $y = 1/(1 + \exp(20\ (0.6 - x)))$
  - G: $y = 1/(1 + \exp(\ 4\ (0.5 - x)))$
- 連続性と見かけカテゴリカルな性質は矛盾しない

見取り図

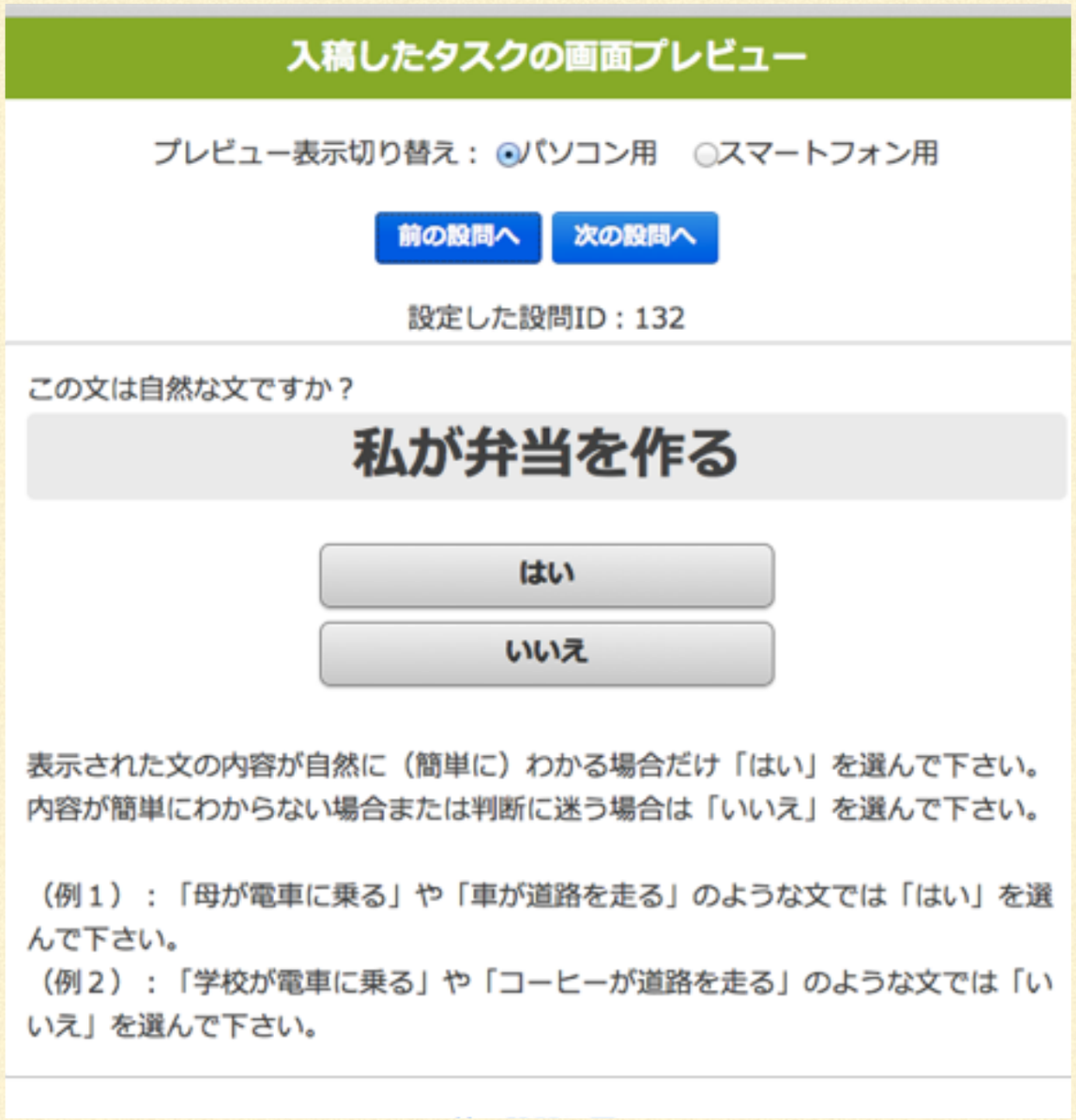
- 以下で，***C*** が暗黙次元である Case 1: ***E*** x ***R*** と Case 2: R x ***E*** の関係を実例で示す

S と *R* の分類の実例

仲村-河原 (2015) の実験の設定 1/2

- Yahoo! クラウドで1945種類の文を，容認できる/できないの2値で判定
  - s1, s3, …, s1935 が実験用
    - d1, d2, …, d10 が検査用
      - d1,…d5 は誰もが容認可能 [1]
      - d6,…d10 は誰もが容認不可能 [0]
    - 刺激文は格フレーム辞書 (河原・黒橋 '06) を元に作成
- 1人当り88.9例で，99例から10例の範囲
- 1文を平均で 9.131 名が評定
- 評定者: 244名
  - 次の2条件で (予稿の191名でなく) 185名に限定
    - 容認率 r: $0.15 < r < 0.85$
    - 評定事例数 n: $80 < n$

仲村-河原 (2015) の実験の設定 2/2



- この文は自然な文ですか？
  - 刺激文
- 選択肢
  - < はい >
  - < いいえ >
- データ処理
  - [はい, いいえ] を [1, 0] に変換

確認用の文 D1–D10

R.COUNT	D-ID	S	RATING
190	1	太郎が学校に行く	1
250	2	太郎がコーヒーを飲む	1
190	3	太郎が本を読む	1
300	4	太郎が自転車に乗る	1
230	5	太郎がボールを投げる	1
200	6	太郎が花を読む	0
160	7	太郎が電車を食べる	0
230	8	太郎が恐竜を寝る	0
160	9	太郎がパソコンを泣く	0
240	10	太郎が黒板を運転する	0

提示文の例 (無作為抽出の10例)

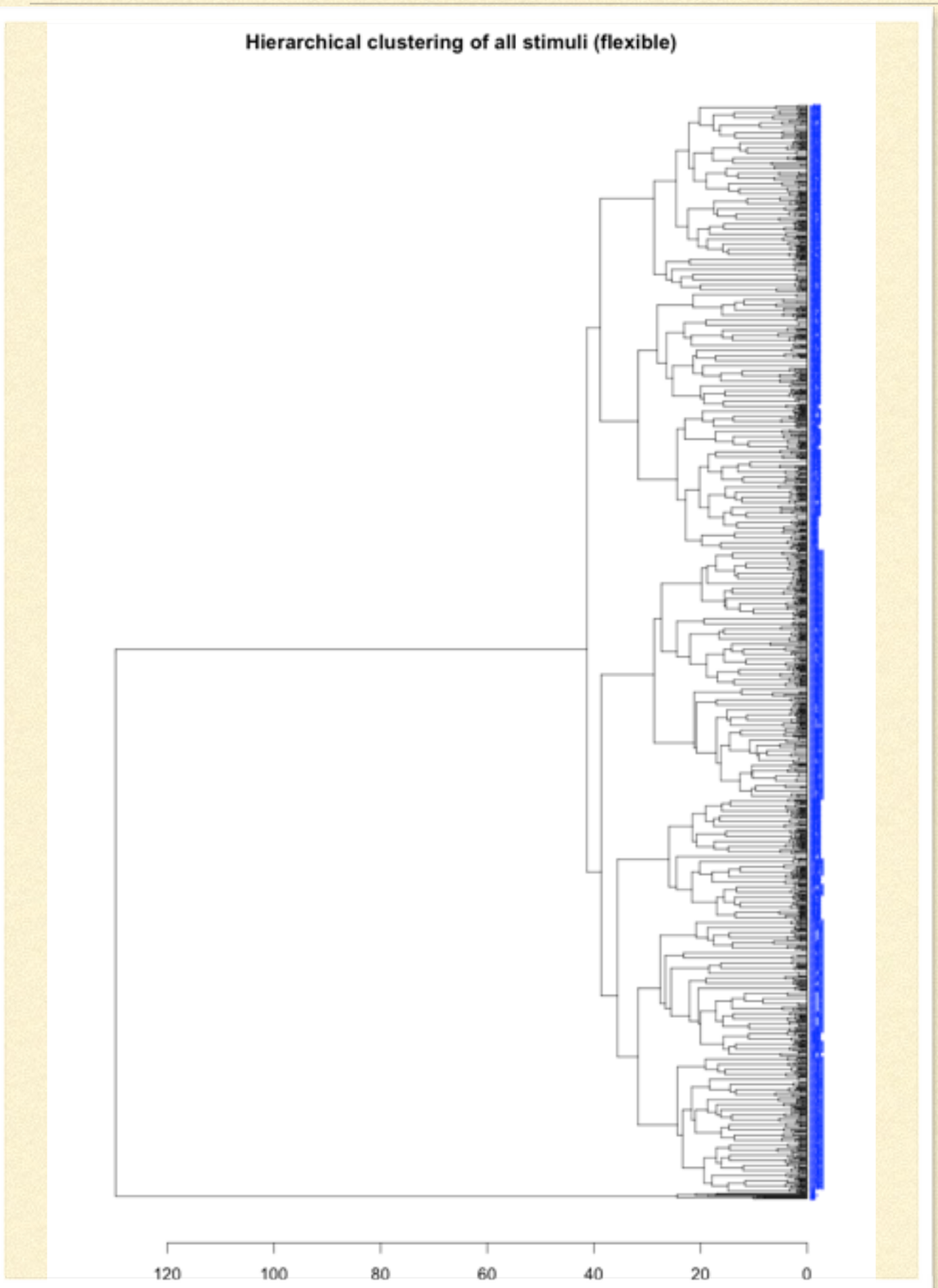
S.COUNT	S-ID	S
10	5	人が出会いを頂く
10	227	人が下りるを待つ
10	442	太陽が顔を出す
10	617	対戦がレースを楽しめる
10	722	署員が煎るを見付ける
10	761	グループが結果を表に纏める
10	1203	人が予備軍を含める
10	1695	人が日祝を除く
10	1757	人が沢を詰める
10	1857	ペンギンが空を実際に飛ぶ

容認度の違いで文は分類できるか？

S の分析

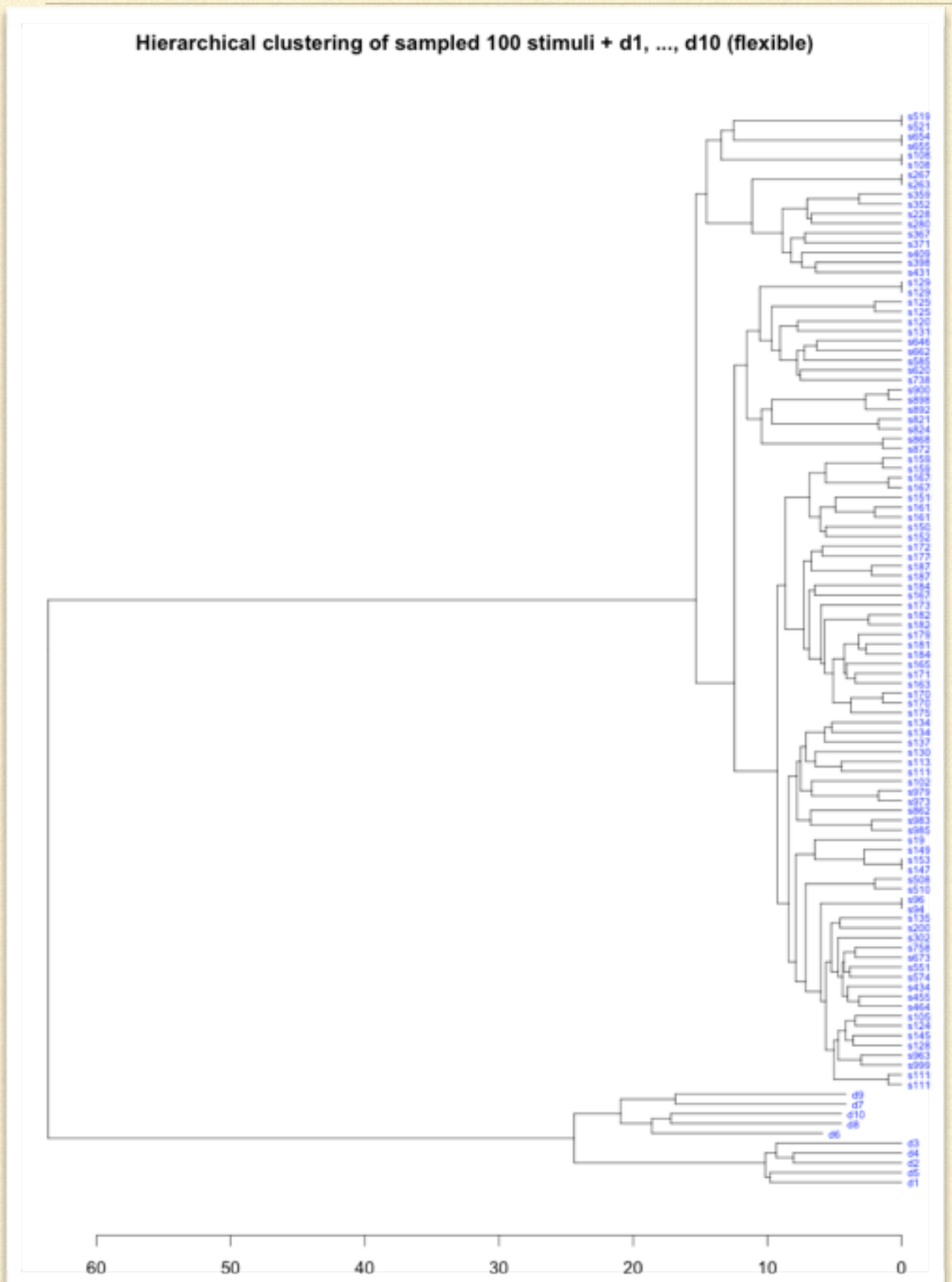
解析法に関する注意

- 設定
  - どの事例にも，評定者の一部 (10名~100名) しか評定値を与えていない
    - 要するにデータは欠損値 (missing values) だらけ
- この欠損値の問題を解決するため，以下の解析では，欠損値ありクラスタリングという手法を使っています
  - R パッケージの cluster (Maechler, *et al.* 2015) が実装
- この解析の結果では
  - 全体の構造の保存は保証されているものの，
  - 枝葉末節の記述は信頼できない
    - 撹乱の影響を受ける
  - 特に個々の事例の帰属するクラスターと，他の事例との距離は (真の帰属や距離に対して) 不正確



