言語学と自然言語処理：統計的な観点から

統計数理研究所
持橋大地
daichi@ism.ac.jp

「言語における系統・変異・多様性とその数理」シンポジウム
パネルセッション
2018-2-2, TKP東京駅大手町カンファレンスセンター

自己紹介

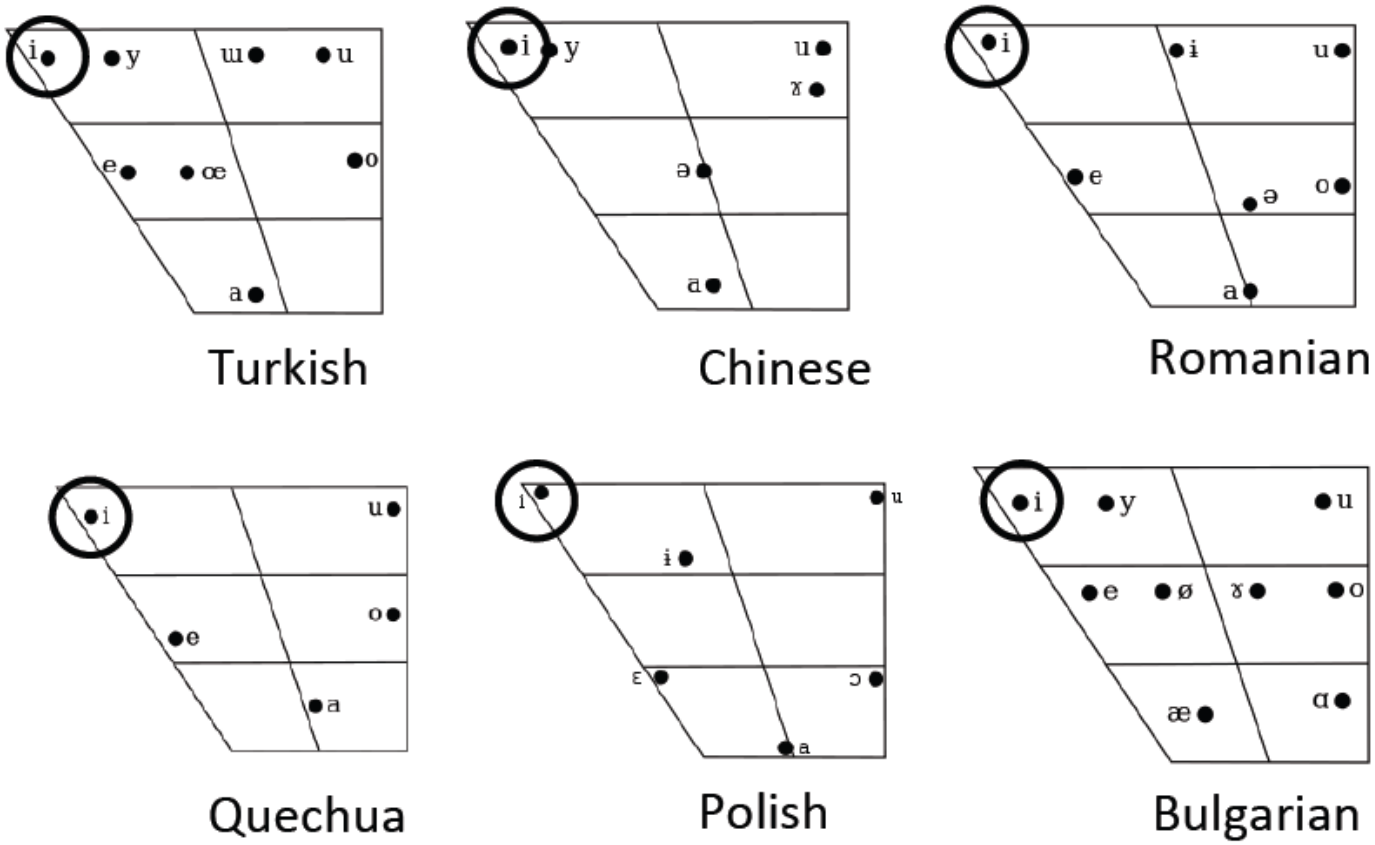
- 1993年 東京大学文科三類入学
- 2005年 奈良先端科学技術大学院大学 情報科学研究科 自然言語処理学講座博士後期課程修了
  博士(理学)
- 2011年～ 統計数理研究所 数理・推論研究系
- 専門：統計的自然言語処理、機械学習

