MACS SG4

種々の実例から考えるパターン理論

発表者：坂田諒一(物理)
メンバー(学生)：藤田和樹(物理)、中村ちから(数学)
(教員)：石塚裕大(数学)、林重彦(化学)

活動内容

✓ **題材** “Pattern Theory”

✓ **内容**

①文章のモデル化

・KL distance

・**n-gram model**　etc.

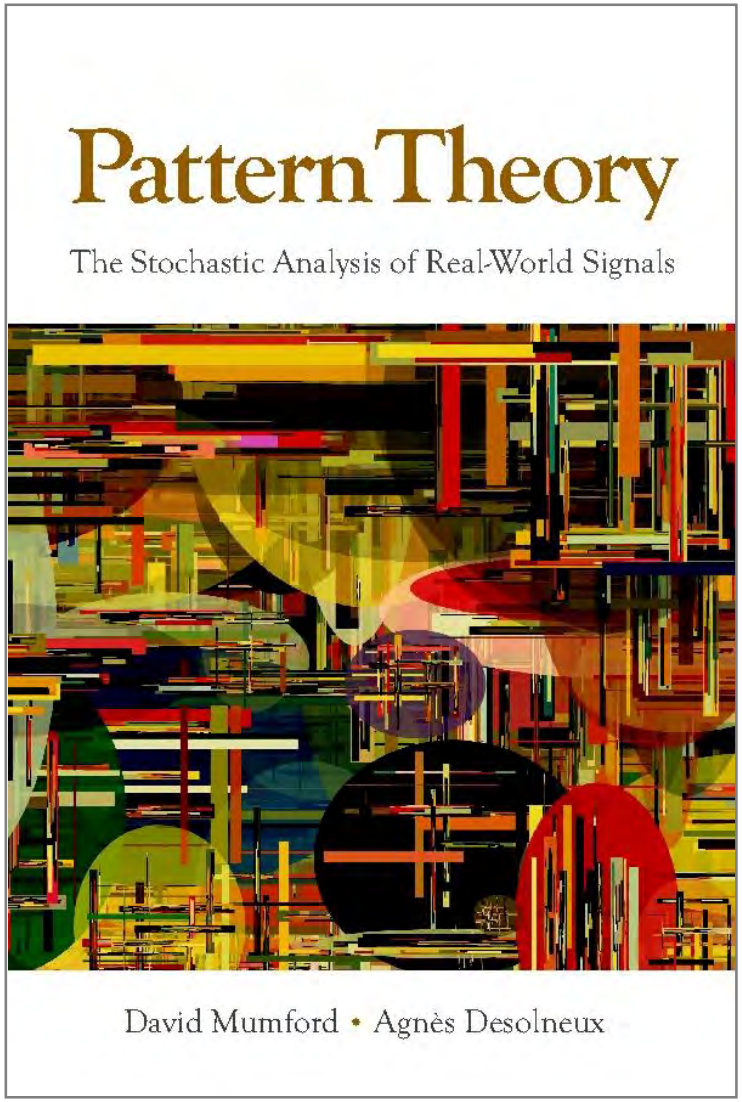
②音楽のモデル化

・Music model

・EM algorithm　etc.

③その他

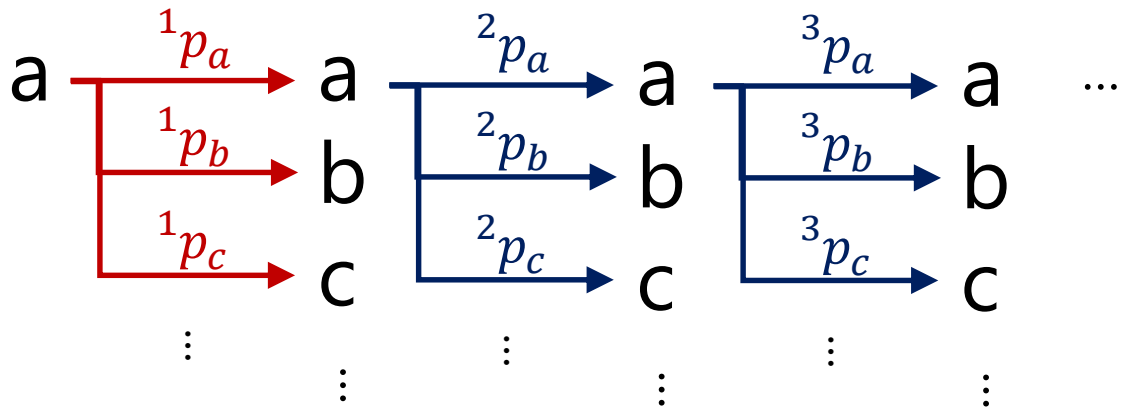
・Brief propagation　etc.



http://www.dam.brown.edu/ptg/MDbook/index.html

n-gram model とは？

- 新しい文章(英語)を生成するプログラムを作りたい
- 1文字のアルファベットから次のアルファベットは予測可能。(連続する2文字間の確率分布を利用)