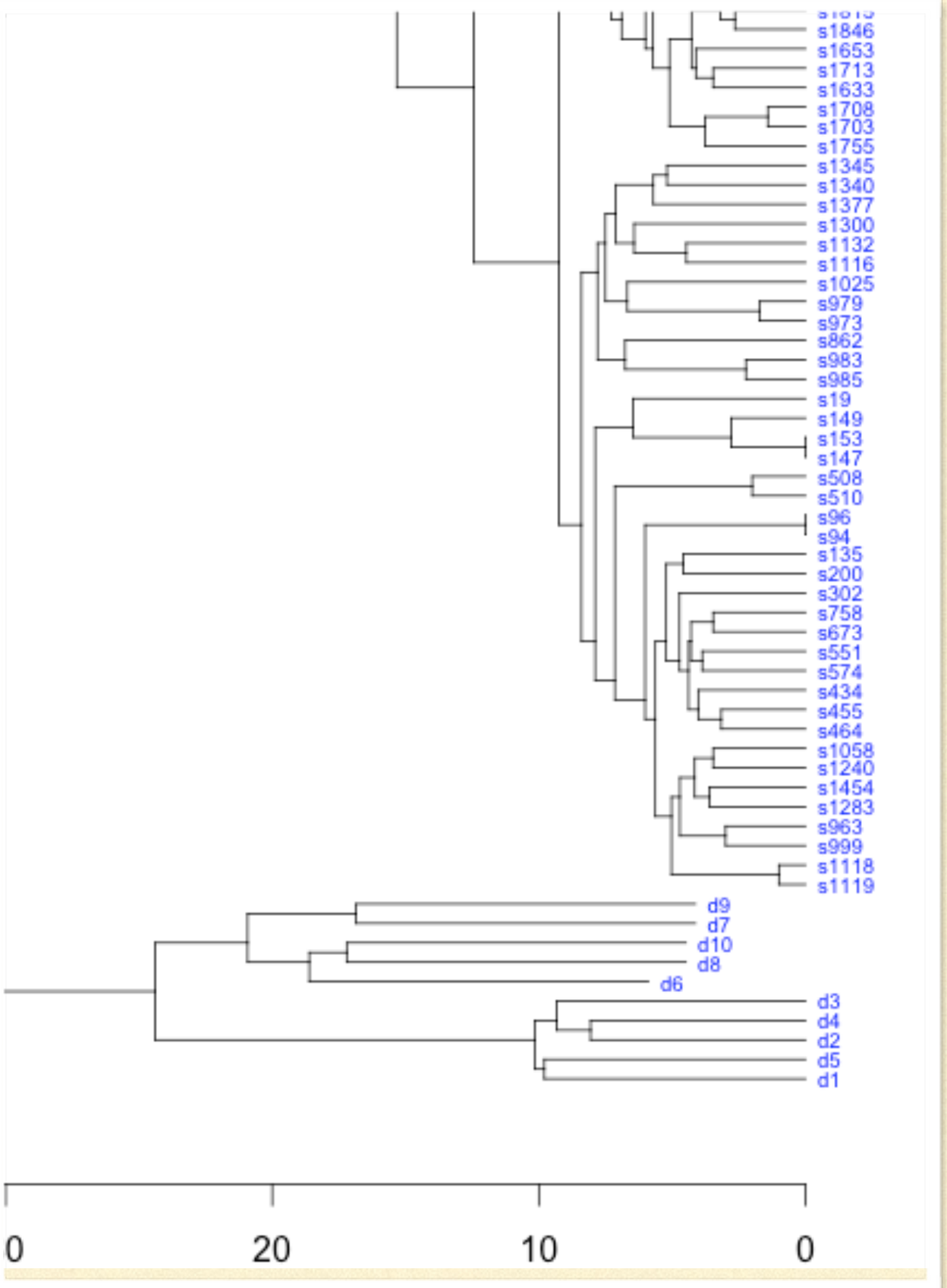
S の階層クラスタリング (全事例)

- 手法
  - s1, …, s1935 + d1, …, d10 の事例を
  - R パッケージの cluster が提供する “欠損値ありクラスタリング” (Maechler, *et al.* 2015) で解析
- 結果
  - d1, …, d10 は完全に分離



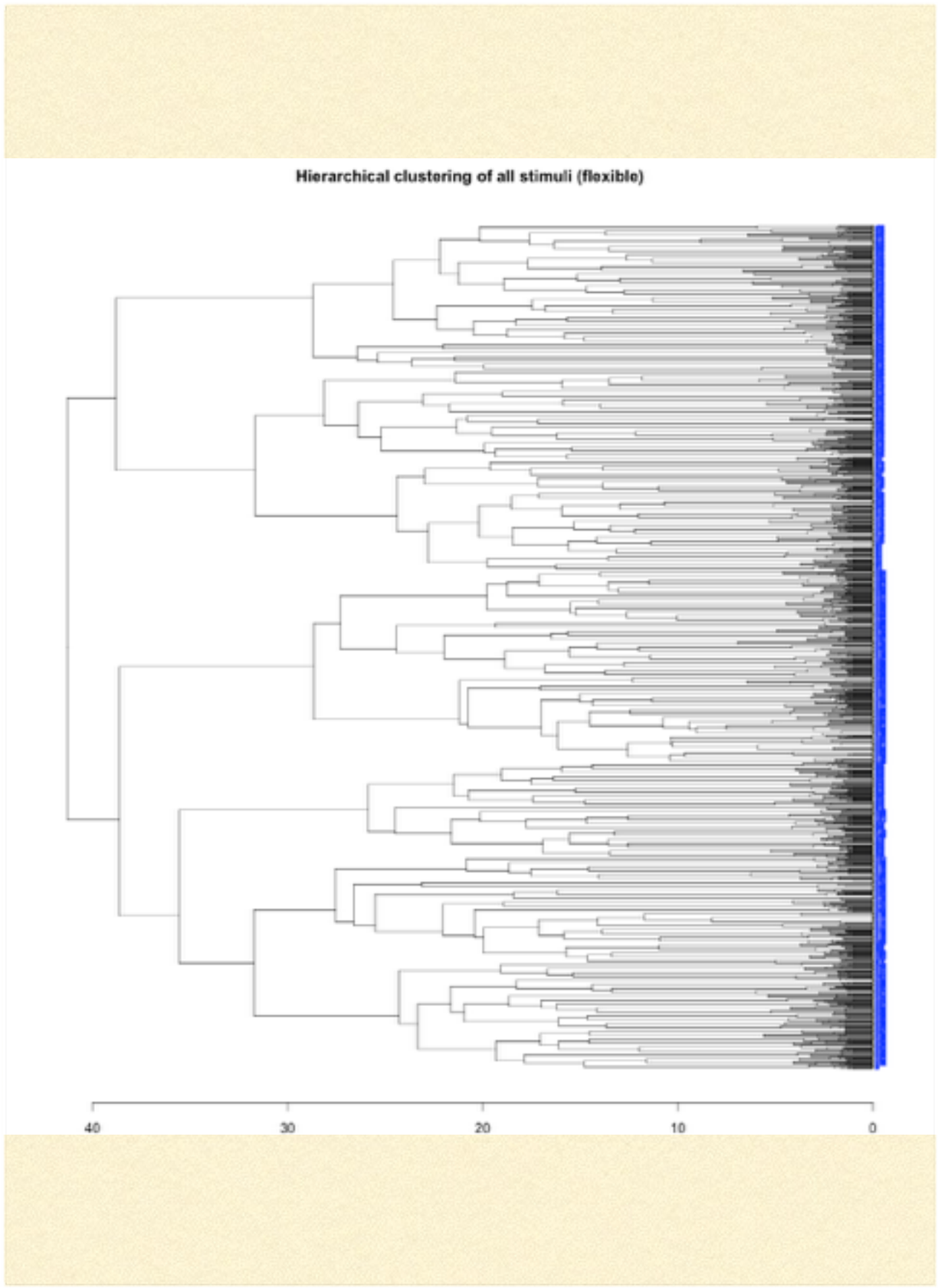
S の階層クラスタリング (全事例 SAMPLED)

- 手法
  - s1, …, s1935 から無作為抽出した100事例に d1, …, d10 を加えて
    - 欠損値ありクラスタリング
- 結果
  - {d1, …, d5} と {d6, …, d10} が完全に分離し異質だとわかる



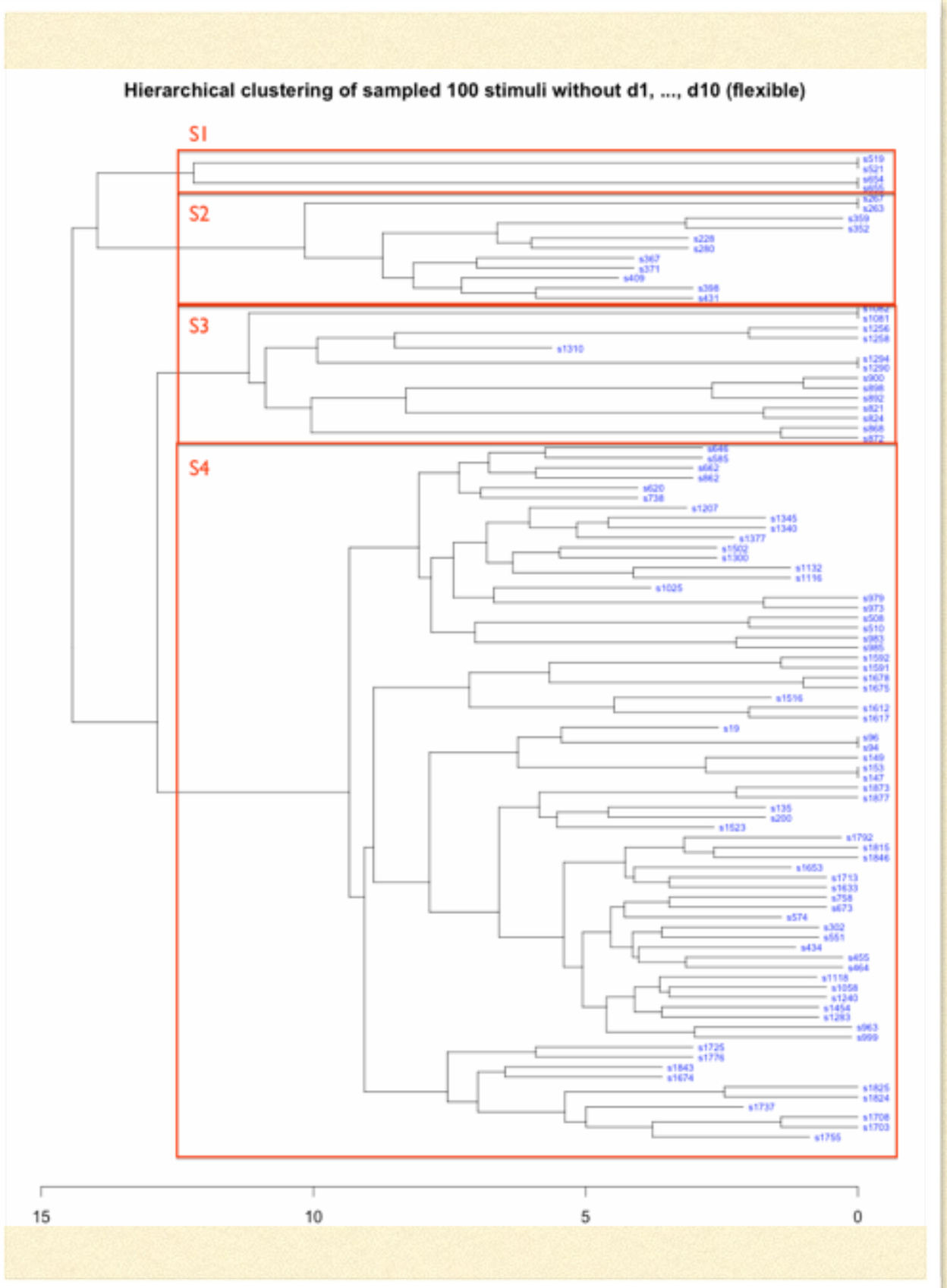
S の階層クラスタリング (全事例 SAMPLED)の拡大図

- 手法
  - s1, …, s1935 から無作為抽出した 100個に d1, …, d10 を加えて欠損値ありクラスタリング
- 結果
  - {d1, …, d5} と {d6, …, d10} が完全に分離し異質だとわかる



S の階層クラスタリング (sNのみ)

- 手法
  - s1, …, s1935 の全事例を “欠損値ありクラスタリング” (Maechler, *et al.* 2015) で解析
- 結果
  - 大きくは G*1 = {C*1, C*2} と G*2 = {C*3, C*4} の二つに
  - 細かくは C*1, C*2, C*3, C*4, C*5 に分離



S の階層クラスタリング (sNのみ SAMPLED)

- 手法
  - s1, …, s1935 から 100個を無作為抽出し，欠損値ありクラスタリング
- 結果
  - S1, S2, S3, S4 の4つ
  - あるいは {S1, S2} と {S3, S4} の2つ