計算言語学と自然言語処理

- ほとんど重なっているが...
- 計算言語学：理学
  - 言語がどうなっているのか、なぜそうなのか、原理は何か
- 自然言語処理：工学
  - 言語をどう処理するか、何ができるのか、性能を上げるにはどうするか
- 計算言語学に興味がある人は1割くらい?
  (若い人は潜在的にはもっと多い可能性がある)

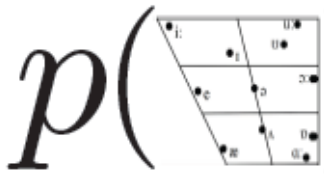
「計算言語学」の研究の例

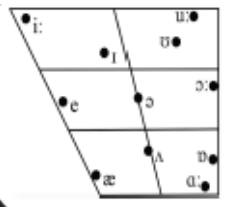
- ACL 2017 Best paper: 言語の持つ母音の分布の学習 (“Probabilistic Typology: Deep Generative Models of Vowel Inventories”)



「計算言語学」の研究の例 (2)

$$p(\ \ ) \propto \mathrm{score}(\ \ )$$

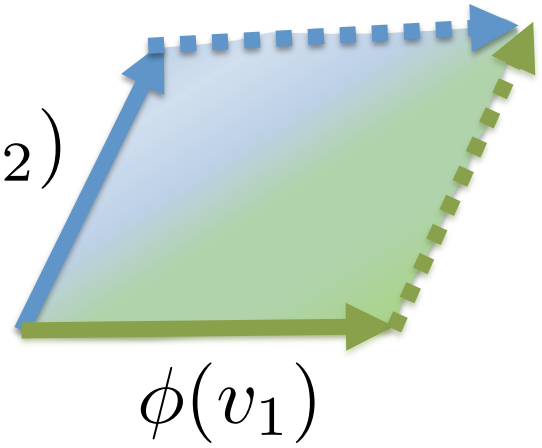




- 行列式点過程で上の確率を定義

$$p(V) \propto \det L_V$$

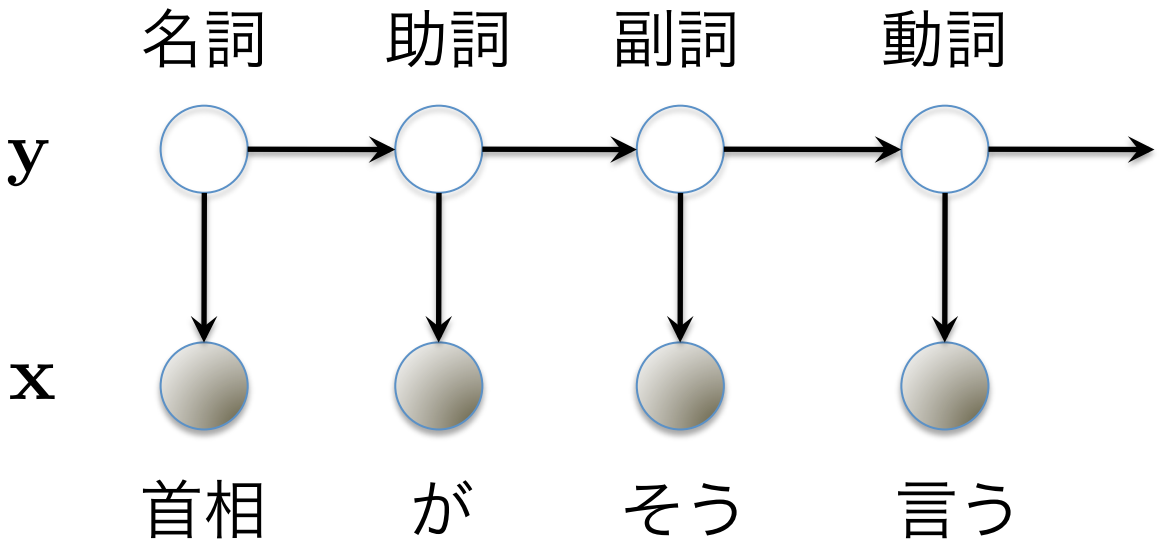
$$L_V = (\phi(v_1), \cdots, \phi(v_M))$$