- さらに、2文字・3文字…のアルファベットを使って次のアルファベットを予測すればより精確な文章になる

⇒連続するn文字間の確率分布を用いるとき
n-gram modelという

n-gram model による文章の生成 (1)

✓ Project Gutenberg から
“ハックルベリーフィンの冒険”の文字列を取得

https://www.amazon.co.jp/Adventures-Huckleberry-Finn-Mark-Twain/dp/1784043176

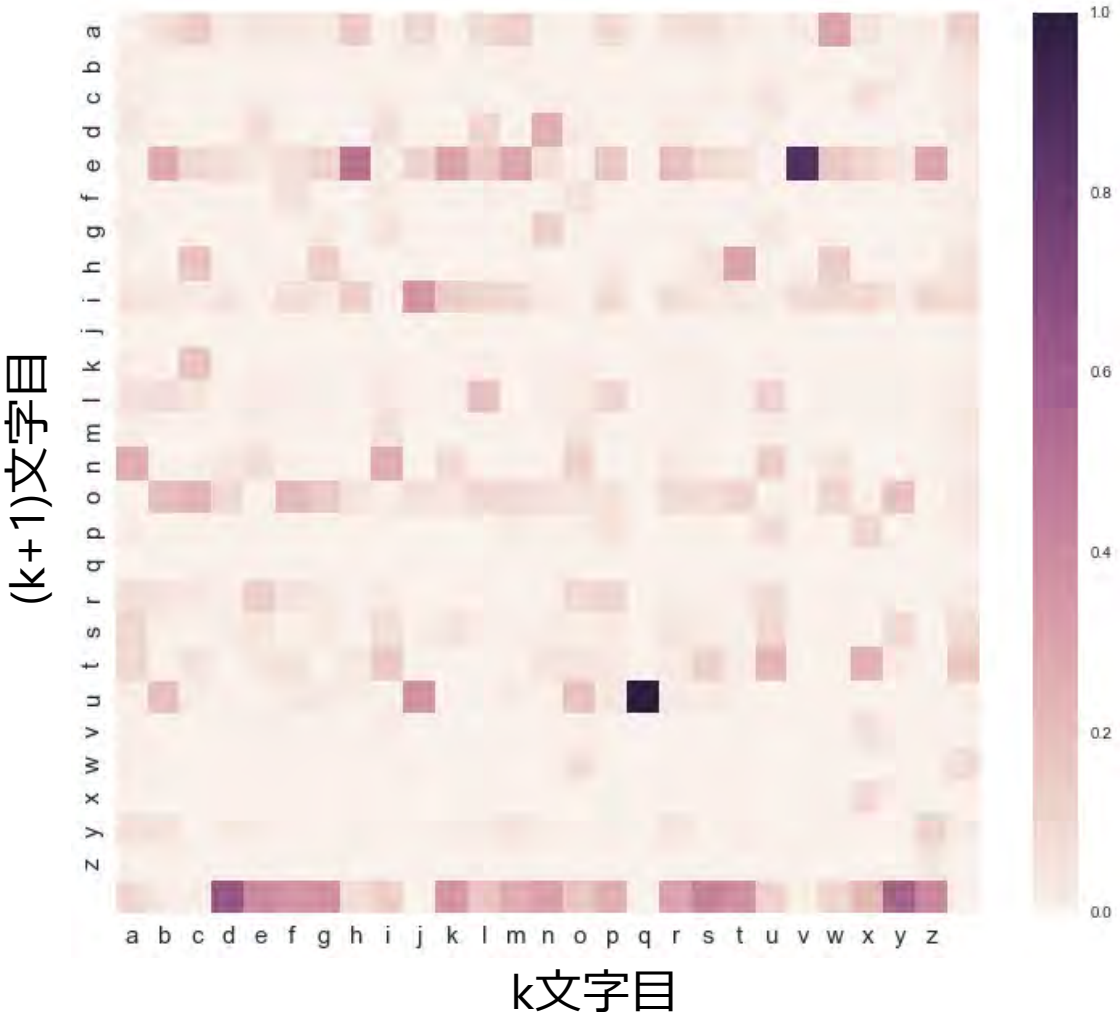
n-gram model による文章の生成 (2)

Step1
文字列を取得し、
ピリオド等を削除

Step2
2文字間の確率分布
を作る(右図)

Step3
同様に3,4文字間の
確率分布を作る

Step4
4文字間の確率分布
を用いて新しく文章
を生成する



結果とまとめ

n-gram model (n=4)

Input

…… So I didn't think much of it. But I never said so. I asked her if she reckoned Tom Sawyer would go there, and she said not by a considerable sight……

Output

……you home told hear a but that wants all this some and just pianoe d dead and we says in a little to get all so ther a bile ain and that i ll rive good this brits soon to me going in lity only dollars i might of the up nevery to said i way broked ……

- ✓ n-gram model において4文字間の確率分布から文章を生成するプログラムを作成した
- ✓ 生成された文章は文法だけでなく、単語自体が間違っているものも多い
- ✓ より多くの文字数に対して確率分布を作れば改善が期待できる