SAMPLE 100 の帰属

- S1
  - s519, s521, s654, s655
- S2
  - s267, s263, s359, s352, s228, s280, s367, s371, s409, s398, s431
- S3
  - s1082, s1081, s1256, s1258, s1310, s1294, s1290, s900, s898, s892, s821, s824, s868, s872
- S4
  - s646, s585, s662, s862, s620, s738, s1207, s1345, s1340, s1377, s1502, s1300, s1132, s1116, s1025, s979, s973, s508, …, s1755

SIの事例

SID	S	Rating Average	Rating Stdev
s519	選手が監督を有るにやる	0.00	0
s521	プロが試合をやるにやる	0.00	0
s654	数々が両方を楽しめる	0.00	0
s655	引きが展開を楽しめる	0.00	0

S2の事例

SID	S	Rating Average	Rating Stdev
s267	情報が改札を画面に出る	0.00	0
s263	記事が出口を上に出る	0.00	0
s359	人がモーツァルトを聞く	1.00	0
s352	人が期間を入れる	0.00	0
s228	人が伸びるを待つ	0.10	0.316227766
s280	行動が域を裏目に出る	0.00	0
s367	人がシリーズを含む	0.00	0
s371	人がカードを含む	0.00	0
s409	人がフォームを含む	0.00	0
s398	人が効果を含む	0.00	0
s431	人が円を掛ける	0.10	0.316227766

S3の事例

SID	S	Rating Average	Rating Stdev
s1082	金が頭をところに使える	0.00	0
s1081	金が頭を使える	0.00	0
s1256	道路が下を真っ直ぐに走る	0.50	0.534522484
s1258	私がトップを主義に走る	0.00	0
s1310	激震が世界を業界に走る	0.00	0
s1294	風が海岸を走りに走る	0.00	0
s1290	ランナーがメートルをトイレに走る	0.00	0
s900	比率が%を超える	0.13	0.353553391
s898	者数が人を超える	0.00	0
s892	人がイヴを過ごす	0.75	0.46291005
s821	人が出汁汁を加える	0.38	0.517549117
s824	客が同店を店に訪れる	0.00	0
s868	人がブラックを調べる	0.33	0.5
s872	人が含むを調べる	0.11	0.333333333

S4の事例 1/2

SID	S	Rating Average	Rating Stdev
s646	海水浴がサーフィンを楽しめる	0.11	0.333333333
s585	人が有るを学ぶ	0.00	0
s662	コントラストが効果を楽しめる	0.33	0.5
s862	センサーが無礼を働く	0.29	0.487950036
s620	こなしがスタイルを楽しめる	0.11	0.333333333
s738	雪解けが距離を進む	0.00	0
⋮	⋮		
s1025	人が臨席を賜る	0.00	0
s979	人が数式を用いる	1.00	0
s973	人が計算を用いる	0.70	0.483045892
s508	社員が仕事をやる	0.78	0.440958552
s510	人が役をやる	0.78	0.440958552
⋮	⋮		
s19	人が土日を使う	0.14	0.377964473
s96	妻が小言を親に言う	1.00	0
s94	妻が小言を言う	1.00	0
⋮	⋮		

S4の事例 2/2

SID	S	Rating Average	Rating Stdev
s1725	人が食べるを見掛ける	0.11	0.333333333
s1776	アルバムがＣＤを作れる	0.25	0.46291005
s1843	コントラストが効果を楽しめる	0.14	0.377964473
s1674	ヒントが回答を貰える	0.50	0.527046277
s1825	資格が元気を貰える	0.25	0.46291005
s1824	人が金を貰える	1.00	0
s1737	人が気付きを貰う	0.44	0.527046277
s1708	人がリストを眺める	1.00	0
s1703	人が短縮化を図る	0.71	0.487950036
s1755	私があなたを殺す	1.00	0

知見

- [1] 述語が同一だと，容認度評定の結果が似る？
  - 順序効果の副作用である可能性が大きい
    - Yahoo! クラウドの仕様 (何と提示順序の無作為化をしない！) が原因らしい
  - 評定者のグループが共有している効果を差し引いてこれが言えたら，興味深い
- [2] S4 を中心とする外心構造がありそう
  - 容認度が下がるに応じて S4 ⇒ S3 ⇒ S2 ⇒ S1 と広がっている
- [3] 逸脱の生じる原因の違いによって反応が異なる
  - S1: 余計な語彙要素がある場合
  - S2: 選択制限の違反がある場合
  - S3: 必要な語彙要素がない場合

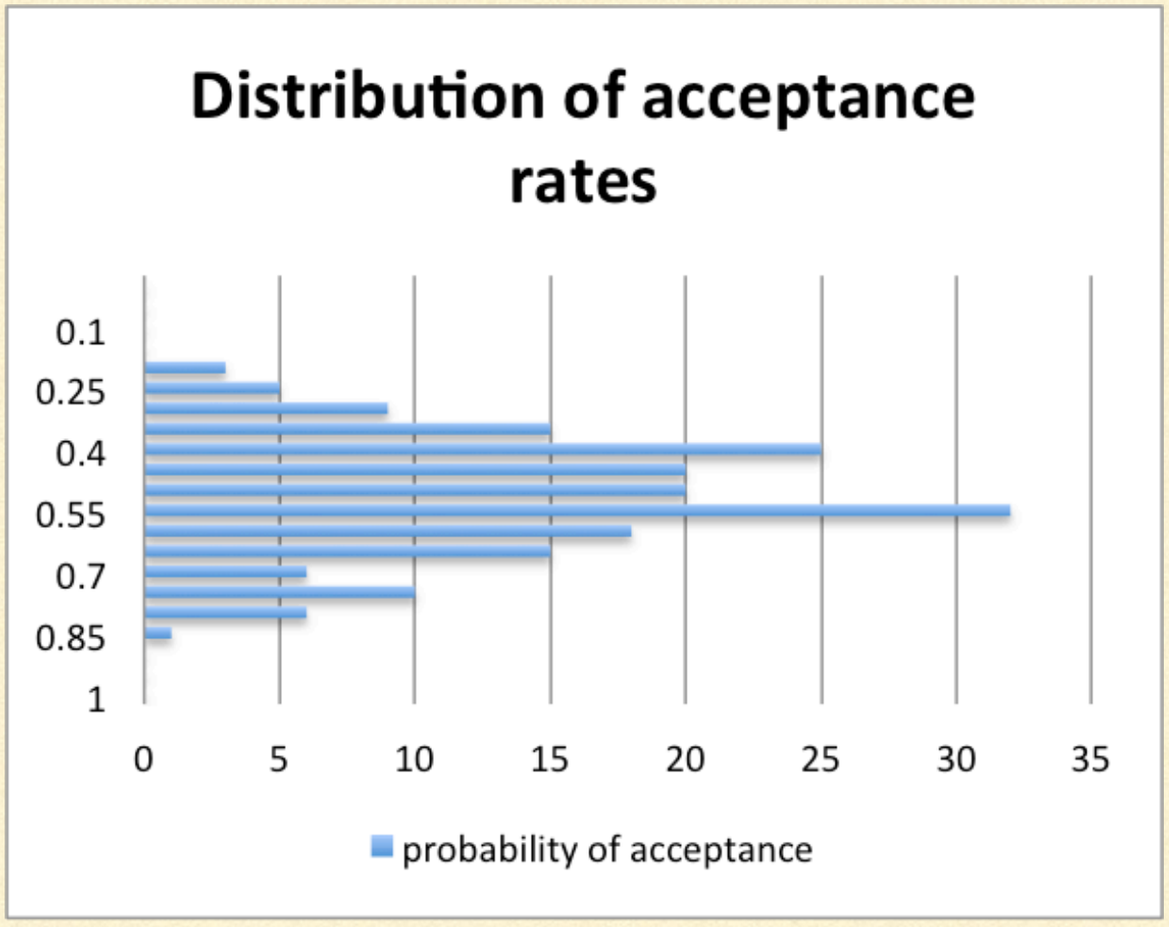
確認できる事

- 提示文=評定対象は一般に
  - A. 異なる評定値の平均値と異なるばらつきをもち
    - A′. d1, …, d10 のような特殊事例=極端な事例を除くと，“正しい反応” と言うものはない
  - B. それを基準にしてクラス分け=タイプ分けが可能で
  - C. その要因は明らかにR の反応の違い
- である
  - A′ にどう対処するか？が早急の課題

容認度の違いで評定者は分類できるか？

R の分類

容認率の分布

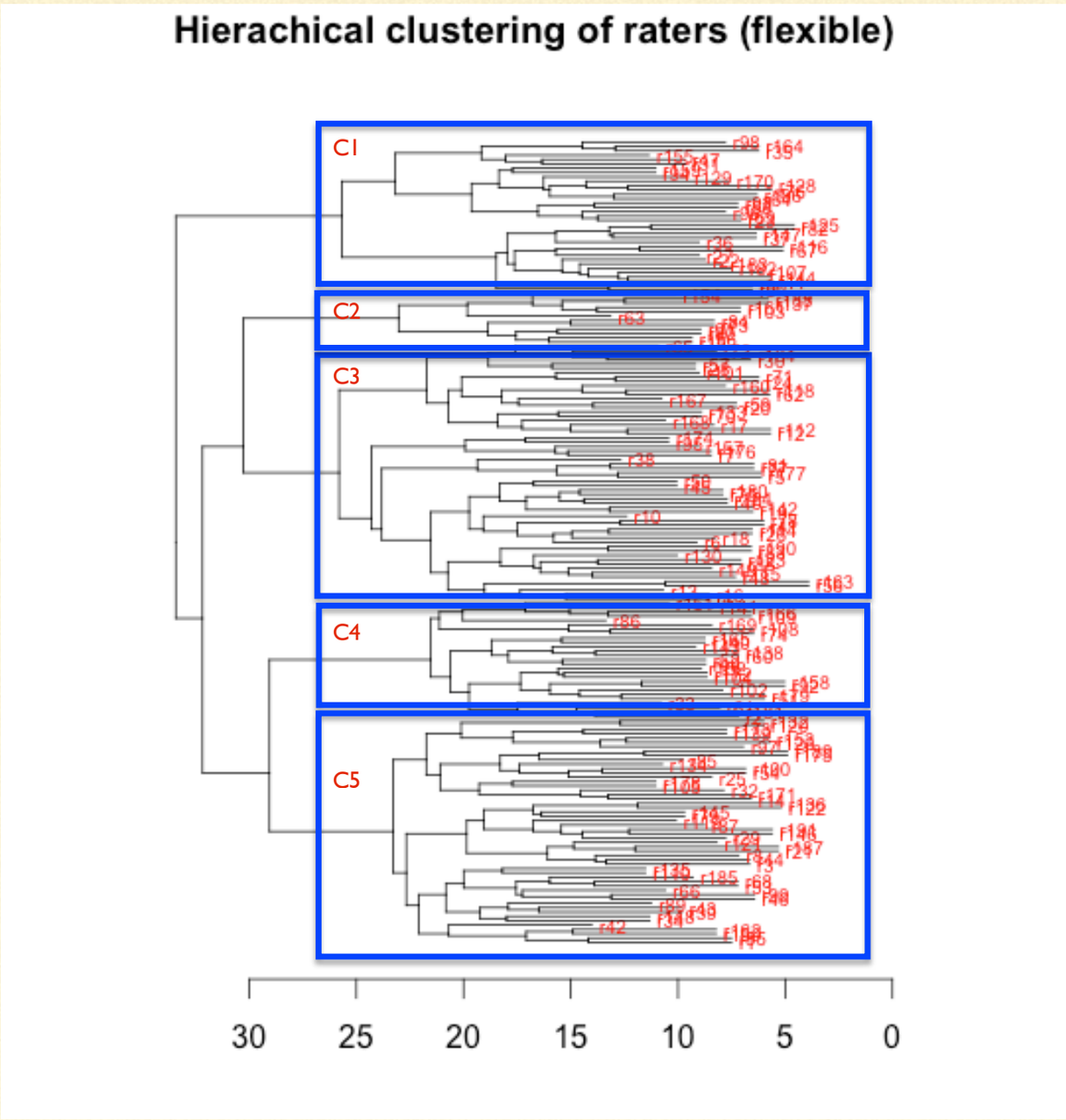


- 185名の評定者の容認度の分布
  - Max: 0.84; Min: 0.16
  - Average: 0.49; Median: 0.48
  - Stdev: 0.14
- 知見
  - 容認度の分布がほぼ Gauss 分布になっているのは予想外
    - 正規性の判定は未実行