- Dispersion-Focalization Theory (Schwartz+ (1997)) を考慮
- 音韻の表現ベクトル $\phi(v)$ を同時に学習

「計算言語学」の研究の例 (3)

- 品詞の自動学習



- 品詞＝隠れマルコフモデルの隠れ状態
- Merialdo (1994)で最初にトライされた
  → 当時は上手く行かないという認識

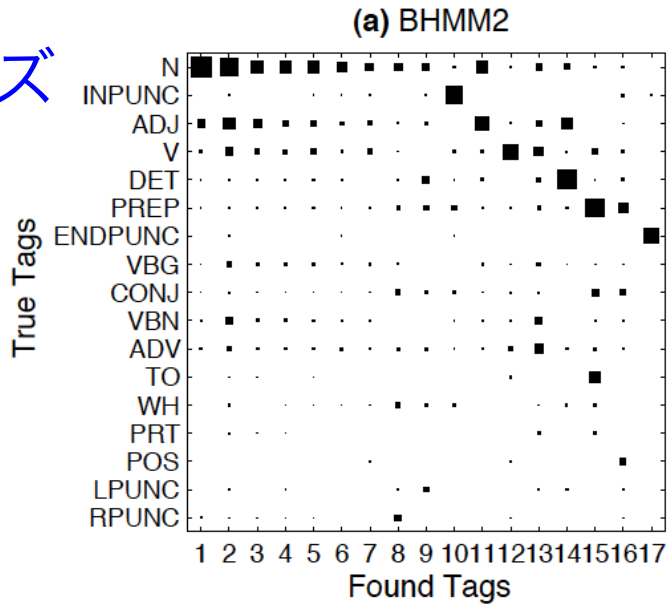
「計算言語学」の研究の例 (4)

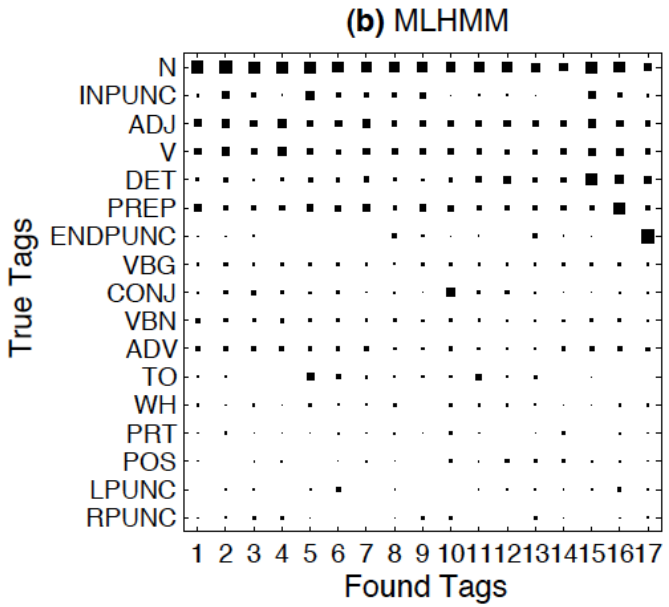
- Goldwater&Griffiths (2007): ベイズ学習で解決

Accuracy	12k	24k	48k	96k
random	64.8	64.6	64.6	64.6
MLHMM	71.3	74.5	76.7	78.3
CRF/CE	86.2	88.6	88.4	89.4
BHMM1	85.8	85.2	83.6	85.0
BHMM2	85.8	84.4	85.7	85.8