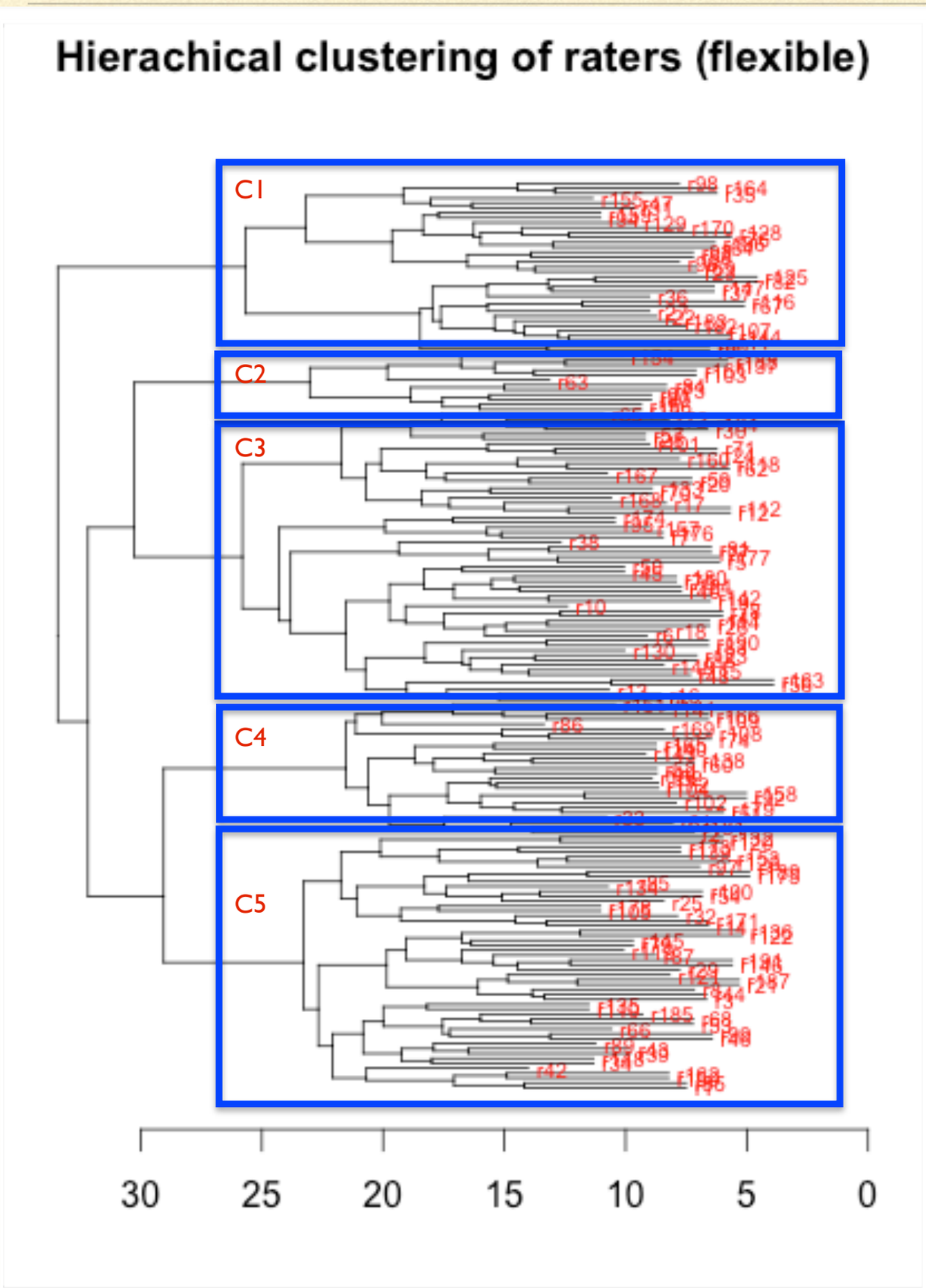
R の前処理

- フィルター1
  - 0.86 より大きな容認率か
  - 0.15 より小さな容認率をもつ評定者を排除
- フィルター2
  - 個数事例数が 80 より少ない評定者を除外
    - 最大値は 99個
- 以上の二重のフィルターで185名に

容認パターンの階層クラスタリング



- R の階層クラスタリングの例 (*n*=185)
  - 予稿の *n*=191 は誤り
  - 属性は s1, …, s1935, d1, …, d10 の評定値から成る 0/1のベクトル
- グループ=クラスター数
  - 欲張れば C_1, C_2, C_3, C_4, C_5 の五つ
  - 控え目なら $C_1, \{C_2, C_3\}, \{C_4, C_5]$ の三つ



結果の拡大表示

- 欠損値ありクラスタリングなので，枝葉末節の位置/区別は意味を持たない
  - ランダムに位置が撹乱される

CLUSTER MEMBERS

- C1 [35]:
  - r98, r164, r35, r155, r47, r31, r159, r94, r129, r170, r128, r76, r186, r64, r93, r58, r96, r69, r23, r125, r82, r147, r37, r36, r116, r67, r27, r22, r183, r182, r107, r144, r11, r90, r9
- C2 [12]
  - r154, r188, r137, r161, r103, r63, r84, r73, r91, r80, r156, r15
- C3 [59]
  - r65, r72, r184, r30, r57, r28, r101, r71, r24, r160, r118, r62, r167, r59, r20, r133, r70, r168, r17, r112, r12, r174, r95, r157, r176, r7, r38, r81, r77, r177, r5, r50, r45, r180, r75, r124, r40, r142, r19, r10, r78, r41, r114, r26, r18, r6, r190, r83, r130, r123, r88, r149, r115, r48, r163, r56, r13, r16, r4
- C4 [27]
  - r151, r141, r166, r109, r86, r169, r108, r74, r165, r140, r143, r138, r60, r52, r49, r152, r172, r104, r158, r92, r102, r179, r51, r33, r61, r181, r2
- C5 [52]
  - r132, r120, r173, r139, r153, r126, r97, r189, r175, r85, r134, r100, r54, r25, r178, r105, r32, r171, r14, r136, r122, r145, r79, r119, r87, r191, r146, r29, r121, r187, r21, r8, r44, r3, r135, r110, r185, r68, r53, r66, r99, r46, r89, r43, r39, r148, r34, r42, r162, r150, r55, r1

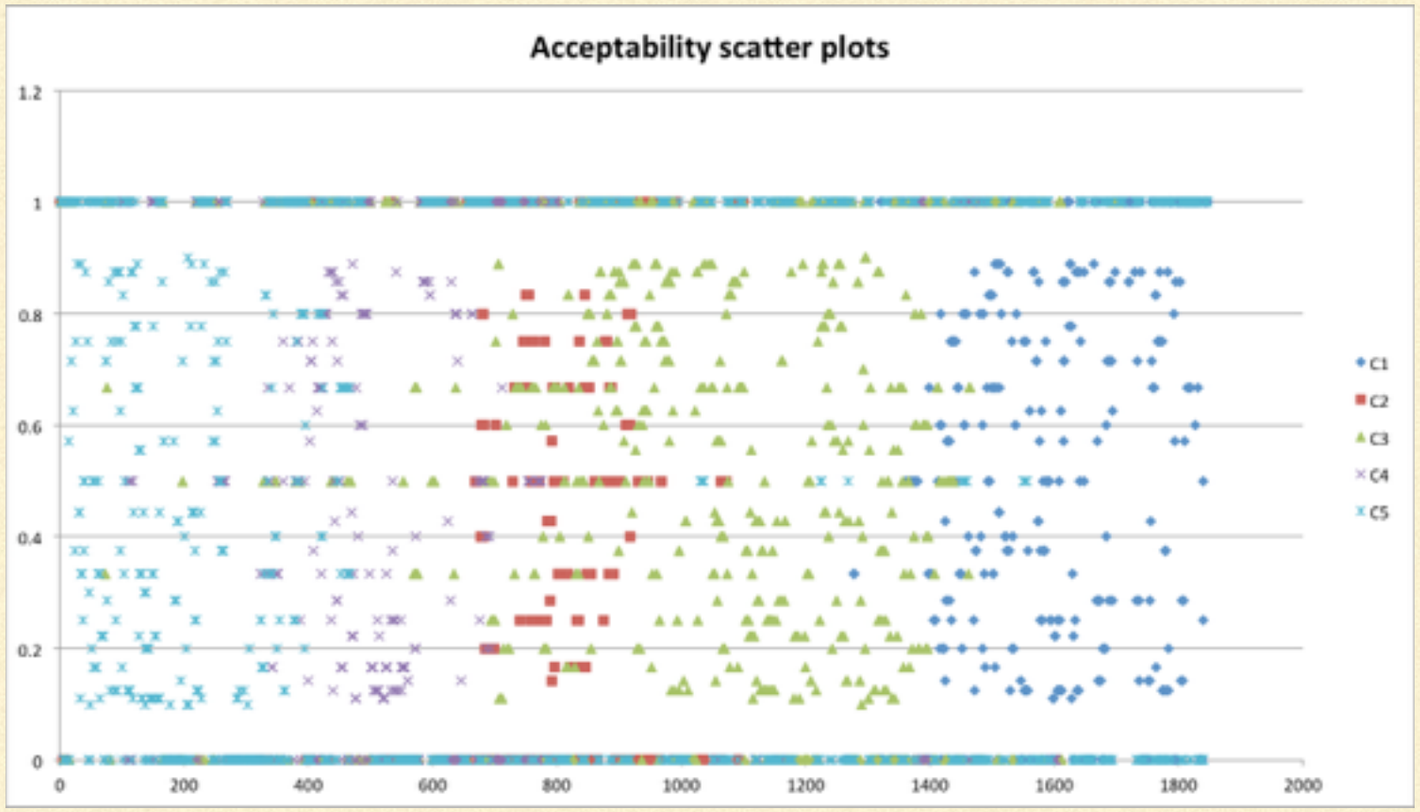
SCATTER PLOT OF AVERAGED RESPONSES

- 手順
  - クラスター $C_1, \ldots, C_5$ ごとに

    s1, s2, …, s99, d1, d2, …, d9, s100,…, s999, d10, s1000, …, s1935 の評定値の平均を求め

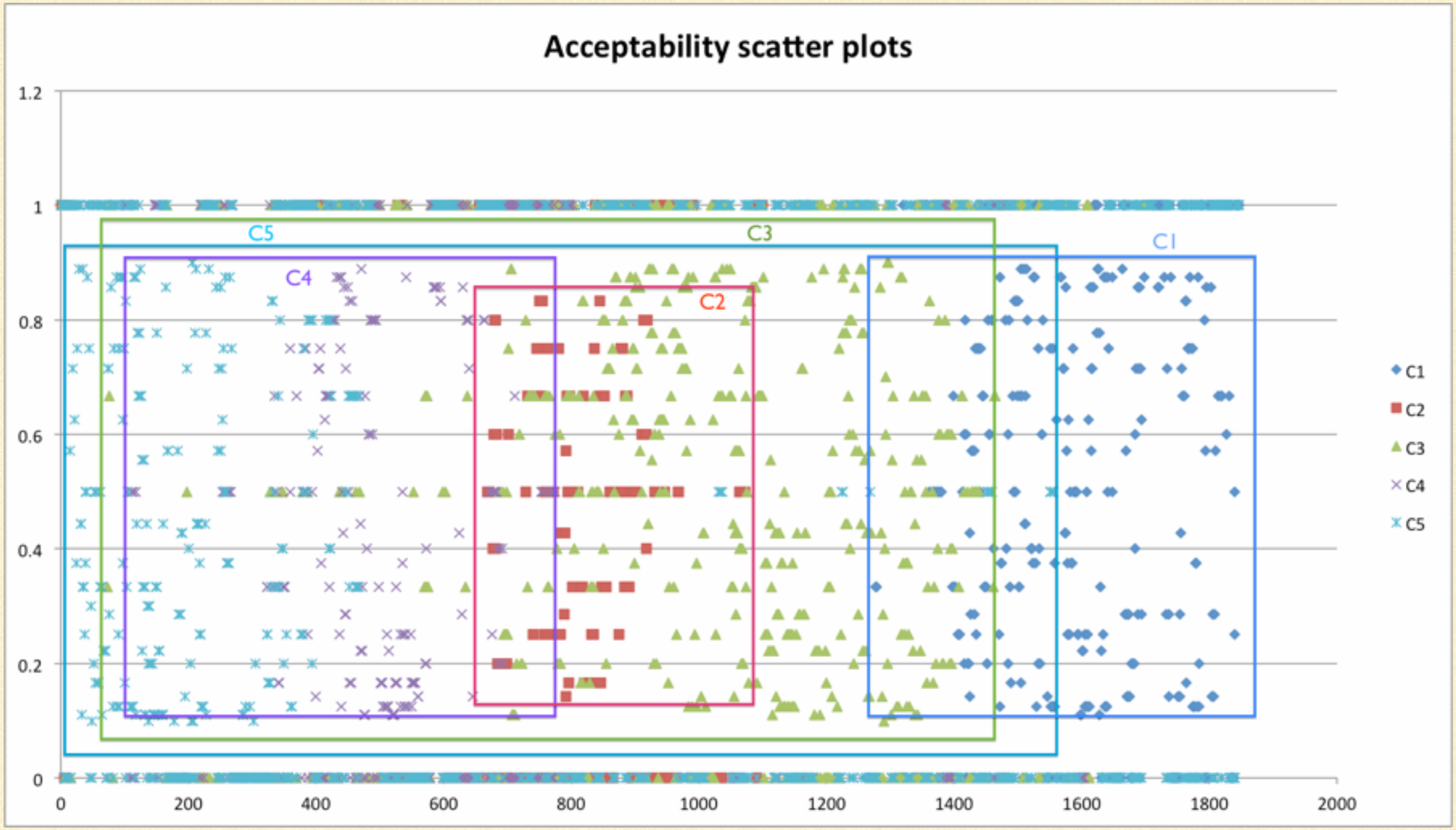
    - 刺激として与えられていない場合は na ⟹ 空欄 とエンコード

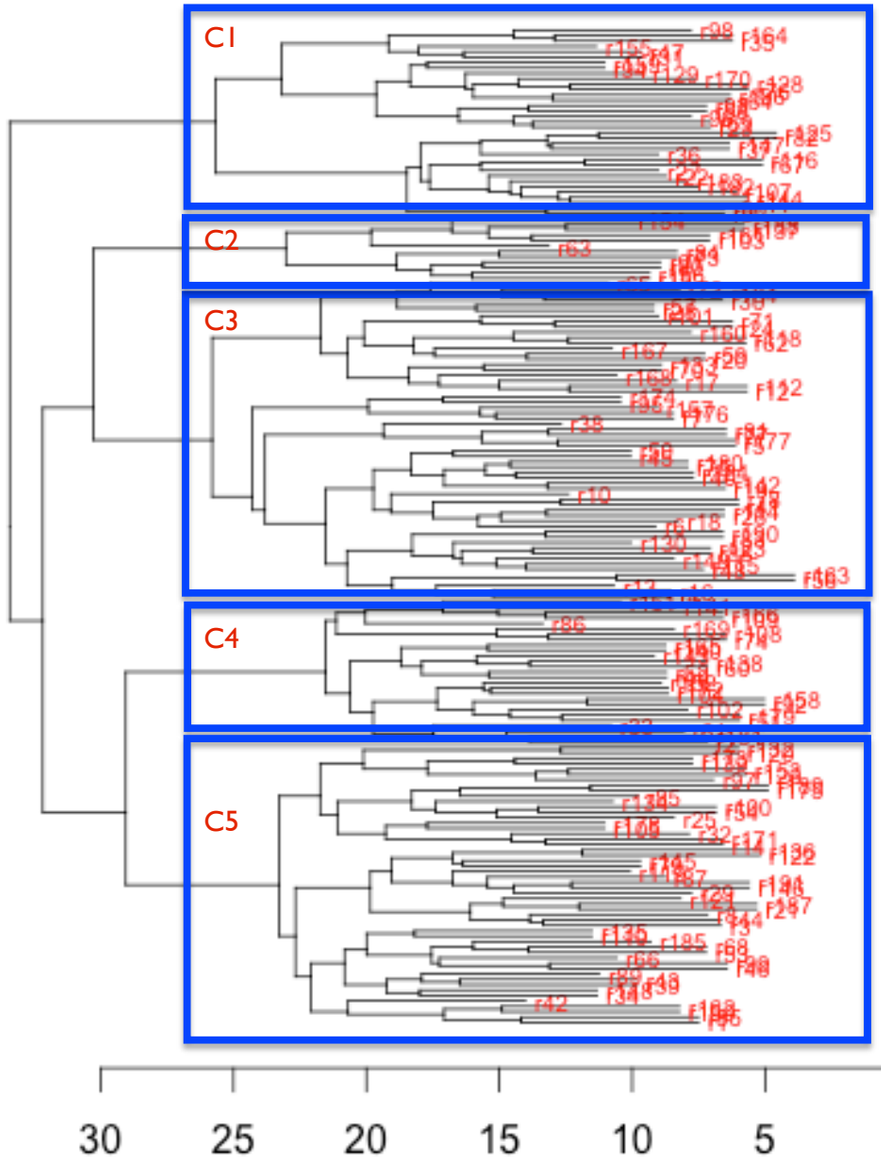
  - その数値列を表現するベクトルとして散布図を描く

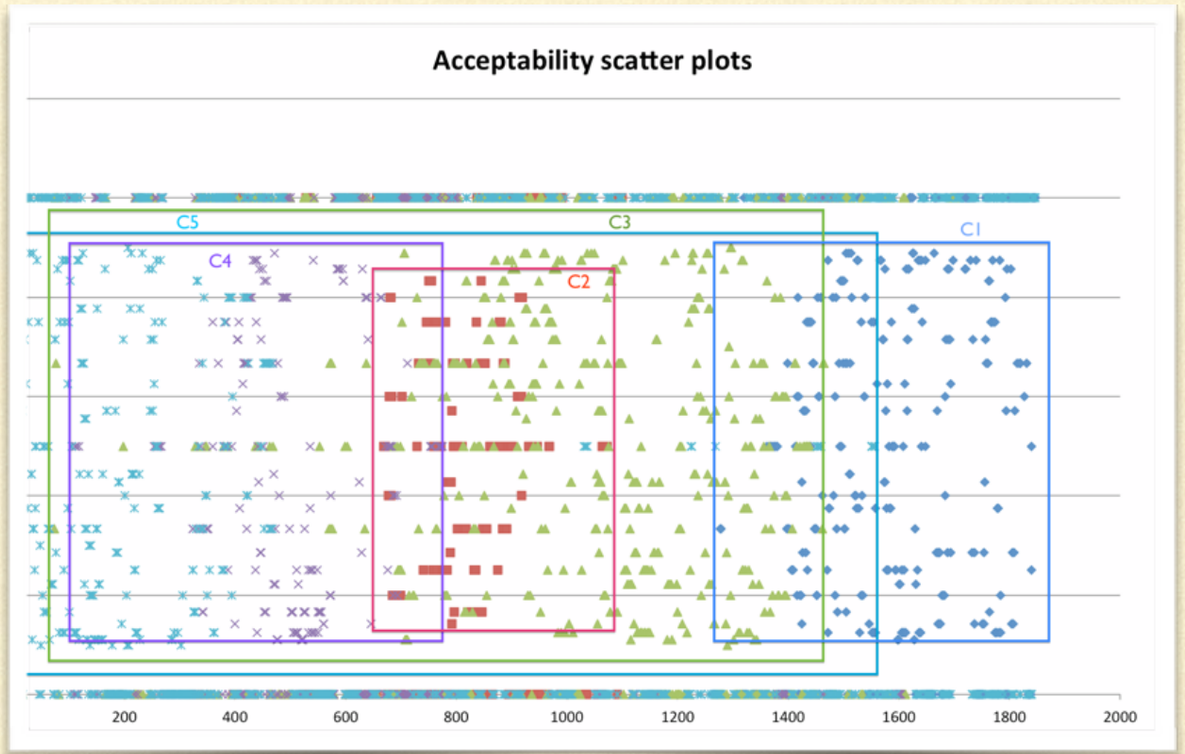


SCATTER PLOT OF AVERAGED RESPONSES

$C_1, C_2, \ldots, C_5$ との対応づけ



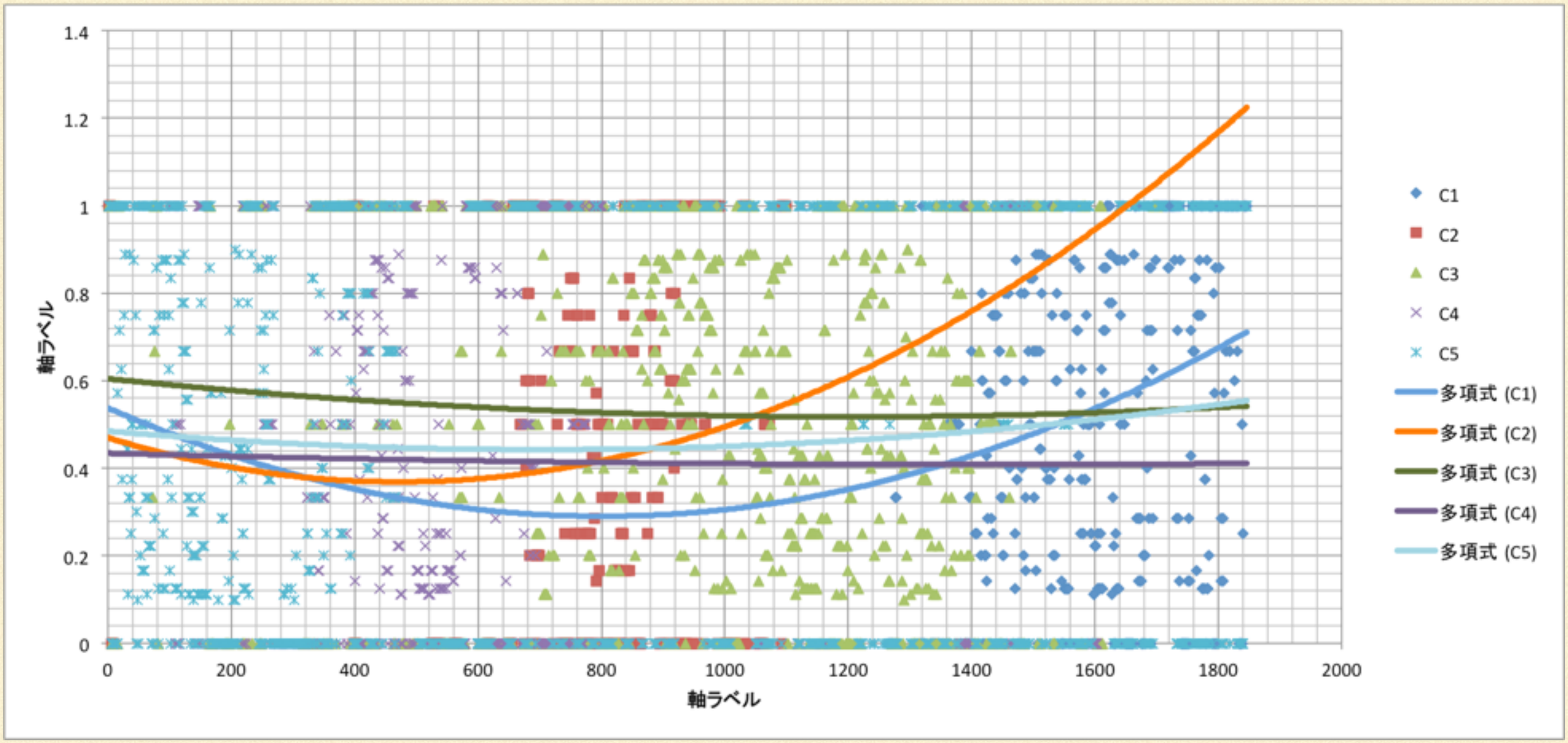




- 散布図と階層クラスタリングの対応
- どうやら順序効果は排除できていないらしい

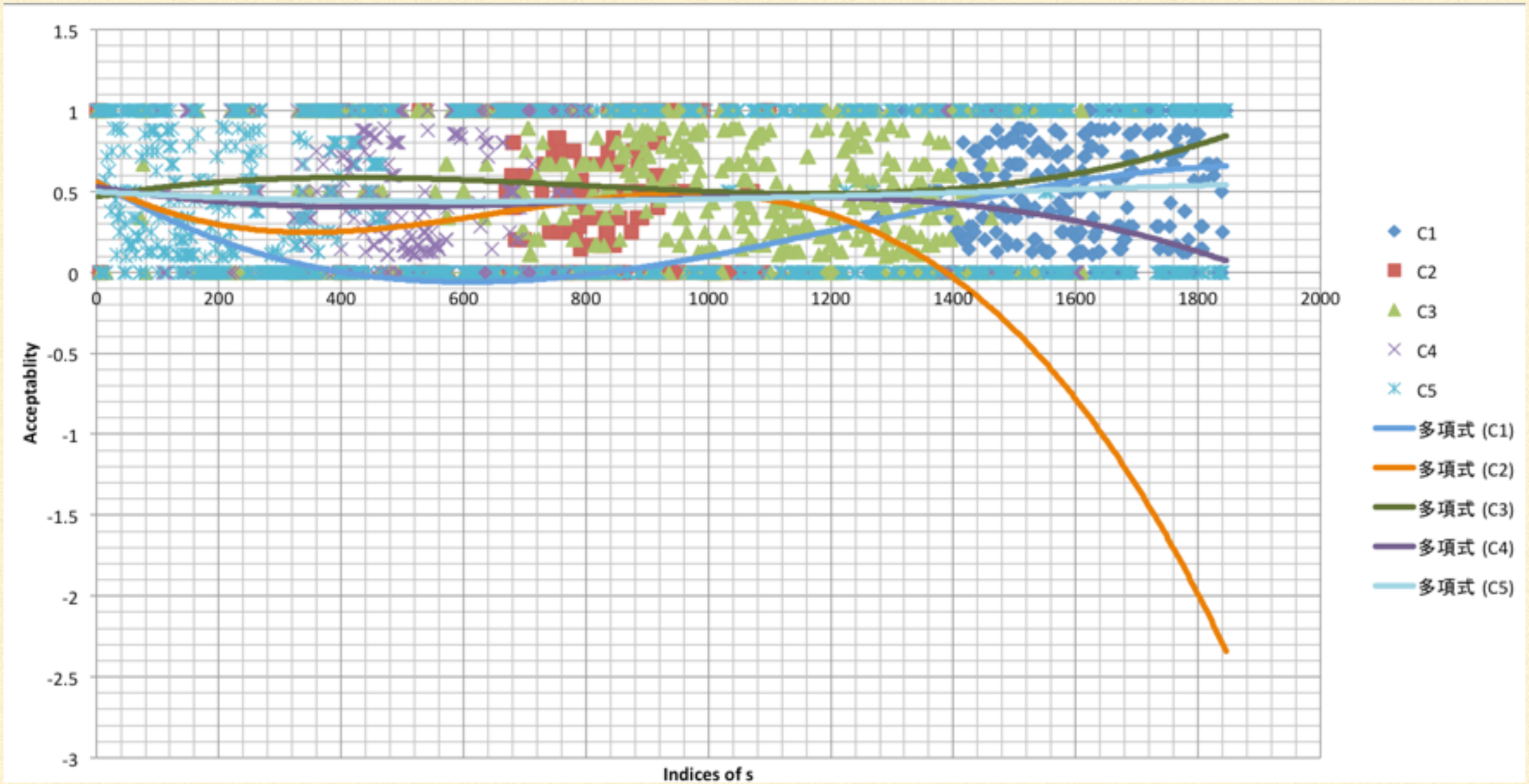
Scatter plot of Averaged responses with Fitting curves (2nd order Polynomial)

Xの並び: s1,…s99,d1,d2,d3,d4,d5,d6,d7,d8,d9,s100,…,s999,d10,s1000,…



Scatter plot of Averaged responses with Fitting curves (3rd order Polynomial)

Xの並び: s1,…s99,d1,d2,d3,d4,d5,d6,d7,d8,d9,s100,…,s999,d10,s1000,…



確認できた事

- 評定者は一般に
  - 異なるバイアスの影響下で評定を行ない，
  - 結果として彼らの与える評定値は異なる分布をもつ
  - それらを基準にしてクラス分け=タイプ分けが可能
- である
- ただし
  - このデータでは，おそらく順序効果の影響が大きくて，他の要因がどれぐらい効いているのか分からない
  - より妥当な結果を得たければ，ちゃんと設計した刺激で実験が必要

まとめ

本発表で示した事 (再掲)

- 論点 I
  - 評定者は（理論言語学の素朴な想定に反して）容認度評定で使われる刺激に一様な応答をしない.
- 論点 2
  - かと言って，彼らの応答はカオス的な訳ではなく，特徴的な応答パターンを幾つか認める事ができる.
  - 別の言い方をすれば，評定者は特定の応答パターンを内在化させている.
- 注意
  - 使用データが別目的に設計されたデータであるため，得られた解析結果は限定的