- 推定された状態遷移行列が最尤推定とは全く異なる

ベイズ

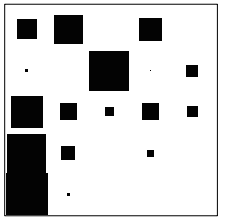




最尤推定

Infinite HMMによる品詞

状態遷移行列

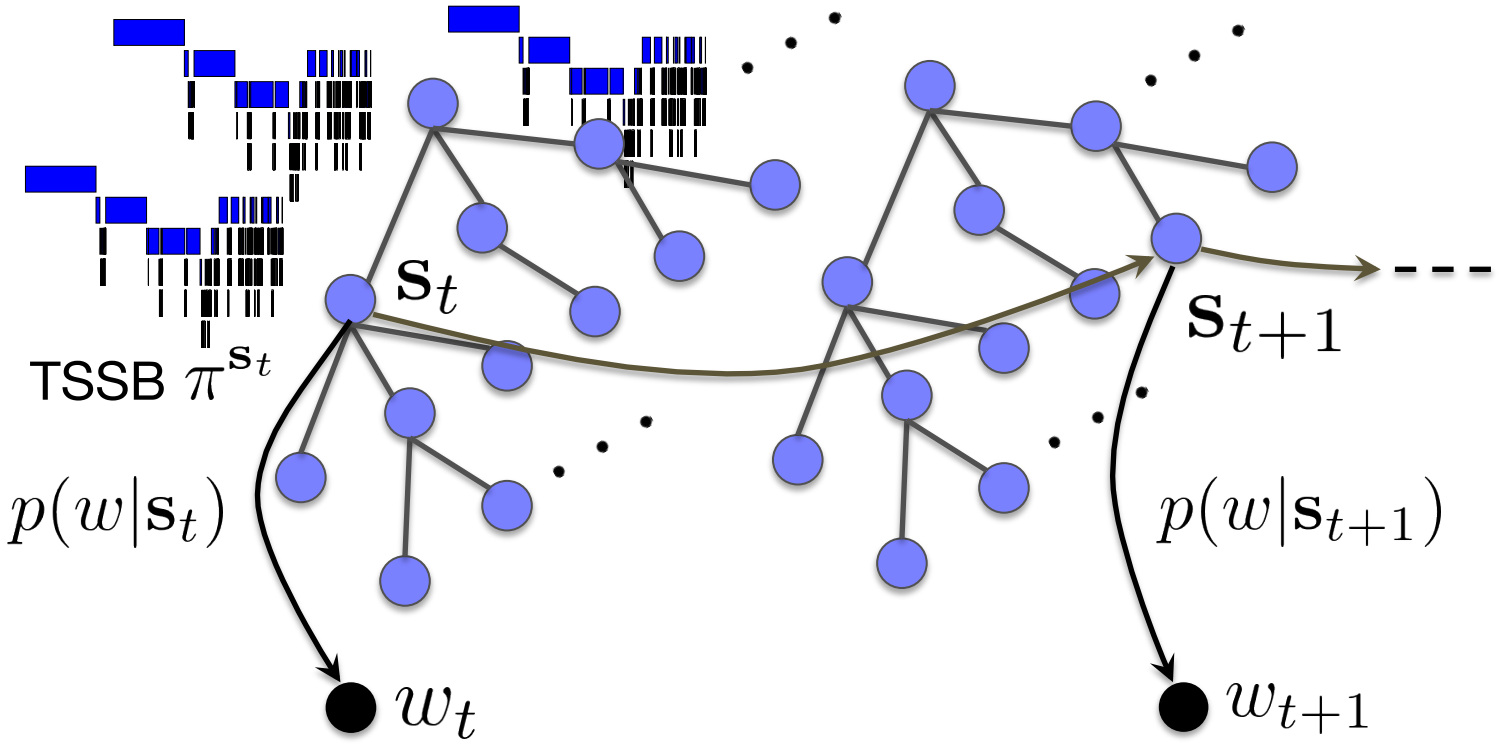


1		2		3		5	
she	432	the	1026	was	277	way	45
to	387	a	473	had	126	mouse	41
i	324	her	116	said	113	thing	39
it	265	very	84	$	87	queen	37
you	218	its	50	be	77	head	36
alice	166	my	46	is	73	cat	35
and	147	no	44	went	58	hatter	34
they	76	his	44	were	56	duchess	34
there	61	this	39	see	52	well	31
he	55	$	39	could	52	time	31
that	39	an	37	know	50	tone	28
who	37	your	36	thought	44	rabbit	28
what	27	as	31	herself	42	door	28
i'll	26	that	27	began	40	march	26