今後の研究で示すべき事 (再掲)

- 論点 3
  - 特定の応答パターンとは課題に対する個人の最適方略の事であり，この数は有限である.
- 論点 4
  - この意味での特徴的な応答パターン=バイアスの存在を認識し，それに対する補正を施さない限り，容認度評定の結果は内在化されたシステム=言語知識の実体を明らかにするものだとは言えない.
- 論点 5
  - 必要な補正の一つは，容認度評定の結果が一定の仕方でバイアスされた多変量データだと理解する事である.
- 論点 6
  - このような理解がない形で利用された容認度(評定)は言語知識に対する妥当な観察データを与えず，逆に観察的妥当性を歪めるデータとなる.
- 追加の論点7
  - 言語の知識が存在するならば，それは個体群に分散的に体現された集合知である

最後に

- 言語研究の現状で証拠の質を上げるために方法論的に意識すべき事
  - 広義の文脈 *C** の効果は存在するが，その内実を解明しないままで，それを隠れ蓑にするのは，単なる手抜き
- 言語研究の今後の方向性への示唆
  - 文法=言語の知識は集合知であり，個人の知識に帰着し得うるものとは思われない
- 容認度評定で，代表性を有する評定者とそうでない評定者の区別は重要
- 容認度評定の大規模の調査を社会調査として実施する必要性
  - 他の多くの属性 (年齢, 出身地, 教育歴 (特に異国語需用歴), 職業, IQ, …) との相関を見るべき

ついでに宣伝

- 次の研究助成金を元にして，日本語の容認度評定データベース *D* を構築します
  - 言語研究者の容認度評定力の認証システムの試作 : 容認度評定データベースを基礎にして (挑戦的萌芽研究)
  - 研究代表者: 黒田 航; 期間: 2016–2018年度; 課題番号: 16K13223
  - 参加者 (7名)
    - 浅尾 仁彦 (NICT), 阿部 慶賀 (岐阜聖徳学園大学), 金丸 敏幸 (京都大学), 小林 雄一郎 (東洋大学), 田川 拓海 (筑波大学), 土屋 智行 (九州大学), 横野 光 (富士通研究所)
  - *D* の構築は標準化のための調査で，
    - 容認度評定が集合知であると前提にし
    - 刺激文の評定値の平均値をばらつきの大きさを知る
  - 事を目的にしています

発表は以上です

付録

付録 I

容認度評定課題の現実

容認度評定の現実 1/3

“言語表現の容認度とは何か? また何であるべきか? から

- 評定対象の文 (16個)
- F1: *x* が *y* を走る
  - F1a 太郎が 校庭を 走る
  - F1b 稲妻が 北の空を 走る
  - F1c 戦慄が 永田町を 走る
  - F1d 汗が 太郎の額を 走る
- F2: *y* を *x* が走る
  - F2a 校庭を 太郎が 走る
  - F2b 北の空を 稲妻が 走る
  - F2c 永田町を 戦慄が 走る
  - F2d 太郎の額を 汗が 走る
- G1: *x* が *y* に走る
  - G1a 太郎が 校庭に 走る
  - G1b 稲妻が 北の空に 走る
  - G1c 戦慄が 永田町に 走る
  - G1d 汗が 太郎の額に 走る
- G2: *y* に *x* が走る
  - G2a 校庭に 太郎が 走る
  - G2b 北の空に 稲妻が 走る
  - G3c 永田町に 戦慄が 走る
  - G3d 太郎の額に 汗が 走る

容認度評定の現実 2/3

- 評定者
  - $r_1, r_2, \ldots r_{16}$ (自分を含む)
    - 旧京都大学Y研究室の大学院生 (2005年当時)
- 評定基準 (4件法)
  - 3点: まったく違和感を感じない
  - 2点: 軽く違和感を感じるが，言おうとしていることは簡単にわかる
  - 1点: 強く違和感を感じるが，言いたいことがまったくわからないわけではない
  - 0点: 何を言っているかわからないか，明らかに異常なことを言っていると思う

容認度評定の現実 3/3

表 1: *E* を *R* で表現 (*R* と *E* はいずれも平均値の高い順に並べた)

Expression	Ind	r1	r2	r3	r4	r5	r6	r7	r8	r9	r10	r11	r12	r13	r14	r15	r16	av.	stdev
太郎が 校庭を 走る	F1a	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	0.00
永田町に 戦慄が 走る	G2c	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	2.88	0.50
北の空に 稲妻が 走る	G2b	2.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	2.81	0.54
校庭を 太郎が 走る	F2a	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	2.00	3.00	2.00	3.00	2.00	2.00	3.00	2.00	2.63	0.50
北の空を 稲妻が 走る	F2b	3.00	3.00	3.00	3.00	3.00	3.00	2.00	3.00	2.00	2.00	1.00	3.00	2.00	2.00	2.00	3.00	2.50	0.63
稲妻が 北の空を 走る	F1b	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	2.00	2.00	2.00	2.00	1.00	1.00	3.00	2.38	0.81
戦慄が 永田町に 走る	G1c	3.00	2.00	3.00	3.00	3.00	1.00	3.00	1.00	2.00	2.00	1.00	3.00	2.00	1.00	3.00	1.00	2.13	0.89
永田町を 戦慄が 走る	F2c	3.00	2.00	3.00	1.00	1.00	1.00	2.00	3.00	2.00	3.00	3.00	2.00	1.00	2.00	2.00	2.00	2.06	0.77
戦慄が 永田町を 走る	F1c	3.00	3.00	3.00	1.00	2.00	2.00	3.00	3.00	1.00	2.00	1.00	1.00	1.00	1.00	2.00	3.00	2.00	0.89
太郎の額を 汗が 走る	F2d	3.00	3.00	3.00	2.00	3.00	2.00	2.00	2.00	2.00	1.00	1.00	3.00	1.00	1.00	1.00	2.00	2.00	0.82
稲妻が 北の空に 走る	G1b	3.00	3.00	3.00	3.00	3.00	2.00	3.00	1.00	2.00	0.00	2.00	1.00	3.00	1.00	1.00	1.00	2.00	1.03
太郎の額に 汗が 走る	G2d	1.00	3.00	3.00	2.00	3.00	2.00	3.00	3.00	3.00	1.00	2.00	2.00	1.00	2.00	0.00	1.00	2.00	0.97
汗が 太郎の額を 走る	F1d	3.00	3.00	2.00	1.00	3.00	3.00	2.00	1.00	2.00	1.00	2.00	2.00	0.00	1.00	0.00	2.00	1.75	1.00
汗が 太郎の額に 走る	G1d	3.00	2.00	2.00	2.00	3.00	1.00	2.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	1.00	3.00	1.75	0.77
太郎が 校庭に 走る	G1a	3.00	1.00	0.00	1.00	1.00	3.00	2.00	1.00	1.00	0.00	0.00	0.00	1.00	3.00	3.00	3.00	1.44	1.21
校庭に 太郎が 走る	G2a	2.00	1.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00	2.00	1.00	0.00	1.00	3.00	1.00	1.06	0.77
	av.	2.75	2.56	2.50	2.19	2.56	2.25	2.50	2.13	2.00	1.69	1.81	2.13	1.63	1.75	1.94	2.00		
	stdev	0.58	0.73	1.03	0.91	0.81	0.86	0.63	0.96	0.73	1.14	0.91	0.96	1.02	0.86	1.12	0.89		

未解決の問題

- 経験科学として考えるべき事
  - この結果を得て，F1a, …, G2d のうちのどれが容認可能で，どれが容認可能でないと判断すべき？
- 例えば
  - 平均評定値が 2.0を越える場合を容認可能とし，それ以下を不能？
  - でも，平均評定値が1.8 の例と1.0 の例の違いはどうなる？
- そもそも評定者は均質な反応をしていると考えるべき？　それとも容認度のポテンシャルに統計分布を想定すべき？
- 観察
  - 0, 1, 2, 3 の分布は，対角線について線対称
    - 評定値の分布構造を記述する事を考えないと何がこの結果を生んでいるのはわからないまま

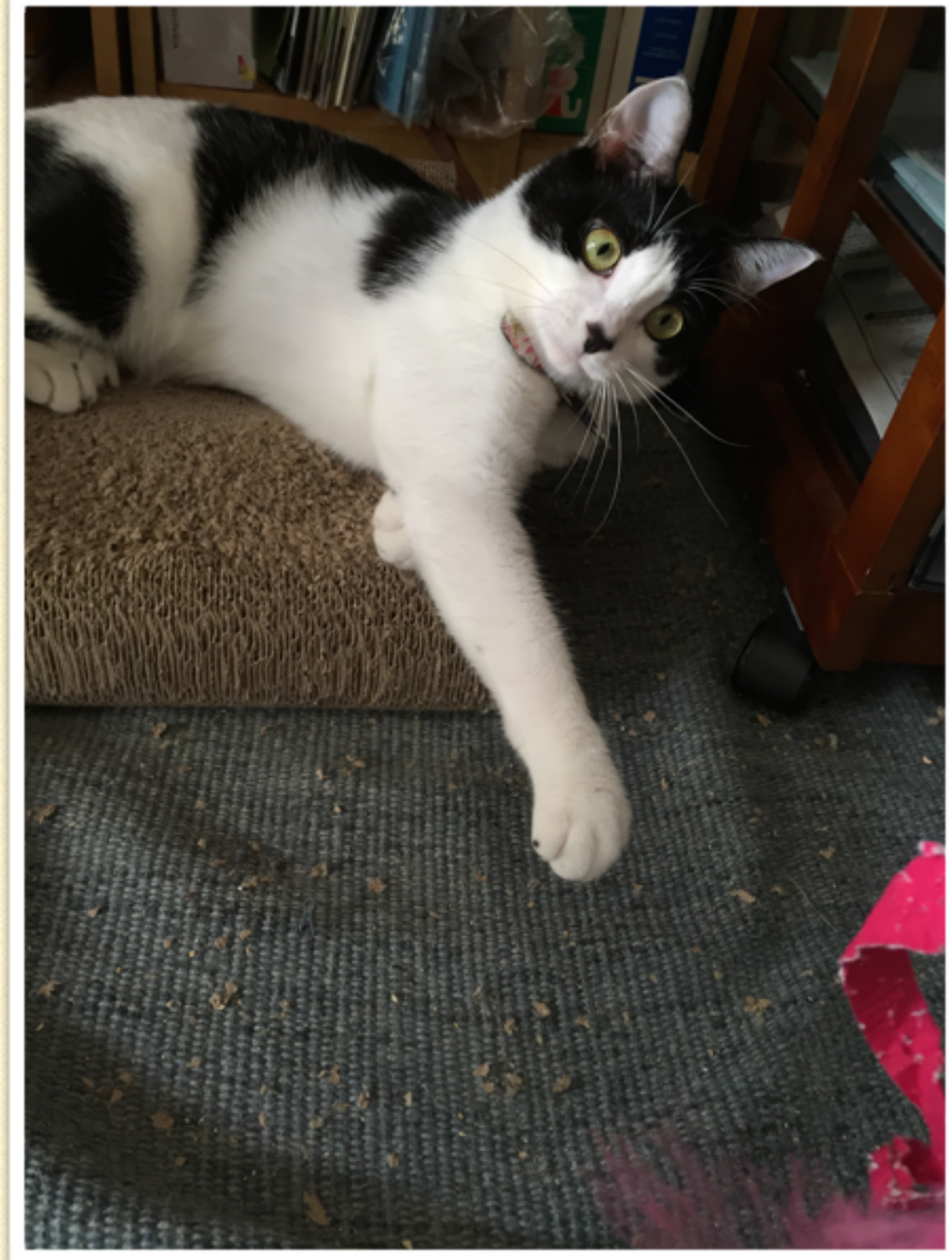
考えるべき分布の構造

表 2: *R* を *E* で表現 (*R* と *E* はいずれも平均値の高い順に並べた)

Rat	F1a	G2c	G2b	F2a	F2b	F1b	G1c	F2c	F1c	F2d	G2d	G1b	G1d	F1d	G1a	G2a	av	stdev
r3	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	2.00	2.00	0.00	0.00	2.50	1.03
r5	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	2.00	3.00	3.00	3.00	3.00	3.00	1.00	1.00	2.56	0.81
r4	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	1.00	2.00	2.00	3.00	2.00	1.00	1.00	1.00	2.19	0.91
r2	3.00	3.00	3.00	3.00	3.00	3.00	2.00	2.00	3.00	3.00	3.00	3.00	2.00	3.00	1.00	1.00	2.56	0.73
r6	3.00	3.00	3.00	3.00	3.00	3.00	1.00	1.00	2.00	2.00	2.00	2.00	1.00	3.00	3.00	1.00	2.25	0.86
r12	3.00	3.00	3.00	3.00	3.00	2.00	3.00	2.00	1.00	3.00	2.00	1.00	2.00	2.00	0.00	1.00	2.13	0.96
r7	3.00	3.00	3.00	3.00	2.00	3.00	3.00	2.00	3.00	2.00	3.00	3.00	2.00	2.00	2.00	1.00	2.50	0.63
r10	3.00	3.00	3.00	3.00	2.00	2.00	2.00	3.00	2.00	1.00	1.00	0.00	1.00	1.00	0.00	0.00	1.69	1.14
r15	3.00	3.00	3.00	3.00	2.00	1.00	3.00	2.00	2.00	1.00	0.00	1.00	1.00	0.00	3.00	3.00	1.94	1.12
r8	3.00	3.00	3.00	2.00	3.00	3.00	1.00	3.00	3.00	2.00	3.00	1.00	1.00	1.00	1.00	1.00	2.13	0.96
r13	3.00	3.00	3.00	2.00	2.00	2.00	2.00	1.00	1.00	1.00	1.00	3.00	1.00	0.00	1.00	0.00	1.63	1.02
r9	3.00	3.00	3.00	2.00	2.00	1.00	2.00	2.00	1.00	2.00	3.00	2.00	2.00	2.00	1.00	1.00	2.00	0.73
r14	3.00	3.00	3.00	2.00	2.00	1.00	1.00	2.00	1.00	1.00	2.00	1.00	1.00	1.00	3.00	1.00	1.75	0.86
r11	3.00	3.00	3.00	2.00	1.00	2.00	1.00	3.00	1.00	1.00	2.00	2.00	1.00	2.00	0.00	2.00	1.81	0.91
r1	3.00	3.00	2.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	1.00	3.00	3.00	3.00	3.00	2.00	2.75	0.58
r16	3.00	1.00	1.00	2.00	3.00	3.00	1.00	2.00	3.00	2.00	1.00	1.00	3.00	2.00	3.00	1.00	2.00	0.89
av	3.00	2.88	2.81	2.63	2.50	2.38	2.13	2.06	2.00	2.00	2.00	2.00	1.75	1.75	1.44	1.06	1.06	1.06
stdev	0.00	0.50	0.54	0.50	0.63	0.81	0.89	0.77	0.89	0.82	0.97	1.03	0.77	1.00	1.21	0.77	0.77	0.77