- 教師なしで、品詞に相当するものとその数が自動的に学習できている!

「計算言語学」の研究の例 (5)

- 階層的な品詞の完全教師なし学習 (持橋 2016)
  - HTSSB-HMM=Infinite Tree HMM (NL226で発表)



「計算言語学」の研究の例 (6)

- 教師なし構文解析

パラメータ	具体例	説明
$p_S(stop\|h, dir, adj)$	$p_S(\neg\text{STOP}\|\text{VERB}, \rightarrow, \text{TRUE})$	VERB が右側に子を持たない状態から，一つ子を生成する
$p_A(d\|h, dir)$	$p_A(\text{NOUN}\|\text{VERB}, \rightarrow)$	右側の具体的な子として NOUN を選ぶ

- 詳しくは→『統計数理』64-2 (2016) 特集
  「統計的言語研究の現在」の能地論文
  “文に隠れた構文構造を発見する統計モデル”

何が足りないか? の例

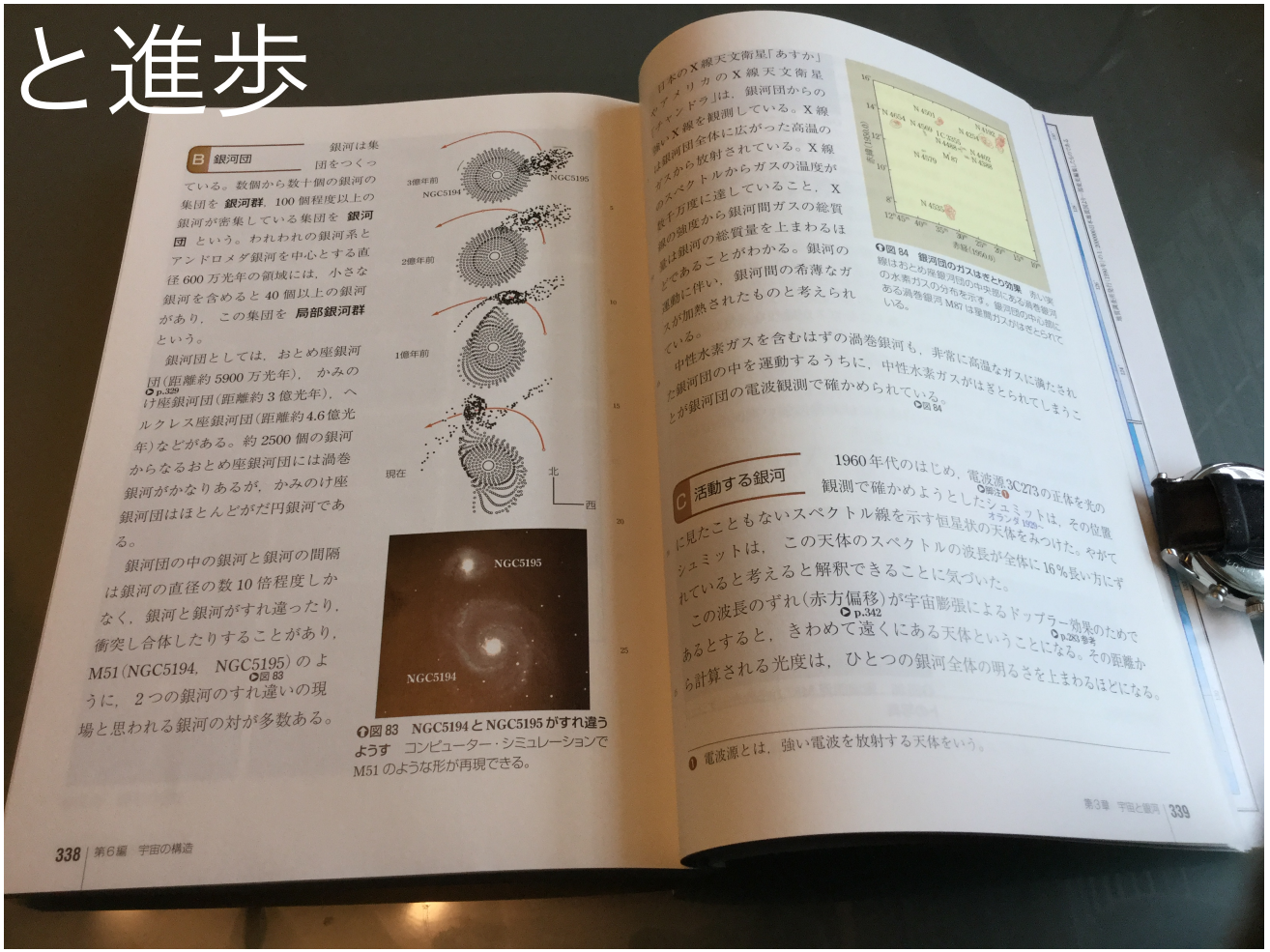
× 主題化、痕跡、モダリティ、反実仮想の数理モデル
△ 系統、変異、多様性のモデル (本シンポジウム)
○ 位相、文脈、形態論の一部

評価の問題：「何のタスクの役に立つのか」が
問われやすい

どこで、誰がやるか?

- 工学部言語学科 → 自然言語処理のこと
- 理学部言語学科 → 文学部言語学科に、数理・理論コースを設ける、自然言語処理の人を適宜呼ぶ
  - eg. 東大経済学部統計コース
  - 文学部心理学科
  - ただし、言語学では個別言語学があるため、心理学と状況が異なっているようにみえる

科学と進歩



B 銀河団

銀河は集団をつくっている。数個から数十個の銀河の集団を**銀河群**，100 個程度以上の銀河が密集している集団を**銀河団**という。われわれの銀河系とアンドロメダ銀河を中心とする直径 600 万光年の領域には，小さな銀河を含めると 40 個以上の銀河があり，この集団を**局部銀河群**という。

銀河団としては，おとめ座銀河団(距離約 5900 万光年)，かみのけ座銀河団(距離約 3 億光年)，ヘルクレス座銀河団(距離約 4.6 億光年)などがある。約 2500 個の銀河からなるおとめ座銀河団には渦巻銀河がかなりあるが，かみのけ座銀河団はほとんどがだ円銀河である。

銀河団の中の銀河と銀河の間隔は銀河の直径の数 10 倍程度しかなく，銀河と銀河がすれ違ったり，衝突し合体したりすることがあり，M51(NGC5194，NGC5195)のように，2 つの銀河のすれ違いの現場と思われる銀河の対が多数ある。

3億年前 NGC5194 NGC5195
2億年前
1億年前
現在
北
西

NGC5195
NGC5194

図 83 NGC5194 と NGC5195 がすれ違うようす コンピューター・シミュレーションで M51 のような形が再現できる。

338 第6編 宇宙の構造

図 84 銀河団のガスはぎとり効果

C 活動する銀河

1960 年代のはじめ，電波源 3C 273 の正体を光の観測で確かめようとしたシュミットは，その位置に見たこともないスペクトル線を示す恒星状の天体をみつけた。やがてシュミットは，この天体のスペクトルの波長が全体に 16 %長い方にずれていると考えると解釈できることに気づいた。