付録 2

Rをモデルに追加して何がわかるか？



R をモデルに追加すると 1/2

- 確実に分かる事
  - 異なる評定者 $r_1, r_2, \ldots, r_N$ が同一の刺激 e_i に対して異なる評定を与える理由
    - 直接的な理由
      - 解釈の柔軟性=追加する前提の範囲，逸脱検知力の高さ，判断の鋭さなどの個性
    - 間接的な理由
      - 慣れや疲れの影響による判断基準の変化

R をモデルに追加すると 2/2

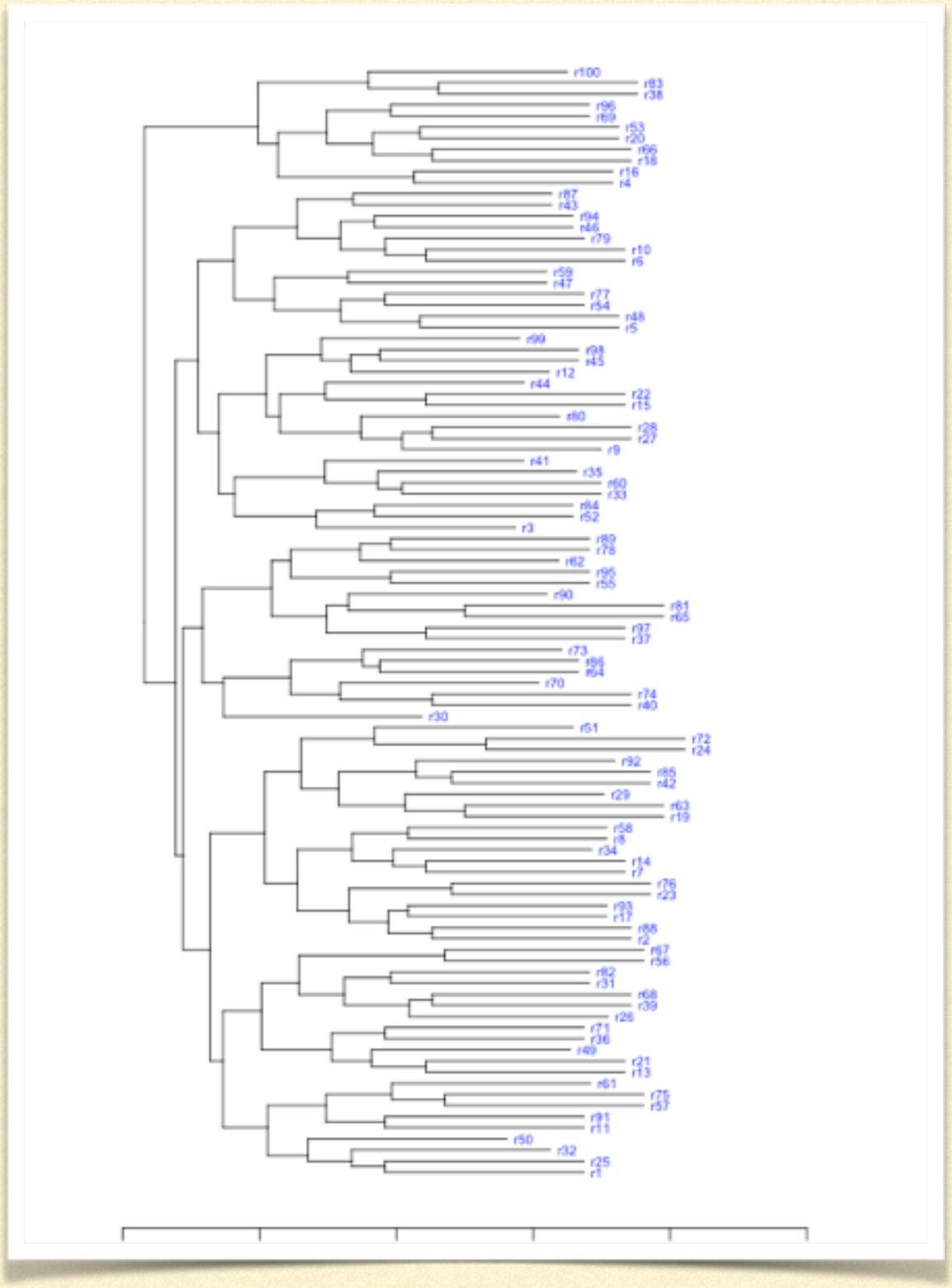
- 分かる可能性が十分にある事
  - 狭義の文脈の効果の実態の解明
    - これは評定傾向が似た評定者を集めて得られた反応を比較しないと正確な記述ができない
  - 社会調査としての容認度の調査

付録 3: R の CLUSTER で欠損値ありクラスタリングを実行するための見本コード

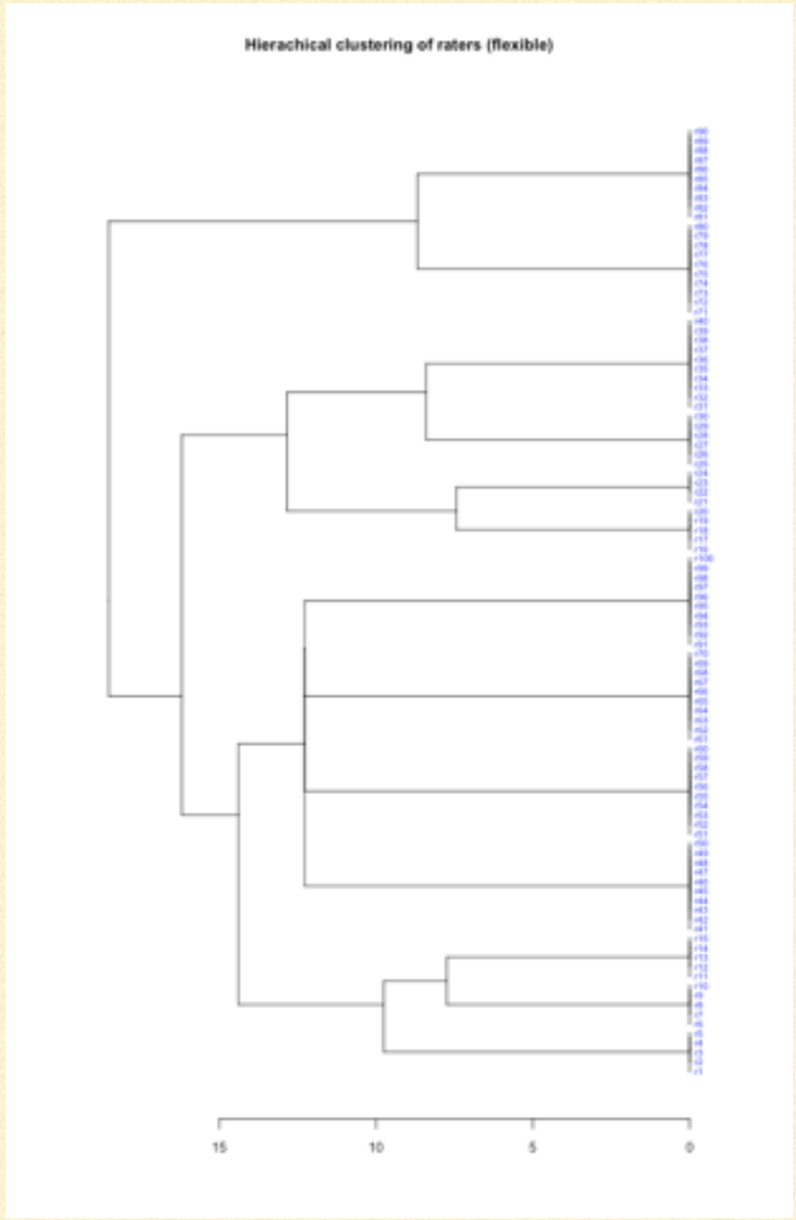
```
# 最初に cluster パッケージの読み込み
library(cluster)
# 次に解析用データをファイル source.tsv から読み込む
# 仕様: 0. source.tsv はタブ区切りテキストファイル,
# 1) 1行目は属性名列, 2) 1列目はデータ名.  3) 欠損値は "na" で表わす.
data <- read.csv("source.tsv", sep="\t", header=T, row.names=1)
# 必要に応じてサンプリング
sampleN=100
data <- data.S[sample(1:nrow(data), sampleN, replace=F),]
# sampleN は可変
# 最後にクラスタリングの実行
# cluster パッケージの agnes と dagnes を使用
# plot 1
agnS.S <- agnes(data, method= "flexible", par.meth= 0.625)
plot(agnS)
# plot 2
dagnS <- as.dendrogram(as.hclust(agnS), hang = 0.2)
plot(dagnS, horiz = TRUE, center=TRUE, main = "Hierarchical clustering of data (flexible)", nodePar = list(ps=20, lab.cex = 0.6, lab.col = "blue", pch = NA))
# nodePar = list(ps=20, lab.cex = 0.6, lab.col = "blue", pch = NA) は描画パラメターの調節
```

付録 4

欠損値の影響


r100
r83
r38
r95
r69
r53
r20
r66
r18
r16
r4
r87
r43
r94
r46
r79
r10
r6
r59
r47
r77
r54
r48
r5
r99
r98
r45
r12
r44
r22
r15
r60
r28
r27
r9
r41
r35
r50
r33
r64
r52
r3
r89
r78
r62
r95
r55
r90
r81
r65
r97
r37
r73
r86
r84
r70
r74
r40
r30
r51
r72
r24
r92
r85
r42
r29
r63
r19
r58
r8
r34
r14
r7
r76
r23
r93
r17
r88
r2
r67
r36
r82
r31
r68
r39
r26
r71
r36
r49
r21
r13
r61
r75
r57
r91
r11
r50
r32
r25
r1

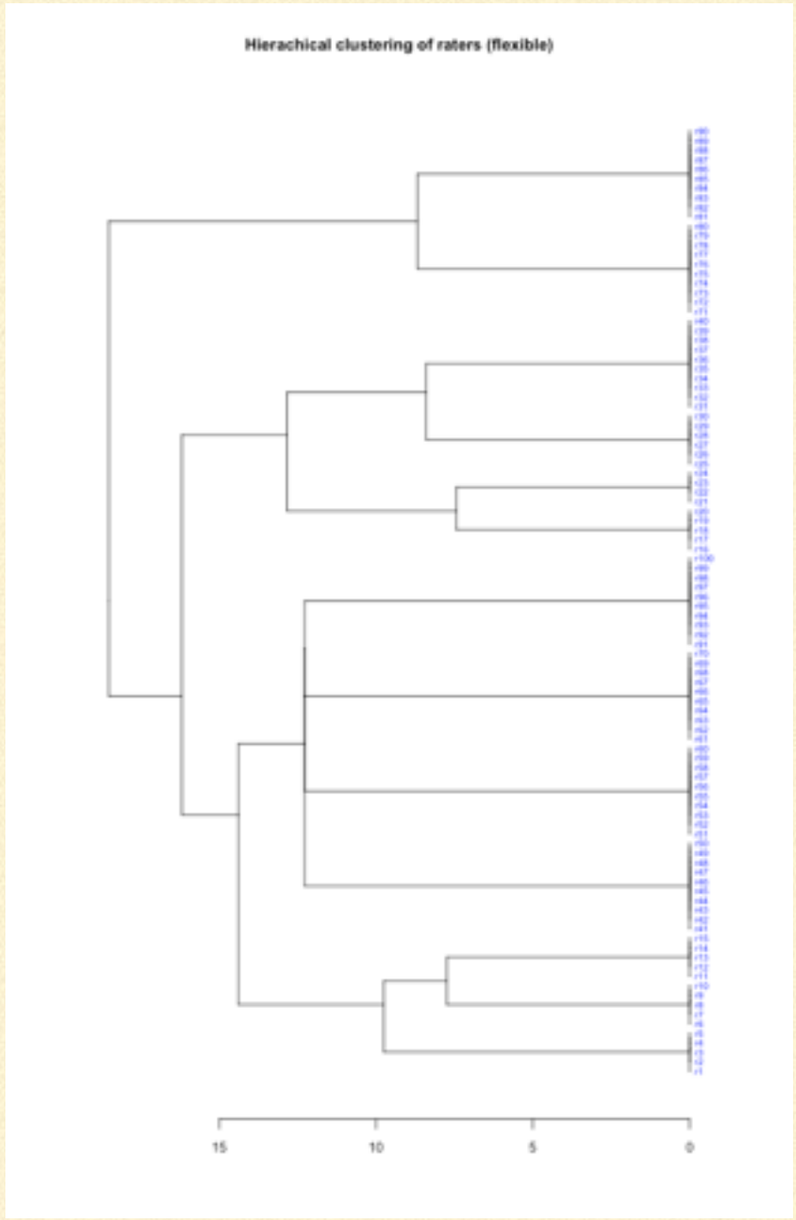
欠損値0%



- 人工的に作ったデータ
  - R: r1, r2, …, r100
  - S: s1, s2, …, s100
  - C1, C2, …, C13 を手動エンコード