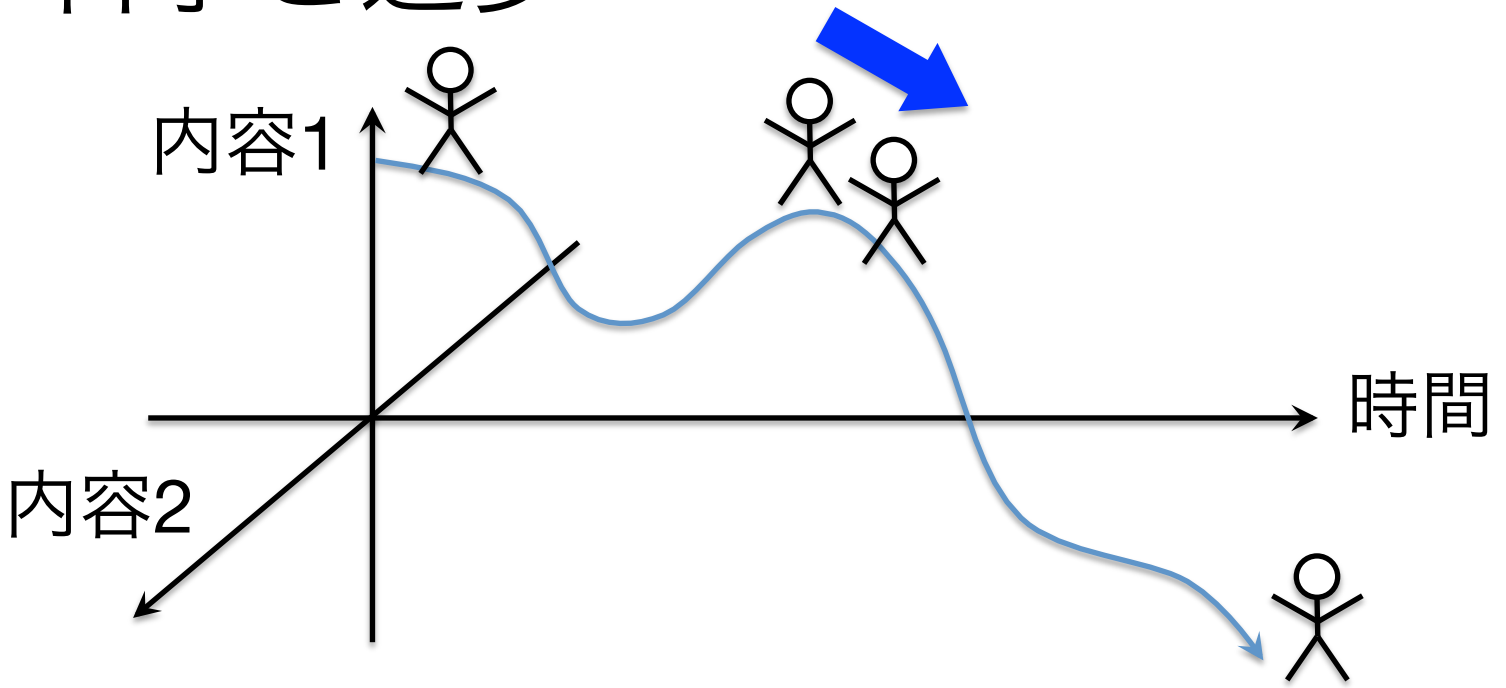
この波長のずれ(赤方偏移)が宇宙膨張によるドップラー効果のためであるとすると，きわめて遠くにある天体ということになる。その距離から計算される光度は，ひとつの銀河全体の明るさを上まわるほどになる。

電波源とは，強い電波を放射する天体をいう。

第3章 宇宙と銀河 339

- 地学の教科書：20年前と大きく様変わり
(宇宙の記述ですら！)

科学と進歩



- 1年や2年では急激に変化しないので，「前と同じでよい」と思い込んでしまう
- **実際は、最前線は全く違ったところにいる!**
  - 過去の内容も含んで止揚されていることが多い

学問の進歩

- 基本的な数理は当たり前になる
  → 数理・データを前提にした言語研究の大枠を描く必要
- (新しい)理論言語学
- 実験言語学、個別言語学
- 参考：物理学のスペクトラム

数学　理論物理　応用物理　実験物理　物質世界

素粒子物理　物性物理