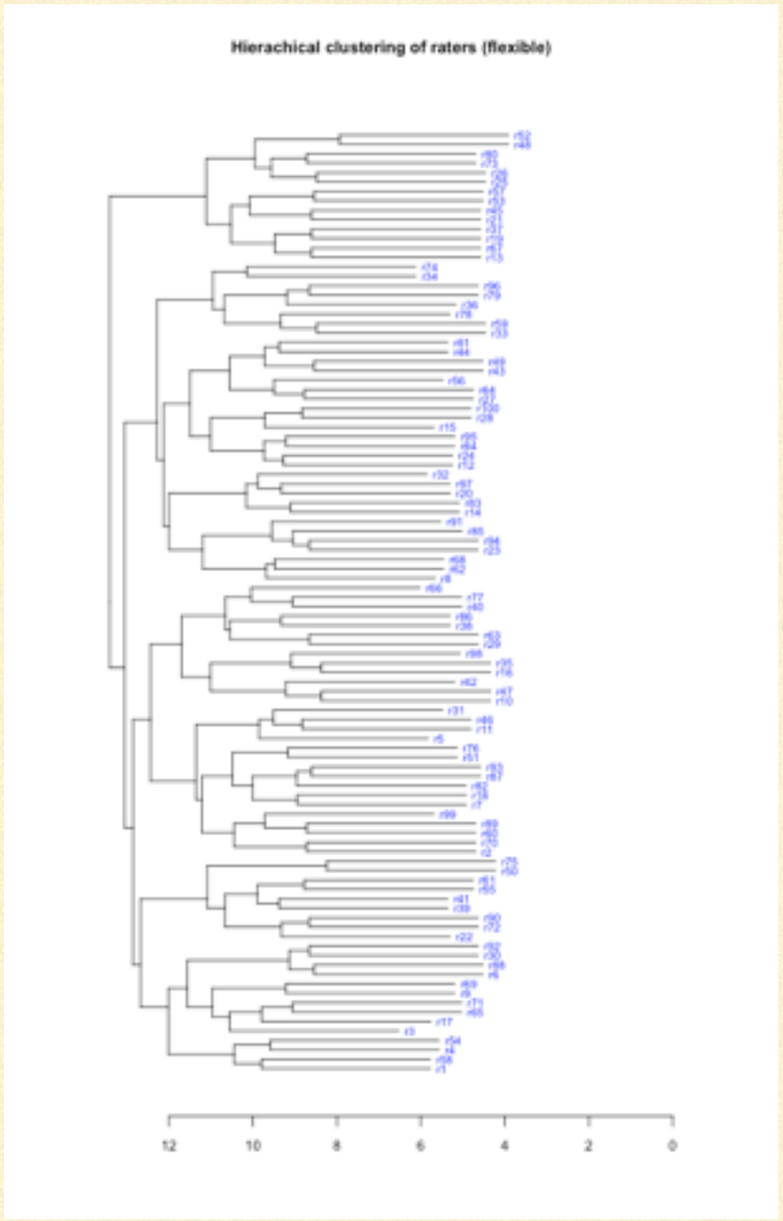
STAT	R.WISE VALUE
AVERAGE	12.60
MEDIAN	10.00
STDEV	4.41
MAX	20.00
MIN	10.00

ランダムな欠損値70%



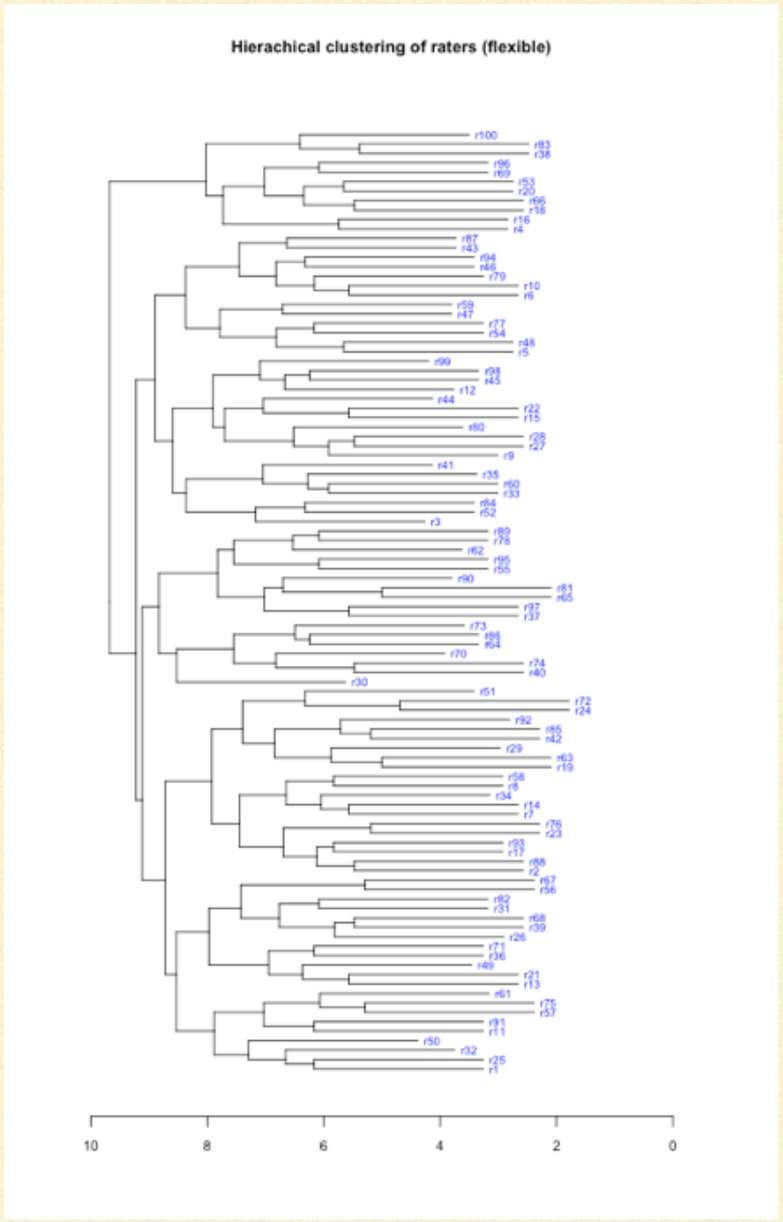
- まったく影響を認めず

ランダムな欠損値80%



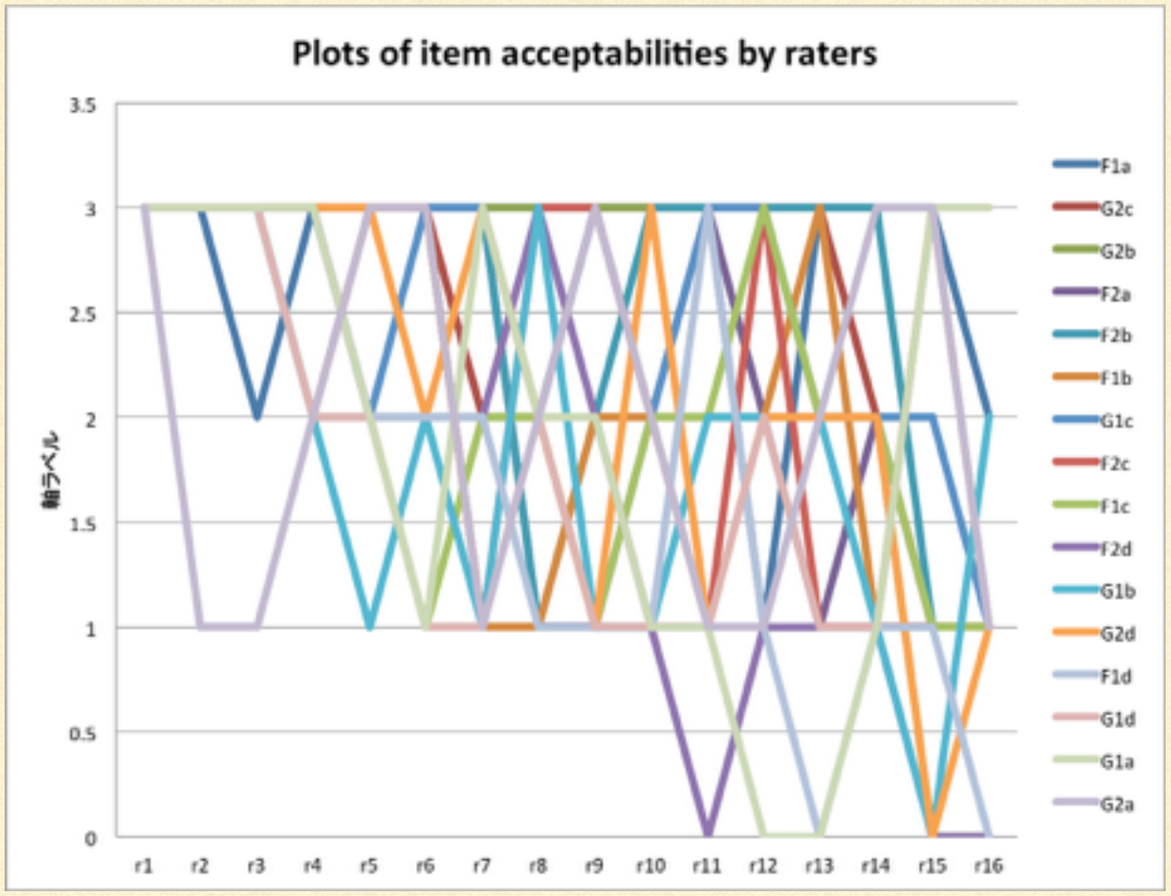
- 枝葉末節に影響がある
- それでも全体のグループ分けは良く保存されている

ランダムな欠損値90%

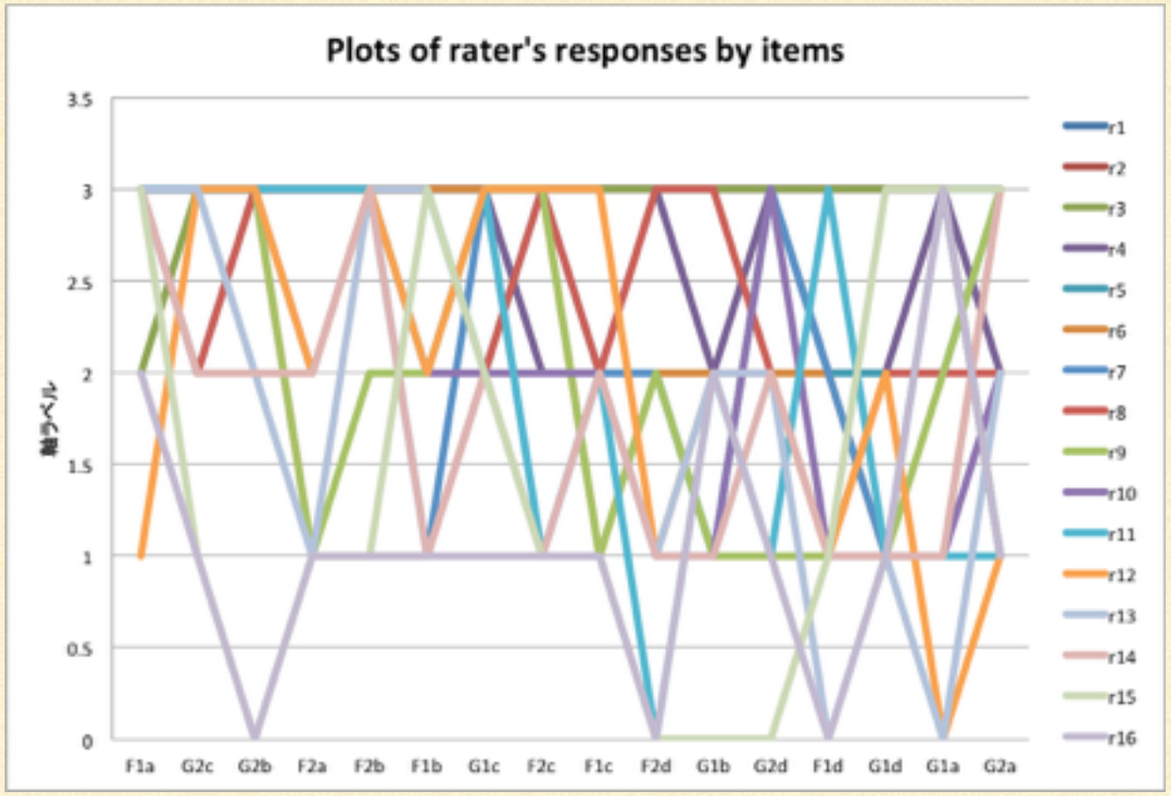


- 更に枝葉末節に影響がある
- それでも全体のグループ分けは良く保存されている

表のデータの解析 1/2


Plots of item acceptabilities by raters
3.5
3
2.5
2
1.5
1
0.5
0
軸ラベル
r1
r2
r3
r4
r5
r6
r7
r8
r9
r10
r11
r12
r13
r14
r15
r16
F1a
G2c
G2b
F2a
F2b
F1b
G1c
F2c
F1c
F2d
G1b
G2d
F1d
G1d
G1a
G2a

表のデータの解析 2/2


Plots of rater's responses by items
3.5
3
2.5
2
1.5
1
0.5
0
軸ラベル
F1a
G2c
G2b
F2a
F2b
F1b
G1c
F2c
F1c
F2d
G1b
G2d
F1d
G1d
G1a
G2a
r1
r2
r3
r4
r5
r6
r7
r8
r9
r10
r11
r12
r13
r14
r15
r16